Modeling Student's Churn in Higher Education in Bosnia and Herzegovina

Dragana Preradović Kulovac

The Business Informatics Doctoral Program

Peter Racsko PhD

Thesis committee members:

1

2

3

4

5

Corvinus University of Budapest
Doctoral School of Economics, Business, and Informatics

Modeling Student's Churn in Higher Education in Bosnia and Herzegovina

Doctoral dissertation

Dragana Preradović Kulovac

Budapest 2024

I would like to extend my deepest gratitude to all those who made this research possible.

First and foremost, I want to express my heartfelt thanks to my mentor, professor Peter Racsko. His guidance, invaluable support, and professionalism over the years have been instrumental to this work. Without his mentorship, this research would not have been possible.

I am also deeply grateful to the esteemed members of my committee—professor Szabina, professor Tibor, and professor Ivan—whose lectures, insights, and constructive feedback greatly enhanced the quality of this thesis.

A special note of thanks goes to Professor Andrea Ko for her pivotal role in approving my admission to this doctoral school as the first Stipendium Hungaricum student. Her support marked the beginning of this incredible journey.

I would also like to acknowledge the University of Banja Luka for providing the essential data that formed the backbone of this research.

To my husband, my most steadfast pillar of support, thank you for your encouragement and belief in me from the very beginning. Your presence made this journey possible.

Finally, I extend my sincere thanks to the Tempus program and the Government of Hungary. Their support was crucial, not only in making this research possible but also in significantly impacting my life.

Contents:

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES IN APPENDIX

LIST OF FIGURES IN APPENDIX

LIST OF ABBREVIATIONS

AI, Artificial Intelligence
B&H, Bosnia and Herzegovina
BHAS, Agency for Statistics of Bosnia and Herzegovina
Bool, Boolean coded variable(s)
CRISP-DM, Cross Industry Standard Process for Data Mining
DT, Decision Tree
EWS, Early Warning System
GPA, Grade Point Average
ECTS, European Credit Transfer and Accumulation System
EDA, Exploratory Data Analysis
EDM, Educational Data Mining
ERIC, Institute of Educational Science
HE, Higher education
HEI, Higher education institution
HGBC, Histogram Gradient Boosting Classifier
IEEE, Institute of Electrical and Electronics Engineers
ML, Machine learning
MLP, Multilayer perceptron
NN, Neural network
NB, Naïve Bayes
LADA, Learning Analytics Dashboard for academic Advisors
LR, Logistic regression
PI, Permutation (feature) importance
RBF, Radial Basis Function
RF, Random forest
ReLu, Rectified Linear activation function
RFC, Random Forest Classifier
ROC AUC, Receiver Operating Characteristics curve Area Under The Curve
RS, Republic of Serpska, entity in Bosnia and Herzegovina
SVM, Support Vector Machine
SMOTE, Synthetic minority over-sampling technique
UAR, Unweighted Average Recall
UNIBL, University of Banja Luka
USA, United States of America

1 INTRODUCTION

With less than 3.2 million of people, good geographical and climate location, with plenty and variety of natural potentials and resources, Bosnia and Herzegovina (B&H) has all pre conditions for wealthy of their residences. Still the macroeconomics parameters shows otherwise: the country is listed on the 121st place at the World Bank's list in 2022 with 7.6 thousand of USD of Gross domestic product (GDP) per capita. To compare, all B&H border countries have higher GDP per capita rank: Croatia (72), Serbia (107), and Montenegro (102) (World Bank, 2023). [1]

In the long-run, a one way of increasing countries global competitiveness is by having more people with tertiary degree. B&H has relatively cheap higher education (HE) with 39 higher education institutions (HEI), and among them eight public universities. Nevertheless, the number of HEI enrolled students' precipitous decreases. Country-level enrollment statistics have been available since 2007 and the number of enrolled students is at the lowest level ever observed. One approach to the achieving the goal of having more highly educated people is to tackle the student's HE churn in the country. B&H has no statistical reports, data, or estimation of student's HE attrition. Country does not have research of the reasons of leaving the HE, or research that explains the enrollment decrease. Also, there is no information of students' trajectories in years after their withdrawal from HEI. One of the reasons of lack the churn information may be the wide administrative and governmental division of the country, with 13 Ministries of educations, each with they own education jurisdiction and education programs, and lack of common database for all public universities. In this research we estimated the student's churn at the sample of 37.6 thousand of students within 12 years of data collected at one of the eight public universities in the country. We employed machine learning (ML) models to predict student's leave at earliest stage. Our model correctly predicts 86 of 100 dropouts. The purpose is to propose and deploy a model of early HE churn detection in public universities, and set up a base of country's HE's dropout research.

A brief description of the repercussions of the education system and the level of higher educated citizens to the countries development are provided in the introductory part of

[1] To compare, Hungary is ranged as 73.

the research. The multidimensional consequences of low level of HE country's population are considered for the country, individuals, and society. Then the problem statement in the Bosnian and Herzegovina's HE environment is given, describing the aims, objectives, results, research questions, and its uniqueness. Education in the long-term for a small country, such as B&H, where the rate of HE citizens is low, can be a one of the important lever of development.

1.1 General overview

The education system of any country has one of the greatest impact on the country's development, progress, and global competitiveness. The level of population education in a country correlates with the factors of development at the state level and at an individual level. Studies have shown that the level of education of citizens has a positive effect on the GDP and tax revenues (Alliance for Excellent Education, 2003e). A strong positive correlation between the growth of GDP per capita and education expenditure was found in developing countries (Appiah, 2017). The same effect of HEIs to the economic growth was confirmed in the EU countries (Pastor *et. al*. 2018), as well in the particular cases, like in Romania, where the research of the effect of skilled workforce on country's GDP also found a positive correlation (Teodorescu, 2018).

At the state level, the more citizens are educated, the more new jobs are created. This increases consumption of real estate and automobiles, impacts quality of life, and the production sophistication, too. The country's tax revenues grow and the government can spend more money on HE, creating more educated citizens. The increase in the education level among a country's citizens reduces the crime rates. The economic burden of one prisoner per year is equivalent to the cost of two to four students (before tertiary education) (US Department of Education, Policy and Program Studies Service, 2016).

Better educated people tend to take preventive measures to decrease their health risks and invest in private insurance, resulting in fewer costs being covered by government health insurance. Another important aspect of the population's education is mortality rates, which were, according to AEE, 2.5 times higher among poor and people with less than 12 years of schooling, than to highly educated in the US (Alliance for Excellent Education, 2011 and 2003b)

The lack of a highly educated workforce has a negative impact on global, national competitiveness and local investment since the gap between high-skilled and low-skilled workers continues to widen as the labor market changes. Studies show that university

graduates earn higher salaries than ungraduated and find jobs more easily. The unemployment rate in the US in 2015 was 21.4 percent for dropouts and 6.0 percent for university graduate students, while this difference was smaller in 2017 (10.4 percent for all dropouts and 5.8 percent for university dropouts), the trend is still present (De Brey *et al*., 2021).

Among the various significant factors that indirectly influence the education level of citizens, the university student outflow is considered an important factor in the education system by European Commission and United Nations (UN), too. The 2030 Agenda for Sustainable Development UN identifies education as one of the most important factors for a sustainable economy. The European Strategy for Smart, Sustainable and Inclusive Growth 2020 also set a target that at least 40 percent of adults aged 30 - 34 should have a tertiary degree. According to EU Commission, increasing completion rates and reducing dropout rates is the key strategy to achieve this goal (European Commission, 2015).

Dropout, attrition, or churn in higher education occurs when students leave the HEI without obtaining their degree or continuing their education elsewhere. The more comprehensive, operative definition of churn is given in chapter 5.1.

Dropout has far-reaching consequences for the country, the universities, as well as for individuals. At the university level, dropout represents the loss of HE resources and opportunity costs, and when high, it can be an indicator of inefficiency in the education system (OECD, 2021). Some countries penalize universities with high dropout rates as part of their assessment of high education (Schnepf, 2014).

At the individual level, in the literature, dropout is portrayed as a waste of time and resources.[2] Income disparities are the second differentiator between churned and non-churned university students. Comparing earnings across Europe, dropouts earn, on average, 8 percent more than those who never entered higher education but 25 percent less than university graduates (Berlingieri and Bolz, 2020). According to OECD 2022 report, the full-time workers with bachelor degree have 44 percent (OECD countries average), 38 percent (EU22 average) higher salaries than full-time workers with high school degree. The salary advantages grow with age: for people aged 45 - 54, the difference is 75 percent comparing to their peers with high school diploma (OECD, 2022).

[2] In the age of the wide availability of accessible online courses for self-study, dropout is still seen as a negative phenomenon. There are proven successful cases of university dropouts, but they are still rare.

Appropriate solutions to reduce dropout rates consist of policies, strategies, and standards to identify, classify, and respond to negative trends. But first, policymakers need to understand the reasons for dropouts. According to the European Commission's main report on student dropout and completion in European HE, the effectiveness of the tools used by policymakers has not been studied comprehensively and in sufficient detail (Vossensteyn *et al*., 2015).

One of the ways to tackle university dropout is by using machine learning (ML) models to prevent attrition. With the advent of ML, numerous predictive models have been developed to identify and classify student retention, success, graduation timing, learning paths, dropout, etc.

1.2 Problem statement

The problem that this thesis addresses is identification of students who are at risk of dropping out of their HE, by employing the ML models, to prevent it on time. At the beginning of research the significant gap in the information base was found: In B&H are no reports of dropouts at the HE level, no estimates of university dropouts, or a warning system for students at risk of leaving at publicly funded HE institutions. There are also no reports of the reasons for leaving the HEI, or for those who continues their education somewhere else.

Not solving the addressed problem may make the current unenviable situation worse. According to The Global Competitiveness Report 2019, B&H is in first place of 141 country by brain drain on the scale 1 – 7, reaching the 1.76 score, where 1 means that all highly educated people leaving the country, and 7 represents the value when they all are staying in the country, (Schwab, 2019). In B&H, the share of people with vocational and HE (college, university, masters, doctoral degrees) in the labor force is only 15.3 percent, while this share in the working-age population is even lower, 9.6 percent by last available data for 2019, (Mijović *et al*., 2019).

In the last eleven years, the total number of students at HE institutions in B&H decreased by more than 40 percent, Figure 1.[3] The consequences of inaction and ignoring high HEI student's churn and decrease of enrolled students have been widely visible in the last

[3] This can be the cause of the visa liberalization process for B&H citizens in some West European countries (Germany, for instance). The adults and young people are leaving the country for a job abroad. There is also a broad range of scholarship opportunities abroad for Bosnian youth, which goes hand in hand with more than a decade of a negative annual increase in population and causes the decrease of freshmen at Bosnian universities.

decade in B&H. Due to a significantly less number of students, universities are forced to reduce the offer of study programs, HE staff, and their budgets causing the long-term negative consequences for the country, as well as for HE institutions. Future freshmen can face higher tuition fees, limited diversity offer of study programs, and lack of quality HE staff. In the long run the B&H may suffer from lack of educated professionals, too.

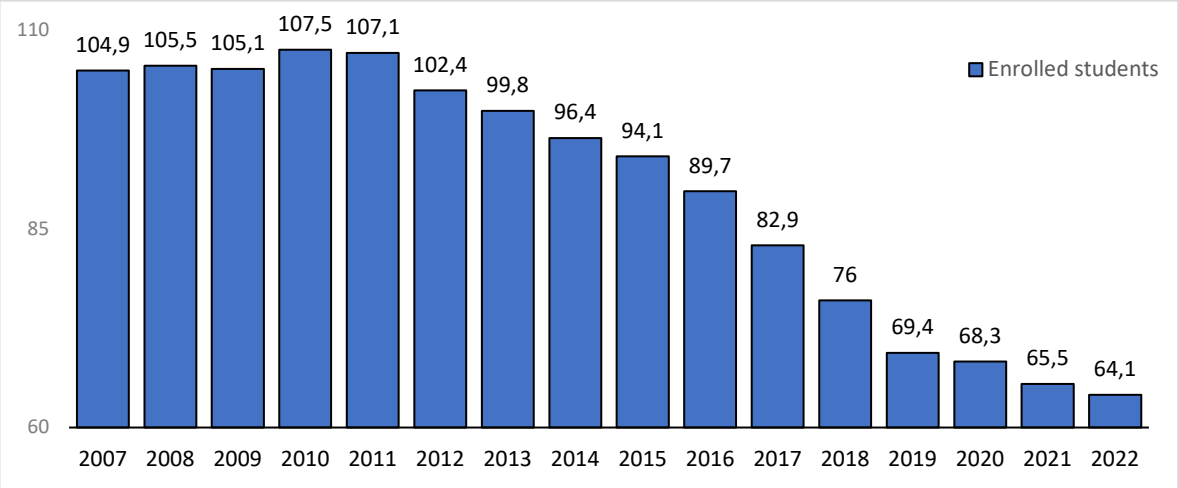


Figure 1 – Enrolled students in all years of study, 2007/08–2022/23 academic year, in B&H (in thousands)

*Data refer to students at universities, schools of higher education, and religious HE. Including students at the bachelor level (first cycle) who study according to the old program and students who study according to Bologna compliant program (first cycle and integrated I and II cycle).

Source for 2021 and 2022: BHAS, Demography and Social Statistics, Higher education in the school year 2021/2022, First release, No.2, Sarajevo, page 2, Table 1. The previous data are from BHAS, B&H in figures 2007-2020.

This research contributes to the enrichment of the field of Education Data Mining (EDM) in the domain of the specificity of the dataset (binary features, missing data, lack of variables) on which we train the variety of ML models and implementing the SHapley Additive exPlanations (SHAP) and Permutation importance (PI). Dataset is obtained from UNIBL, with dominantly categorical and very modest amount of socioeconomics, academics and secondary school variables, and significant amount of missing data. The research brings additional benefits by providing country related dropout data and reasons for churn, for the first time at one university in B&H.

Identified stakeholders of this research are students, the University management, the Ministry of Higher Education, and the public. The highest impact of the research results is for the UNIBL management, since University loss significant amount of money due to

students churn, each year. The results are also part of broad research which is going to spread on the five[4] public universities in the country.

Since this is the first HEI attrition research in B&H and is limited to only one public university, the need for additional research is inevitable. The model has to be enriched with important missing variables, which have been confirmed by other researchers to play an important role in prediction. Also, it is necessary to expand the time frame of the collected data to observe the trends that affect the prediction in the longer run.

1.3 Research aim, objectives and results

Despite a decrease of more than 40 percent in the number of HE students in the country, there is no information about HEI student's attrition or prevention strategies. The main goal of this thesis is to identify students who are at risk of dropping out at UNIBL in challenging dataset. Through the following list of objectives, it is briefly explained how the thesis aim will be achieved:

- Conducted desk-research of national and UNIBL rules, and laws valid between 2007 and 2018 school year, related to HE study – to understand the enrollment, study and drop out process, to apply it in a code stage of research.
- Data collection at UNIBL: obtaining the data set of University student's study data between 2007 and 2018 school year, to model the attrition.
- Data collection among students who dropped out: interviews and surveys, to set the most precise dropout definition, and to better understand attrition.
- Transformation and preprocessing of data using a Python Jupyter and Google Coo-laboratory, to prepare it for modeling.
- Appling the variety of ML prediction models, to find the best one for churn prediction at UNIBL.

Results:

- Model that identifies at risk students at UNIBL in the early stage of education: prior to enrollment, in enrollment week, and before their sophomore year of study.
- The list of variables with the highest impact on the churn in the current data sets by all models with passing of time.
- The first comprehensive dropout estimation for UNIBL within 12 observed years.

[4] Data for University of Sarajevo, University of East Sarajevo, University of Zenica, University of Džemal Bijedić are collected and being processed. Obtaining the data of the University of Mostar is still ongoing.

- The first research of reasons for leaving the HE in the country.
- Recommendations to the UNIBL database Center to improve the data collection process regarding the dropping out.

By implementing the Early Warning System (EWS) to support students at risk of attrition, the UNIBL has the highest benefits in this research. Beside the implementation stage, the University's management has to establish steps that follow after the identification of at risk students. In that manner, as the second largest university in the country, the UNIBL will be receiving the significant amount of funds by government, and reducing the outflow.

The policymakers in the country can benefit from using this research as a ground base for creating an effective HE attrition prevention policy. Besides the benefits of using the EWS for the student at risk of dropout, we presented the reasons for churn at UNIBL and recommendations to the university database center in order to improve the data collection process for further attrition investigation.

At individual level students benefit of not dropping HE are multiple: In the first place, there are financial benefits in the form of higher incomes and potentially better employment opportunities, followed by better health, life quality and longevity.

Practitioners and individuals can benefit from the study by using the data for further research to build upon and comparisons since our dropout definition distinguishes between students who dropped out permanently and continued their studies. Also, there is a case of imbalance data when the definition of churn was changed and their affect to model's performance and feature importance.

Having in mind that B&H's society has scientifically proven benefits from having highly educated citizens, the thesis contribute in a way to start HE dropout research in the country. Prior to this study, there were no country-level data on university dropout in B&H. With this study, we also contribute to the European network of dropout research by presenting the level and structure of dropout from HE institutions at the UNIBL, for full-time bachelor students between the 2007/08 and 2018/19 school year.

1.4 Research questions

The research questions are formulated to address the needs of the University management, policymakers, as well as practitioners and theoretical researchers regarding the attrition

of full-time undergraduate (bachelor) students at publicly funded UNIBL, B&H. To overcome the country related attrition information gap, the first research question is:

1) What is the dropout magnitude, structure and reasons for leaving the higher education at University of Banja Luka, Bosnia and Herzegovina, between 2007/08 and 2018/19 academic year?

   Without knowledge of the reasons, structure, and magnitude of the outflow, universities and policymakers cannot address the problem. Answering those research questions, we want to set a solid base to raise further dropout research at HE institutions in the whole country.
2) How well does a model perform when trained on a data set that is almost entirely binary, with missing data, and lack in socioeconomics, academics and secondary education features?

   This question addresses which model provides the best performance in predicting the HE dropout on the current dataset given the predefined metrics, and reliability in the earliest stage of education. The variables in three time points were selected and gradually added into the models (pre-enrollment, at enrollment week, and at the end of first year). The ML models: Histogram-based Gradient Boosting Classifier (HGBC), Random forest (RF), Artificial Neural Network (ANN), and Support Vector Machine (SVM) are implemented to answer on previous research question. The challenging dataset determine the costs and risk factors of classification students setting the second research question.
3) How can the explainability and interpretability of a black-box model be effectively enhanced, overcoming its inherent limitations to provide a clearer understanding of its outputs?

   Pre-hoc and post-hoc interpretability and explainability techniques were utilized to investigate differences among models trained at three different time stages. The objective was to improve the interpretability of prediction models within the EDM field, thereby enhancing the models' overall reliability

To address these research questions, the dissertation is structured as follows: After Introduction chapter, the next one brings explanation of the complexity of HE churn definition and presentation of attrition rates in Europe, narrowing down to existing research in the EDM field from around the world, through Balkan's countries to domestic

ones. The third chapter briefly explains the literature review process, followed by existing research that meets the thesis features such as source of data, reasons for churn in HE, ML model choice, achieved classification task performance, change in definition, imbalanced data, and explainability and interpretability of ML models. The research gap and contribution close the third chapter.

The fourth chapter contains description of applied methodology which is document heavy, and research design, churn estimation approach, data description, and feature engineering steps. The results chapters are chapter five, that presents attrition reasons and magnitude at UNIBL, and chapter six with evaluation of applied ML models and their explainability and interpretability by SHAP and PI. The seventh, discussion chapter brings the interpretation of dissertation results in align with research questions, and existing literature, together with research limitations and recommendations. The Conclusion chapter concludes the document by summarizing the key findings and demonstrating how the research aligns with the overall objectives of the thesis.

2 HE DROPOUT CHALLENGE AND TRENDS

Having in mind the repercussion of the HE system any country has on its economy and development, the following chapter will be discussed the challenges of dropout definition and comparability of reports, as well as the magnitude of HE dropout in Europe, ex Yugoslavian countries and, in the end, the size and the structure of dropout at UNIBL, together with reasons for churn.

2.1 HE attrition statement

In the following part of the research, we will present the definition of HE dropout by OECD, scientific researchers, the current law in an entity where UNIBL is placed, and cases in everyday practice at the University. The researchers are aware of distinguish among dropouts and not-dropouts HE students challenge due to non-existence of HE dropout standard. A lot of factors contribute to the complexity of attrition precise measurement and unsuitability of comparison: The permanency of dropout decision, different types of attrition, HE law and study regulations among different countries and HE institutions, the goals of dropout research, stakeholder's report needs, the source and the choice of the data, etc.

Dropout rates at tertiary education level by OECD, are defined as difference among total number of enrolled students (value 1) and completion rates within the official study duration time, (OECD, 2008). Choice of study duration time influence the completion and dropout rates due to occurrence of variety reasons for temporary leaving the education: illness, pregnancy, switching between programs with different study duration, study abroad, personal reasons, and part-time students usually are pulled in this category (Schnepf, 2014). Some authors highlight different types of HE outflow due to possibility of comparison like involuntary, voluntary, dropout across persistence, formal, informal and transferred students (Kehm *et al.*, 2019). Xavier indicated the issue with dropout definitions, where dropout is "*broadly defined as the student's failure to enroll for a definite number of successive semesters*" and notices the two main approaches in the literature: the synonym approach, where dropout is described as attrition, withdrawal, non-completion, churn, non-graduation, outflow, abandonment, and another approach that using rates of completion, success, graduation, retention, continuance or persistence, (Xavier and Meneses, 2020, p. 4). According to researchers from Spain, "*the most*

extended dropout definition in Spain, identifying dropout students as those having started a particular university program and decided to do not re-enroll during two subsequent academic years", (Bernardo *et al*., 2017, p. 3). Another definition defines attrition as *"the cessation of the relationship between the student and the training programme leading to a higher education degree before the degree is achieved. An event of a complex, multidimensional and systemic nature, which can be understood as cause or effect, failure or reorientation of a training process, choice or obligatory response, or as an indicator of the quality of the education system",* (Guzmán *et al*., 2021), from ALFA GUIA Project DCI-ALA/2010/94, 2013, p. 6. Other researchers count dropouts only in the second year of study, like (Modena *et al*., 2020, p. 5): *"dropouts are those enrolled as first year students in the academic year t who did not enroll at any university in the following academic year t+1*."

2.2 HE dropout in Europe

After an extensive web search of the tertiary level dropout rates in Europe, by country, in order to conduct a comparable overview, the results are sorted alphabetically and presented in Table 1. With the exception of Finland, which has a detailed annual report in English on official website, the annual dropout reports by official countries' institutions are rare or exist (but are available) in local languages (such as Austria, Spain, Hungary, and Denmark). In other cases, the HE dropout rate is estimated by researchers in published scientific papers like in Croatia, Estonia, and Denmark or not available at all like for Greece, Ukraine, Montenegro, etc. In the cases where we could not find the dropout rates in English, or such estimation does not exist, we used OECD estimation, like in the case of Russia.

Beside the lack of reports, the next challenge in presenting a comparable dropout overview in Europe was the approach of estimation. Some reports use a cross-sectional approach, while others use a longitudinal (panel data model) or some third way of estimation. For reports using panel data, there are differences in the expected time of school completion (5 or 6 years) that affect dropout rates. As the time of graduation used in dropout estimation is shorter, the dropout rates are increasing.

Another challenge of comparison is the school year used for calculations and the time within the churn is occurring. Studies use different time frames to report the churn, like during the first year of study, after the first year, or within two years of enrollment, etc.

The student's churn is not always presented at the country level. More often there is a dropout at one university or science area.

Table 1 - HE dropouts in Europe, by school year and country, in percent, as of May 2022.

Country	Dropout rate (%)	Year of estimation	Description and sources
Austria	22.1	2009/10-2012	(Bianca Thaler and Martin Unger, 2014)
Bulgaria	Low	Not specified	(European Commission. Directorate General for Education and Culture, 2015)
Belgium	12 – 8.6 6.7	2008-2018 2021-22	(Statista, 2022.)
Croatia	40	2007/08	6.6 years (Avg time for graduation), (European Commission. Directorate General for Education and Culture, 2015)
Cyprus	Low	Not specified	(European Commission. Directorate General for Education and Culture, 2015)
Czech Republic	No data		
Denmark	30	2014-15	44 percent of dropouts continue HE. Included all leavs with and without another HE degree (Troelsen and Laursen, 2014), (European Commission. Directorate General for Education and Culture, 2015)
Estonia	13.5	2009-10	Early leavers 18-24 in education system which did not receive the HE in the last four weeks of survey. (Statista, 2022)
Finland	5.9	2017-18	In detail for every year, (Hiltunen, 2018)
France	22	2008-09	The 50% of all dropouts are the first year dropout (Rajski; *et al.*; 2018), (OECD, 2012.)
Germany	28	2013-14	(Heublein, 2014)
Greece	No data		
Hungary	36-39	2020-21	(Baranyi *et al.*, 2020), (Licskay, 2021)
Ireland	13; 12; 9	2017-18–19	Tableau source (Completion Data Release March 2021)
Iceland	23.7	2011-14	
Italy	40	2013-14	(2003-06; 8.2%; 2007-10; 7.6; 2011-13; 6.7%). Dropouts are those enrolled as first year students in the academic year t who did not enroll at any university in the following academic year t+1 (Modena, et al., 2020)
Latvia	27	2012-14	University of Agriculture, after 1st year (Paura and Arhipova, 2016)
Malta	14.9 79.5	2015-16 2017-20	*(ESLU 2017)* (Torou *et al.*, 2022c), Malta College of Arts, Science and Technology
Netherlands	17.9-14.1	2008-2014	First 2 years, (CBS, 2016.)
Norway	6; 11.5*	2015-2020	Dropped within 1st year; Dropped after 1st year
Poland	37; 38	2010-2011	University of Warsaw, Math and life sciences 48; 48; Social, political, end economics 31;32; Humanities and language studies 39;39; (Zając and Komendant-Brodowska, 2019)
Portugal	No data		(Tavares, et al., 2018)
Romania	43.8	2015-2020	Percentage of students enrolled in the first year of a 3-year bachelor program dropped out from the university within 5 years. (Herţeliu, et al., 2022)
Russia	21		OECD estimation, no primary data. (OECD, 2015)
Serbia	10.5	2017-2018	Estimation, no primary data. (Stepanovic Ilic, et al., 2020)
Slovakia	42-51	2005-2010	(Stiburek, et al., 2017)

Country	Dropout rate (%)	Year of estimation	Description and sources
Slovenia	35	2014-15	(European Commission. Directorate General for Education and Culture, 2015)
Spain	19.6	2009-10	Catalonia, public universities, (Arce, et al., 2015)
Sweden	17-21	2012-2014	Engineering, Arts, Nursing, Social Work (Sofia Berlin Kolm and Fredrik Svensson, 2017)
Switzerland	28.1	2002-2008	5 years long period (Wolter et al., 2014)
Türkiye	92*	2017-2018	The total dropped after 1st year in the country. Cross section.
United Kingdom	6.3*	2017-18	Computer sciences dropout rate at 9.8%, with medicine, dentistry and veterinary science 1.5%. ("Universities With Highest and Lowest Dropout Rates")

*Unofficial source (web news, or portals).
Source: A literature review.

The countries with the lowest HE attrition rates are Bulgaria, Belgium, Cyprus, Finland, Norway, the UK, and Serbia. Germany, Hungary, Italy, Poland, Romania, Slovakia, and Switzerland have the highest dropout rates. The Türkiye is standing out due to extremely high rate of churn.

Of the Balkan countries, in the last ten years, Slovenia has only one HE dropout report (2014-15), which is 35 percent. Croatia has a report from 2007 (40 percent). Serbia has an estimate for 2017/18 school year (10.5 percent) based on a mix of documents with no primary data. Bulgaria and Cyprus have no report but only claim that the dropout rate in HE is low, in an EU report.

Comparability of our dropout data even with border countries is challenging because, for example, Croatia uses 5 years long period of time to distinguish among churned students at the country level (study of 2007/08 school year), while for estimation of UNIBL outflow we used at least 6 years of time. Montenegro and Kosovo do not have HE dropout reported, while Serbia, as the following border country, doesn't use primary data in research, which may lead to underestimation of attrition.

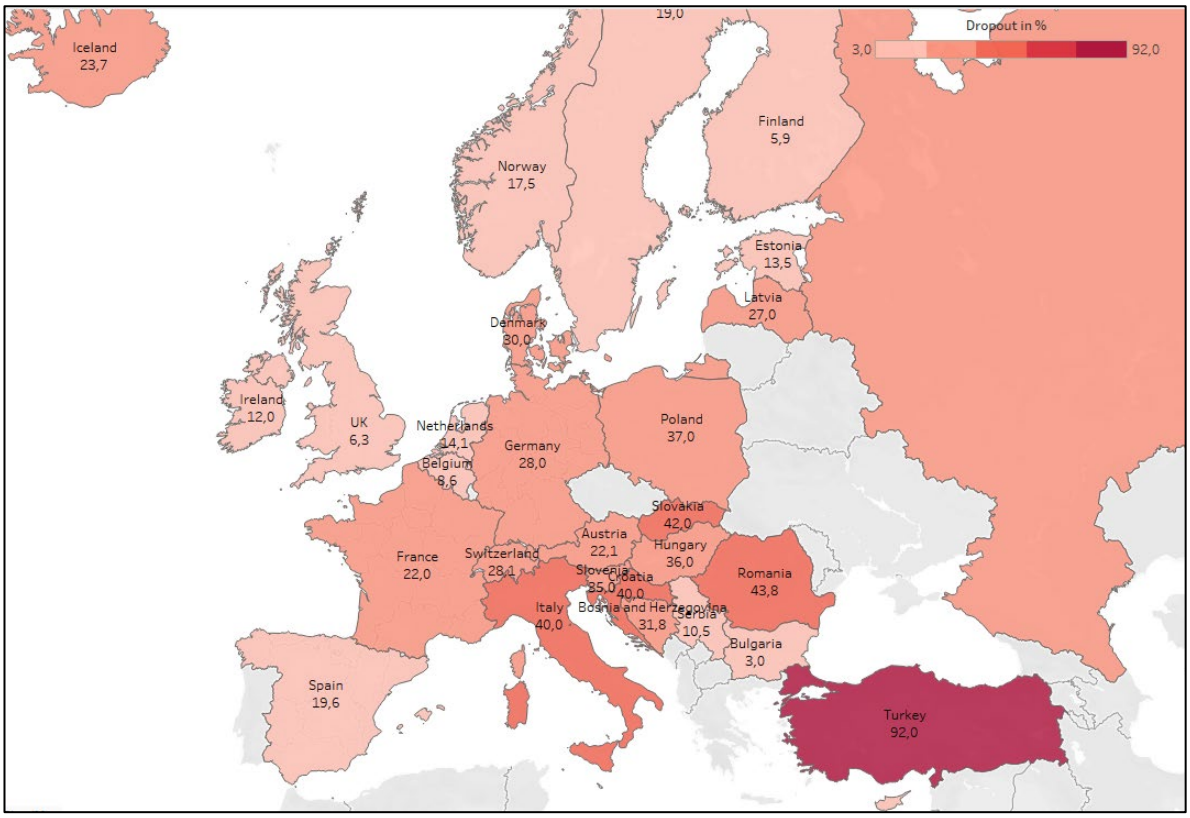


Figure 2 – Bosnia and Herzegovina mapped at HE dropout in Europe.

Source: Author's contribution.

Investigation of collinearity between dropout rates and education system, HE fees, governmental support and possible penal of HEI with high churn in European countries is interesting but overcome the scope of this research.

2.3 Bosnia and Herzegovina's HE environment

The B&H HE environment is constituted by private and publicly funded universities, higher schools, religious universities, and official state institutions like ministries and the Agency of education. The complex political and administrative structure overflowed into the HE sector.

One of the original goals of this research (to estimate and report the country-level publicly funded university's dropout rate) shrank to UNIBL, since upon today we are in the process of collecting the data for eight public universities in the country. The main reason of such long collecting data process is that the country still does not have a united (central) database of HEI. Each publicly funded university has its own database, which architecture and implementation differ from the others. Some universities do not have a central database, yet each faculty member has their own central database (University of Mostar).

The data which are collected from students in hard copy (written on pre-defined forms) at the beginning of each semester were placed in the database by administrative staff, often filling only the required fields. Differences in data collection forms from students are also noticed by the country's two entities. There are cases where the central database is not fully implemented even at the faculty level, and the majority of the database is on the paper form.

To grasp the complexity and challenges associated with conducting research and disseminating findings in B&H, particularly in modeling tertiary education dropout using primary data, the following sections will provide a brief overview of the higher education environment in the country.

According to the World Bank, B&H is an upper middle-income country with a population of 3.5 million (Census, 2013). Even such a relatively small country is divided into two entities and the special district of Brčko (town). Broad division in the country's administration caused the complex structure of HE jurisdiction of official institutions.

Both entities and the Brčko District have their own institutions of education:

- The Ministry of Scientific and Technological Development, Higher Education and Information Society of the Republic of Srpska, for entity RS.
- Federal Ministry of Education and Science, for entity Federation of B&H. The Ministry has a coordinating role because the Federation of B&H entity is divided into 10 cantons, with each canton having its own Ministry of Education.
- Department of Education of the Government of Brčko District of B&H, for District Brčko.

Besides that, there is the Agency for Development of HE and Quality Insurance and the Education Sector of the Ministry of Civil Affairs of B&H. The Agency's role is to define and monitor the quality standards of HE institutions in the whole country regarding the accreditation process, establishment and closing, development, and implementation of those standards. The Agency does not collect any dropout HE data, like there are cases in the EU countries. The Education Sector mainly has a coordination role, representing the country in the domain of exchange of activities and data with domestic and international institutions responsible for the field of education.

In addition to the Brčko District, two municipalities, Žepče and Usora, have their own educational permits, which they obtained in accordance with the decision of the

Constitutional Court of the FB&H in 2010. It means they can create their own educational program independently, but in those two municipalities there are no HEI.

Eight state-funded universities are accredited with the following number of enrolled students, presented in Table 2.

Table 2 – Public funded universities in Bosnia and Herzegovina, by the number of students.

University	Number of students	Share of students (in %)*	Year of source:
University of Sarajevo	26,233	33	2018
University of Banja Luka	15,000	19	2022
University of Mostar	12,000	15	2021
University of Tuzla	8,318	11	2018
University of East Sarajevo	8,049	10	2018
University of Zenica	3,173	4	2018
University Džemal Bijedić	3,030	4	2018
University of Bihać	2,952	4	2018

*Share of all students enrolled into those eight public universities, without population of students enrolled into other HEI.
Source: Official university websites and Google short feedback.

The largest university, with the longest tradition in the country, is the University of Sarajevo, with an average share of 33 percent of enrolled students. The UNIBL is the second largest university in the country, with an average share of 19 percent of enrolled students.

Research and development spending as a percentage of GDP is trending downward from 2013 to 2021, from 0.32 to 0.19 percent of GDP (Agency for Statistics of Bosnia and Herzegovina, 2021).[5] B&H does not have dropout reports for private or state-funded institutions of HE. Dropout reports are available for primary and secondary schools (high schools) in local languages.

The tuition fees at UNIBL are relatively low, a bit different from one faculty to another, and did not change much over time. Fees are very favorably, especially for domestic citizens, full-time students, the university scholarship holders (Table 3). All students who are co-financed and self-financed full time students, may gain of switching to the category of university scholarship holder if they passed all exams in given school year. In that way the government stimulates students to finish studies on time, by minimal cost of 43 euros per year.

[5] EU27 average in 2021: 2.27 percent of GDP, where the highest percent goes for Sweeden 3.35, and the lowest for Romania, 0.48.

Table 3 – The tuition fee at Faculty of Economics, bachelor study, UNIBL, 2007-2023.

Tuition fee for two semesters (one school year) in EUR	Students study's finance status
43	University scholarship holder, full time
225	Co-finance (part of scholarship), full time
614	Self-finance, full time
767	Part time student
1023	Foreigner, full time

Source: The tuition fees at UNIBL, official document at University's website.

2.4 The use of ML in HEI dropout prediction

An important instrument for preventing HE attrition is by predicting the dropout before it occurs or before a student even knows that he/she will drop out. The researchers went far beyond theoretical and statistical models, using a variety of data sources, and implementing the ML models to predict churn at time, i.e. the earliest as possible. With the increase of computational power, the warning systems for at-risk students gain an important role with other tools that are at disposition to the management of HE institutions.

2.4.1 The development of the educational data mining field

The first studies dealing with student retention and dropout in HE were social in nature and mainly theoretical, with the aim of identifying and better understanding the reasons for attrition. The most frequently cited theoretical model is the book *Leaving College: Rethinking the Causes and Cures of Student Attrition* (Tinto, 1987), which forms the basis for other theoretical and statistical models. Tinto contends that student retention is influenced by pre-entry characteristics: a) family background, b) skills and abilities, and c) prior schooling. Pre-entry elements related to academic, social, and external factors influence the decision to drop out. The second well-known model considers the relationship between retention and dropout variables (Bean and Metzner, 1985) and relies mainly on Tinto's model, introducing endogenous (commitment, institutional fit, and grades) and exogenous variables (academic, social, and environmental factors).

The next generation of research in this area focused on explaining dropouts using statistical models. The most commonly used models were logistic regression, discriminant analysis, and structural equation modeling with the goal of identifying essential variables for university student success, retention, academic outcomes, achievement, and dropout risk. The conclusions were that traditional admission criteria

and learning style have low or limited predictive power for grade point average (GPA), student satisfaction, and attrition (Thammasiri *et al.*, 2013).

Things accelerated after 2006 with the research of Hinton, LeCun, and Bengio in machine learning algorithms (Hinton *et al.,* 2006), (Bengio and LeCun, 2007), and the ground base for a new field, the Education Data Mining (EDM), which was developed in the consecutive years. This and the increase of computing power and Internet of Things enabled the use of artificial intelligence (AI), that is, machine learning and deep learning techniques to collect and process the increasing amount of data. The complexity of EDM stems from big data coming from various sources such as e-learning platforms, student cards, school information system databases, and student surveys (self-assessment, depression, satisfaction, motivation, etc.). Studies focus on identifying learning styles like slow learners and other student problems, such as depression. Books and course recommendations are written based on machine learning models, web data is analyzed, measurement and prediction of student performance, grades, graduation success, and time, engagement, and enrollment are considered.

The current step in the field development is toward interpreting the ML models by explainable (visual) techniques. There is a need to explain how and why "the black box" models make "decisions." The field is in its rising part, and we expect its further expansion. During the literature review, we identified nine papers that tackles the HE dropout using explainable ML implemented in Python.

2.4.2 Overview of domestic research in EDM

We identified seven papers in total referring to EDM in B&H from 2012 to 2022 (Table A1 in the Appendix), two of which addressed the dataset of high school students to classify student performance (Osmanbegović *et al.,* 2014) and predict high school student affiliation (Osmanbegović *et al*., 2013). Other studies targeted university students to predict success in a first-year course at the Faculty of Economics using Bayes, neural networks, and decision trees (Osmanbegovic and Suljic, 2012), success in the course at the Faculty of Pedagogy (Simeunović and Preradović, 2014) using if-then rules; recognition of 4 learning types in the context of predicting success at the Faculty of Economics (Kacapor and Lagumdžija, 2020), and prediction of final grade (Gašpar *et al*., 2015). Reported measures of accuracy ranged from 50 to 76 percent. There is no research on dropouts at HE level in B&H up to July 2024.

Overview of Ex Yugoslavians countries' EDM in HE research: There are three EDM university research in Croatia: prediction of passing or failing courses (Kovač and Oreški, 2018), prediction of failing the course Programming 1 using Moodle data (Sisovic *et al.*, 2016), and dropouts in the junior years of HE using logistic regression, decision trees, and neural network (Jadrić *et al*., 2010, p. 35) identifying the high school as one of the most critical factors influencing in the model. The authors assumed that "*students drop out of their studies by choice after the first year, while students drop out after the second year mainly because of failing exams*." The study showed that women were less likely to drop out of their studies than men. EDM research at three universities in Serbia seeks to predict the average study grade (Išljamović, 2013), final course outcome (Jovanovic *et al.*, 2012), to improve the web-based reports of an e-learning system (Blagojević and Micić 2013), and even to build the EWS that will predict students at risk on time, but in secondary education (Jovanović *et al.*, 2017).

Montenegro and North Macedonia have only early leavers' statistics available, but we can't rely on it since it presents the share of youth aged 18-24 which did not receive any education or training in the last four weeks prior to survey they were part of. Republic of Kosovo has one research with dropout reported in 2022, for Faculty of Electrical and Computer Engineering, University of Prishtina, which are ranged 53-78 percent, (Kabashi *et al.*, 2022).

2.4.3 Churn prediction at HE institutions

In this section will be presented cases of EWS utilization at HEI in the world, with their prime goals, input data, employed ML models, evaluation metrics and performances, as well as their observed outcomes. There are only a few rare cases were the HEI in the world have implemented EWS, while there is far more examples of building and testing EWS that demonstrated good or excellent HE churn prediction, but remained in their experimental phase.

The example of implemented EWS based on ML classification method, that uses unique input data is found at the Institute of Information Engineering, Hangzhou University, China, (Wang *et al.*, 2018). The main goal of this EWS implementation was to reduce the number of dropout at HEI by on time identification and informing the students, parents and teachers. The input data covered time frame between 2009 and 2016 school year: students score, attendance, personal data, entrance of the dormitory and library, as well as library borrowing data. After cleaning and preprocessing they had 1,712 students in

total. Principal component analysis has been identified the key factors of dropping out: arithmetic mean, course credits obtained, the average score, and the library books lending, which achieved the cumulative contribution rate of 86 percent. The ML classification was done in Python, using the 70 percent data for model training, by employing decision tree (ID3 algorithm), neural network, and Naïve Bayesian model. The evaluation metric used on 20 times tests running was average accuracy: accuracy of DT was 84-87, NN decreasing 77-75, and Naïve Bayesian 85-88 percent shows the best performance.

(Gutiérrez *et al*., 2020) implemented The Learning Analytics Dashboard for academic advisors (LADA) which was used at two universities, in Europe and Latin America, as a tool which helps academic advisors to make better decisions to assist students at risk of dropping out the course or degree. There were two groups of advisors: the experienced experts and those who are not trained to provide academic advising. The second group, called laymen, was introduced the LADA, and compared to the experts group. LADA uses student grades, list of available courses, the courses booked by a student and the number of course credits, as input data. Data from previous cohorts are base for prediction of current students. The model implementation was done in Python (scikit-learn library), using Adaptive Multilevel Clustering technique, while the client side was implemented as Web application. The tool was positively evaluated by users, where the most valuable was ability to analyze huge amount of data, exploring the different scenarios in a short amount of time.

Another example of utilization of ML models in predicting HE dropout was found at Vrije Universiteit, Amsterdam, (Plak *et al*., 2022). In the school year 2016/17 was implemented EWS to identify freshmen who are at risk of failing the course and dropping out of study in their freshmen year. The EWS is made by selecting the best performing model between logistic model, additive logistic model (showed the best performance), SVM, and RF model, which all were trained on 85 percent sample data, by 5-fold cross-validation method, and evaluated by mean absolute error for eight time periods: between the beginning of the first year of study and before enroll into the second study year, where mean absolute error was ranged from 36 percent at the beginning of the first year, up to 6 percent during the summer break. Models were fed by demographic and student progress data of the freshman of two previous freshman's years. Data were used at weekly basis in form of percent of dropout risk by student counselors to mitigate the risk of failing the course and dropout risk. The sample included 758 freshmen bachelor students, which received the student counseling. Even there are important benefits of early identification

of at risk students, the study shows no significant difference between student counseling assisted with EWS information and counseling as usual, that is there were no changes in dropout risk or increase in obtained ECTS course credits, between those two samples.

The most cited example of EWS in HE is at Purdue University, Indiana, USA, (Arnold and Pistilli 2012), named Course Signals, since 2007. This EWS is not based on ML, but on mining and weighting the huge amount of data (that includes students' performance, effort, prior academic history, and personal characteristics) by statistical models. In the beginning the main goal was to ensure reaching the maximum course potential for each student, but grew into a tool that has been embraced by students and teachers, plus it was demonstrated high performance: decrease in course dropout and increase in high grades, comparing the courses that did not used the Course Signals. Although the Course Signals at Purdue University does not grounded at ML, it deserves to be mentioned here as example of effort that HEI make to help students to gateway courses, which leads to dropout risk mitigation.

The same goes for University of Michigan, who established personalized learning solution (ECoach) to help students to pass STEM courses, based on learning analytics (Wright *et al*., 2014). Meanwhile five other universities in USA and Singapore adopted the ECoach. An example that is also worth mentioning is Student Explorer, designed to shortened the time between identification student in risk of course dropping out and advisor's intervention, at University of Maryland-Baltimore County and University of California, Berkley, (Krumm *et al*., 2014).

We have observed rare HEI instances of implementing a system to assist students in overcoming challenges related to the heightened risk of dropping out of their courses and studies. Examples of ML models implementation through EWS are even rarer. Most cases involve constructing experimental models, such as the one presented in this thesis.

3 LITERATURE REVIEW

The main purpose of the literature review is to be used for constructing, validating, and setting the ground base of the research domain. It is a process which result is pointing and filling the research gap which govern contribution of the research. We conducted the literature review in a way to find what other researchers found regarding the HEI churn and how far the B&H went in this domain.

The section starts with brief description of literature review methodology and searching techniques applied, then continues by presenting what other researchers in the EDM field discovered related to our research questions, and ends with description of quantitative models and metrics used for churn classification, as well the research gap.

3.1 Methodology of literature review

The following data sources are identified by conducting the literature review: Google Scholar (as a broad, unspecific database), the Institute of Educational Science (ERIC) as a specialized database related to the education field, the Institute of Electrical and Electronics Engineers (IEEE) database for best fitting articles in machine learning domain and Elsevier database. The specific conferences and journals tightly related to machine learning in the educational data mining field are The Journal of Educational Data Mining, and the International Conference on Educational Data Mining. Both sources are indexed in Scopus and Web of science until the moment of checking.[6]

Description of sources:

Database: Google Scholar[7]. Searching techniques include synonyms, boolean operator AND to narrow the search, and OR for including the synonyms, truncation method (*), use of brackets, and excluding sign (not) "-". In the searching process:

- Keywords "*student churn prediction",* which generated 22.700 results
- "student" AND "*churn prediction"* – 21.600 results
- We have narrowed the search by adding more synonyms and excluding terms using dashes: "churn" OR "dropout" OR "attrition" AND student* AND "prediction model" AND "higher education" and to exclude -e-learning, -bank, -

[6] June 27, 2022.
[7] April 29, 2019.

telecom. Filtering search for the results from 2009 until today, in English – 832 results were the first review base.

From 832 retrieved documents in the first phase, it was extracted 90 articles matched the title keywords criteria. In the next stage, from 90 articles, we pulled 58 articles by keywords in abstract content. In the third phase, from 58 articles, 16 articles of high interest and 15 articles related to this research are identified (predicting student retention, success, grades etc).

Institute of Educational Science (ERIC) contains a collection of 1.160 journals related to education from high school to post-graduate level. The searching technique was the same as in the previous case, respected the database searching rules – including brackets: ("drop out" OR "churn" OR "attrition" OR "retention") AND student* AND ("tree" OR "neural network" OR "SVM" OR "random forest"). The result was 29 retrieved documents (3 of them matching the Google Academic search) and an additional 4 generated in this search.

The Institute of Electrical and Electronics Engineers (IEEE) database: searching for the keywords (using truncation method combined with Boolean) in the title: student* AND predict*, which resulted in 280 retrieved documents. The second step was filtering the time period (last ten years), which narrowed the search to 262 documents. The third step included adding the index terms: pattern classification, decision tree, random forest, neural network, SVM, big data, regression analysis, Bayes method, data analysis and pattern clustering which narrowed the search to 149 documents. The last step – identification 6 papers highly related to the research (3 were the results retrieved previous in Google Scholar search), and 3 articles more were identified. Results of the search are presented in the Appendix, Table A2.

To update our literature review, another search of the Corvinus University Library database was conducted in June 2022 with the following parameters:

- To cover all terms and synonyms: dropout OR attrition OR at risk OR churn
- To target only higher education synonyms (AND): university OR college OR higher education OR tertiary education
- To specify the desired domain (AND): machine learning
- To exclude (NOT): telecom OR bank OR cancer OR eLearning OR surgery OR brain OR injury OR disease OR bacteria* OR virus OR psychiatric OR water OR bleeding OR diabetes OR toxic*
- To narrow the search (NOT): predict success OR high school

– From 2020 up to 2023.

This search generated 472 results, where it is noticeable that results are somewhere doubled.

3.2 Previous research of ML modeling the tertiary level education dropout

While conducting the literature review we focused on the studies that use the same models, the same data sources and have the same goal, to classify dropouts in tertiary education, as we do. The following chapter summarize the most important findings of other researchers which are corresponding to the scope of our study contribution.

The following papers deals with tertiary, i.e. HE attrition prediction, mostly using the first-year students as target group, due to fact that the highest amount of HE churn occurs at freshmen year. The presented studies use data at the time of enroll and first university year, while only few of them use questionnaires and data of sophomore and following years for train of ML.

The main features of dropout identified by ML models where university database was the main data source:

The **financial reason and grade point average** (GPA) are commonly presented in papers coming from United States of America (USA), as variables of strongest impact to the prediction model. The GPA, family income, age, scholarship, and a bank loan are the most significant factors of dropout at Tennessee, USA, among freshmen, while the less significant variables are family size, having or not dependents, and living in campus, (Baghernejad, 2016). The different types of financial support like tuition scholarship, received aid, student's loan, with ethnicity and major declared are the strongest predictors of dropout at USA, for first year students, using large university database from 2005-2011, (Thammasiri *et al.*, 2014). Financial support and HE performance were the most important predictors of dropout in USA, in a large university dataset, (Delen, 2010). Another large university dataset in USA, which analyzed the attrition of the freshmen found that GPA in mathematics, English, chemistry, and psychology are the strongest churn predictors, (Aulck *et al*., 2016).

The strongest predictors of HE outflow may be connected with **family and carrier plans**: For the whole university students, the GPA during HE and ECTS score ware the best predictors of will and when the dropout happened, followed by factors like the number of family members, number of family members at college, county of residency, age at the time of entrance the university, ethnicity and high school GPA, (Ameri, 2015). Plans of

starting family, marriage, or pregnancy, together with carrier change, university change and profession are the best prediction variables of dropout in Mexico, while the number of supported children, gender and age are variables with the less predictive power to the model, (Pérez *et al*., 2019).

Some studies shows the **importance of place of residence**, like in Taiwan, India and Italy: The second-semester grade, together with study major (or field), place of residency and loan status were the main features on the first-year student dropout and retention in Taiwan (Weng Fu Mei, 2010). For the first year students at technical university in India, the best predictors were high school GPA and living location (Pal, 2012). Among the most important predictors is the distance from the university, while the gender and scholarships are the less important in dropout prediction, in Italy, using survey among 810 freshmen in a health care profession, (Siri, 2015).

Sometimes **race, or ethnicity**, are, among others, presented as strongest dropout predictors. For the first years for Philippines students, the high school GPA, admissions test score, race, or ethnicity are the strongest predictors of dropout, while gender or attendance of private or public school was the less important, (Patacsil, 2020). Another Philippines research identified the first semester grade as the strongest prediction of churn at the end of freshmen year, and the lowest prediction power had religion, (Barbosa, 2017). Gender, financial condition, age, ethnicity, education, work status, disability and study environment are the strongest dropout predictors, for IS students in Bangladesh, (Mustafa *et al*., 2012).

The gender is sometimes important attrition predictor, and we presented both. The studies which claim opposite, and for example the (Aguiar, 2015): The gender and the GPA in the last grade of high school are the most important predictors of freshmen engineering student's dropout.

There are also studies which considered **nonstandard attributes**, like (Yang, 2021): The quality, quantity and mobility of academic partners, the seating position in classroom, dormitory study atmosphere, the addiction level of video games, the English scores of the college entrance examination and the degree of truancy are the best predictors for academic risk, while the less contribution to the model provides living in campus or not, work-study, lipstick addiction, student leader or not, number of lovers, and smoking status.

The churn classification ML models performances achieved by other researchers in existing literature:

Analyzing the conducted HE dropout ML modeling literature, that correspond to the ML models employed in our research, we witnessed of the high model performances, which are often related to DT algorithms (Table 4).

Table 4 – ML model performances used to predict dropout at HE institutions.

Model	Performance (in %)	Measure	Sample size	Source
DT LR NB	94 92 87	Accuracy	802	(Perez *et al.*, 2018)
LR kNN DT – CART Ensemble (AdaBoost)	92.2 80.50 89.7	10-fold cross validation, confusion matrix	7,936	(Barbosa, 2017)
DT – J48	91.2	Accuracy	170	(Pérez *et al*., 2019)
DT – ID 3 DT – CART	90.9 86.0	Accuracy	1,650	(Pal, 2012)
RF NN DT – J48 NB	86.14 81.67 80.18 63.4	Accuracy (reported here, but it was used recall, and precision, too).	671	(Rodríguez-Maya *et al*., 2017)
NN	85	Accuracy	300	(Ahmad and Shahzadi, 2018)
NN	76-84	Accuracy	810	(Siri, 2015)
NN DT LR	81.19 78.25 74.33	Accuracy	25,224	(Delen, 2011)
SVM LR	80.1 72.5	Accuracy	2,353	(Weng Fu Mei, 2010)
DT Forest tree DT – J48 Ensemble models AdaBoost	 70.49 63.56	Accuracy (reported here, but it was used recall, and precision, too).	2,401	(Patacsil, 2020)
LR NN DT	62; 57 56; 54 45; - -	Recall; precision	1,453	(Agnihotri and Ott, 2014)
Regression LR, SVM Support Vector Regression Adaptive boosting DT	Cox – the best performance, DT – the worst	10-fold cross validation, accuracy, F-measure, AUC	11,121	(Ameri, 2015)

Source: Authors contribution, based on the literature review.

The highest accuracy was found in the papers of (Perez *et al*., 2018), by DT of 0.94, and (Barbosa, 2017), by LR of 0.92. It seems that sample size does not have crucial role in achieving high prediction performances, rather the data quality and reach in feature data build the strong model. It also noticeable that there is only a few papers that uses datasets larger than 5,000 students. Majority of authors report accuracy, while some report sensitivity (recall) and precision.

The impact to the model by high dimensionality of features, missing values, and imbalanced data:

The dimensionality of data set play important role in achieving the targeted performance, but in the next example, with the time component, **adding more features improved the prediction model:** Chai and Gibson experimented with four different time models to predict first year attrition at Curtin University, Australia, gradually adding variables to the ML models: pre-enrolled model that uses 17 variables, enroll model adds 5 more, in-semester model uses 24 variables, and finally end of semester model uses 28 variables in total, which ware encoded into 164 selected features. Dataset of 23,291 students from 2011-2013 is imbalanced, 83:17, where minority class are dropouts at the end of freshmen year, no missing data. Models LR, DT (CART) and RF ware evaluated only by precision and recall, using imbalanced data. Results proofed that adding more variables to the model improves prediction measures, significantly. The highest precision had RF (71 percent), and the highest recall had DT (37 percent), (Chai and Gibson, 2015).

Generally approach to imbalance data set definition means that data are imbalanced when one of the target classes is more represented than others. In this particular case it means that we have more students which are not dropouts than those who are. This may affect the model in a way that scores a high accuracy by classifying the students who are not dropouts. According to Li and Sun, (Li and Sun, 2012), the dataset is considered as imbalanced when the representation of minority class is less than 35 percent of the whole dataset.

While there are many approaches to data imbalance problem, (Barros u. a. 2019) classify them as data level, and model level techniques. Authors defined the three major approaches that are represented in the literature to the "Accuracy Paradox", the case when high ML model accuracy, at the same time, does not mean the high model's quality.

Using 670 Mexican's high school pupils' data set of 77 variables, with imbalance ratio of 91:9 for pass and fail the preparatory course for the university, authors have proved that **feature reduction and balanced techniques may improve** the model performance. They employed the most popular five DT and five rule induction algorithms (if-then) in WEKA, which were evaluated using overall accuracy, TP rate (sensitivity, or recall), TN rate (specificity) and G-mean. Models were applied on four different datasets: 1) prediction using all 77 variables, 2) prediction using selected 15 variables, 3) prediction using SMOTE balancing on 15 variables, and 4) prediction using cost-sensitive classification on 15 variables, as another approach to balancing. The best overall model

performances come from dataset number 3, i.e., **prediction using SMOTE** balancing approach, keeping in mind prediction of failing the course, (Márquez-Vera *et al*., 2013). With the time new balance techniques are developed and tested: Authors (Waheed *et al*., 2021) introduce the additional resampling approaches, **adversarial networks**, to find the best one that pushes up the performance of ML model at imbalanced data set. Data set is obtained from 32,593 UK students, 2013-2014, and multiclass outcome is transformed in two classes: pass of fail the course. The evaluated measures are precision, recall, accuracy, AUC, and F1 score. Results show that the highest performances were achieved by GAN resampling approach by both classifier, RF and NN.

Although we have binary prediction model, the researchers suggest that balance techniques improve ML models performances on multi-class prediction, as well, (Ghorbani and Ghousi, 2020).

Different studies which use different datasets show opposite findings regarding the best balance model. For example, (Hasan *et al*., 2015) stress that ensemble models as approach to the balance problem (AdaBoost and Bagging) improve the ML model performances. At low dimensionality, no missing values data set, author Patacsil seeks to predict will student dropout or not (graduated) during the first study year at the university due to high dropout in freshmen year (more than 80 percent) by employing the ML models to enrollment data. **Ensemble models** (vote, bagging in combination with DT, RF, J-48), **which ware applied to balance the data, made precision and accuracy worse, but improved the recall** to 78.69. AdaBoost as ensemble algorithm had poorest performances than ensemble models (Patacsil, 2020).

The next research use only overall accuracy and precision as imbalance measures: Using high dimensionally data set of 39 variables, and 16,606 records of data from 2004-2008, no missing values, with imbalance ratio of 78:22, to predict university dropout at freshmen year, author compared results of imbalanced, over sample balanced and ensemble ML models, using overall accuracy and per-class minor (precision) class (not retained) accuracy. The per-class minor accuracy with imbalanced data employing NN, DT, SVM, and LR was less than 50 percent. **The balanced data set showed high increase of** per-class accuracy (**precision**) up to 86 percent, and ensemble models followed with 75 percent, (Delen, 2010).

Alternative metrics which are effective in imbalanced data:

We saw by now that solving the imbalance problem, overall accuracy, recall, and precision were used for evaluation of the ML models. The consideration of the alternative

metrics in this case is important because in high imbalanced data, model can observe the minor class as noise, so the model far better generates prediction on dominant class, where measures like overall accuracy, and ROC AUC, when False Positive (FP) is relatively low (small sample) continue to show high prediction accuracy. This why some papers suggest that we must add measures which highlight the FP error:

Lee and Chung, (Lee and Chung, 2019) accent the difference between use of ROC curve and **precision recall (PR) curve** in case of high imbalanced data (Figure 3). The authors compared ROC curves and PR curves, to determine better metric when the imbalance ratio is 98:2, at a large, rich in features, high school data set, using more than 165 thousand of Korean's pupils, for train the dropout model. PR curve, unlike ROC curve, for x-axis use recall (sensitivity), or true positive rate, and for y-axis use precision. The RF and boosted DT were used as classifiers, and the SMOTE for under and over sampling was used, combined with RF and boosted DT, as well. All four models achieved high ROC, more than 0.98, which may imply that we can use whatever model we choose. But only boosted DT achieved 0.898 value at PR curve, while others had far lowest performance.

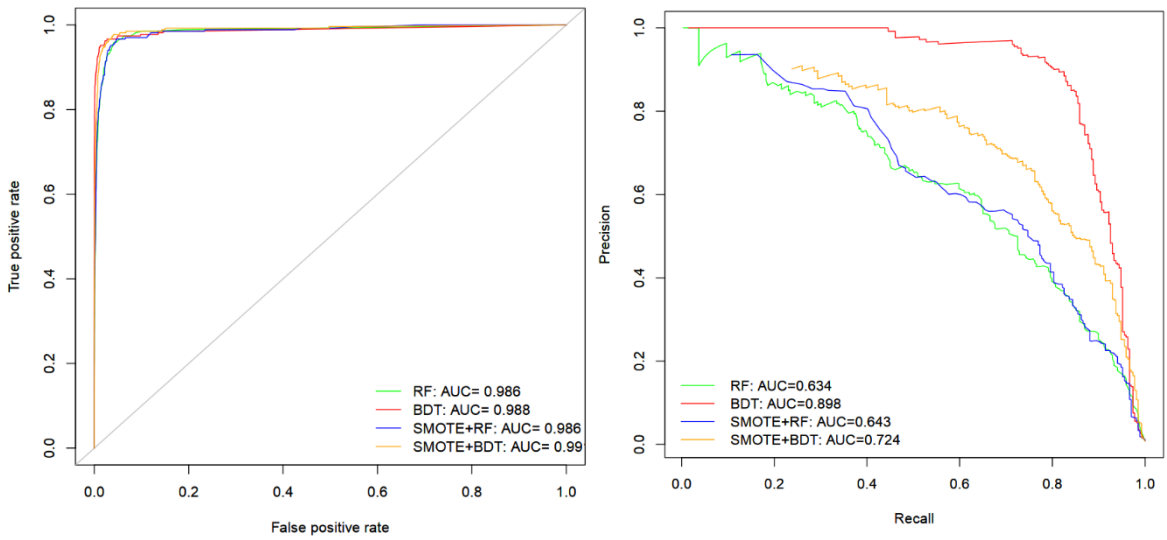


Figure 3 – Compare of ROC curve and PR curve at imbalanced dataset.

Source: (Lee und Chung 2019), page 10/14.

The high dimensionality university dataset[8] of 21,654 records, and 34 variables, imbalanced 79:21, where the minor class represents dropouts at the end of first year of university study within the 2005-2011 school years, was used to build ML models which

[8] Data set is from several disjoint databases.

predict attrition in freshmen year. *Missing values* were represented within two variables in amount of 4 percent and those records were removed. Employed ML models LR, DT, NN, and SVM with imbalanced data ware combined with random over-sampling, under-sampling, and synthetic minority over-sampling technique (SMOTE) due to imbalanced dataset. To ensure comparability 10-fold cross validation was used, sensitivity (recall), specificity, F-measure, and accuracy as standard measures, together with alternative metrics calculated from confusion matrix for imbalanced datasets: **precision±, FP-rate, G-mean, and correlation coefficient**:

$$precision^{+} = \frac{TP}{TP + FP}$$

$$precision^{-} = \frac{TN}{TN + FN}$$

$$FP - Rate = \frac{FP}{FP + TN}$$

$$G - mean = \sqrt{sensitive * specificity}$$

$$CC = \frac{TP * TN - FN * FP}{\sqrt{(TP + FN) + (TN + FP) + (TP + FP) + (TN + FN)}}$$

Introducing balanced data was significantly improved the specificity and precision-, while recall was slightly increased from 87.8 to 88.5 percent. Only SMOTE improved accuracy, from 86.4 with imbalanced data up to 90.2 percent. The best overall performance had a SVM with SMOTE, (Thammasiri *et al*., 2014).

There is a valuable contribution of (Solis *et al*., 2018) who highlighted the importance of **impact that attrition definition has to prediction model**, and testing the new imbalance data metric. They tested 4 different types of attrition definitions to eliminate time noise of previous semesters and active students who did not graduated at the time of observation, at the Instituto Tecnologico de Costa Rica, within 14 semesters, 2011-2016 school year. Those four definitions of dropout had different imbalance ratios, but authors did not use any balance techniques. The sample size decreased from 80,527 records (not students) for the first, the most comprehensive definition, end ended with 7,936 records for the last, and the least noise represented definition, with imbalanced ratio 44:56 in favor of dropout students. For the evaluation purpose they implemented **Kappa**, **sensitivity, specificity**, the probability of correctly detecting dropouts (positive) and probability of correctly detect non-dropouts (negative) rates for RF, NN, SVM, and LR. Authors accent that the most important measures are sensitivity (recall) and positive

probability. By far the best performance had the least noise represented definition, where the RF, with highest Kappa score (85 percent) which indicated the highest sensitivity (recall) (93 percent) and specificity (94 percent) scores were chosen as the best model. The conclusion is that when "*train the dropout prediction algorithm, it is convenient to exclude active students, who may add noise because it is not known beforehand if they will dropout or graduate in the future. In essence, the problem is that they may have a dropout pattern, but they have not been classified as such*." (Solis *et al*., 2018, p. 5)

Utilization of explainable ML in HE attrition modeling:

We found nine papers in total that use explainable AI or interpretable ML techniques in the domain of EDM, especially in classification churned students, by June 2024. Two of them are focused on secondary education, namely in Chile (Rodríguez *et al*., 2023), and Japaniese online courses school (Katsuragi and Tanaka, 2022). The remained seven papers focusesd on tertiary education.

The first authros who opened the door to EAI in modeling the tertiary student's churn was from Hungary. (Baranyi *et al*., 2020) used SHapley Additive exPlanations (SHAP) library in Python to explan the model prediction of will student drop out or not at the Budapes University of Technology and Economics, among 8.319 students, since 2013 up to 2019 school year. Authors used SHAP for identificaton of important feature at model level and their permutation importance.

One of the authors who utilized the explainable tool to elaborate the most influental variable to university dropout was (Yang, 2021). Except the displaying the most significant variables of dropout at one private university in China, the SHAP showed dependancy between the quality of academic partners, seating position in classroom, dormitory study athmosphere and dropout risk. At the end, Yang demonstranted the individual level of dropout risk explanation by showing the marginal contribution by single variable for a particular case.

(Dass *et al*., 2021) uses data from Open edX platform for self-paced math course of Arizona State University, collected between 2016 and 2020 to predict course dropout date. At the end of reserach author use SHAP to presented feature importance for the whole model and two cases at individual level.

(Nagy and Molontay, 2023) in their comprehensive work presented and compared XAI tools for a black box ML models of dropout at Budapest University of Technology and Economics at two levels: globaly and locally (individualy) visualization and explanation. They compared LIME, SHAP, permutation importance (PI), and partial dependance plots

(PDP) to explan the most important factors of dropping out in general and in four individual cases. Results were evaluated by users at the University and showed positive feedback.

(Delen *et al*., 2023) employed SHAP to reveal the most influental variables that affect student dropout at the end of the first study year at one mid-western university in USA, using the data since 2009 up to 2018, with 39.470 freshmen, and more than 60 variables from several combined data sources. After training the NN and usage of SHAP for the global and individual dropout explanation, author stressed out the importance of understanding the results of the black box model for future research and model improvemnets.

(Kim *et al*., 2023) uses SHAP a bit different than others: at the beginning of building the student dropout prediction system at Gyeongsang National University in South Korea for extraction the most influential variables. The SHAP combined with permutation importance was able to detect 27 essential of 40 variables in total that were at desposal for model building.

The last research that was found in this domain was by authors from Mexico, (Alcauter *et al*., 2023). They build several ML models to classify student at risk of leaving the STEM faculties at one university in the country, and among them the RF showed the highest accuracy. The LIME was applied in two particular cases to evaluate student's dropout estimation in detail. In this way authors stressed out the benefits of LIME in better understanding variables that influence risk in two direction, for taking the proactive measures for students in risk of attrition.

The reasons for leaving the university education by literature:

It is wide known that HE attrition is a one thousand variable question, due to multi-causality of this phenomena. The reasons for leaving the HE studies are numerous. Among them the important place takes does student come from urban or rural area[9]. **Students from urban areas are three times more prone to dropout** from university than their peers from rural areas, in Germany (Behr *et al*., 2020).

(Guzmán *et al*., 2021) presents its systematic literature research results of investigating 59 dropout explanatory variables, for students located or coming from rural areas in Europe, where 35 percent of dropout is explained by individual determinants, 25 percent carries socioeconomic reasons, 27 percent is caused by academic, and 13 percent by

[9] According to Census in 2013, B&H has 57.3 percent of citizens who live in rural areas, (Agency of Statistics B&H, 2014).

institutional determinants. Among **individual determinants** which are cause of desertion, Guzman identified gender, especially woman dropped out more; age, where are different reasons for leaving by younger and older students; unemployed adults who studying; education level of the parents; student's mental health; minorities (ethnicity, or race, or belonging to some of the social groups like illegal immigrants). Rural students may to experience higher stress and pressure due to new educational environment, commuting, academic demands and creation of new personal relationships. In the cases of single-parent, or extended large families where grandparents, relatives and siblings live together, due to family environment and pressure, students can make decision to leave the university. Procrastination, fear of failure, which is the most present at freshmen year, and lack of self-autonomy have negative consequences to dropout. Students who work part-time, or full-time must do tradeoff between hours on work, and hours for study which lead to pressure and dropout.

Guzman noticed that **socioeconomic determinants** make difference in the countries where the tuition fees are small, or free of charge, and where social differences are not so intensified, thus the family income or student's employment status do not affect HE dropout. Accommodation during study can be an obstacle for rural students with low family income, and lead to demotivation to pursue HE, because students are forced to find cheap accommodation far from university, and to commute on daily basis.

Guzman also highlighted the student's **pre-academic background** like the knowledge of declared major, GPA in high school, or achievement in special courses (like medicine, engineering) that are highly connected to academic determinants of churn. Academics GPA is the strongest indicator of churn, the more higher academic's GPA is, the less are chances of dropping out. Family pressure, or lack of information on mayor declared, cause the selection of faculty, which is not suitable for student, and increase dropout risk. Absence from classes from many (un)objective reasons is among strongest criterion of dropout.

Institutional programs like persistence or graduation plan significantly improve students' autonomy, GPA, and learning skills, comparing to students who do not participate in those programs. Diversification of communication channels, especially in the case of virtual studies decrease the dropout. Another important factor is recognition of pre-academic knowledge, courses, or work, which encourages retention.

Unlike the Guzman, the systematic review of empirical literature in Europe, (Kehm *et al*., 2019) divides the reasons for churn at those at the university side, students side and external conditions, which are not influenced by HE institutions (Table 5).

Table 5 – Determinants of dropout at universities in Europe

Determinants of churn in HEI	Affected by
Study conditions at university:	Influenced by university side. HE institutions with higher institutional resources meets lower dropout rates. – Curriculum, study structure, exam organization – Teaching and exam methods – Learning environment – Support and counseling services – Peer effects – Major declared
Academic integration at university	Student and university side of influence.
Social integration at university	Student housing (living in campus).
Personal efforts and motivation	Interest in the subject, Time used for self-study, Interest in the future job
Informational and admission requirements	Pre-enrolled entrance: admission via tests which are graded, and which are not graded. Study demands prior to university enrollment.
Pre-university education	School achievement, subject focus, private or public-school type. Grades in Mathematics.
Personal characteristics of student	Age, gender, learning approach and conscientiousness.
Sociodemographic background	Education level of parents and belonging to social class (occupational level) of parents.
External conditions	Financial status of student (family). Working status of student while study (part-time).

Source: Author, based on (Kehm *et al*., 2019)

Different classification of dropout reasons summarized by Kehm et al., are portrayed due to understanding of complexity and chaos in the field. The most comprehensive state of the art of university dropout determinants, collected between 2000-2020, only from peer-reviewed journals, limited to Europe's countries, in English, come from (Behr *et al*., 2020). Authors summarized dropout reasons by institutional and individual level, showing a huge number of dropouts describing variables, and demonstrate the necessity of field development in a following ways: a) existence of the large and longitudinal countries data studies for comparison, comprised not only the university data, and b) accent to the large and longitudinal detailed survey of dropout students.

3.3 Description of the ML models for churn classification

The four machine learning models were used to find the one that best fits our data and provides the highest performances in identifying the dropout. We started with an advanced DT, an ensemble of trees, a Histogram-based Gradient Boosting Classification

Tree (HGBC). The other models are random forest, neural network, and the support vector machine with 7 different kernels.

3.3.1 Decision tree base model

We began with a decision tree of the type often used to support decision-making processes in supervised machine learning. In the first step of applying the HGBC tree, we trained our tree on 80 percent of the dataset. The training was stopped when the model reached the required level of accuracy on the training data. The trained model is then applied to the test data, i.e., the remaining 20 percent of the dataset. The general approach to the classification learning task using a decision tree is presented in Figure 4.

One of the challenges in machine learning is overfitting the model. It occurs when the model has high performance on the training dataset but poor performance on the test data. We wanted to be transparent about the results, so in the summary of model metrics, we report the results for both the training and test datasets.

The "industry standards" among decision tree algorithms are the C5.0 algorithm, XGBoost, other popular methods are MultiBoost, Classification and Regression Tree (CART) - which uses the Gini index as a metric, Iterative Dichotomiser 3 (ID3) - uses the entropy function and information gain as a metric, Chi-squared Automatic Interaction Detection (CHAID), Decision Stump, M5 and Conditional Decision Trees.

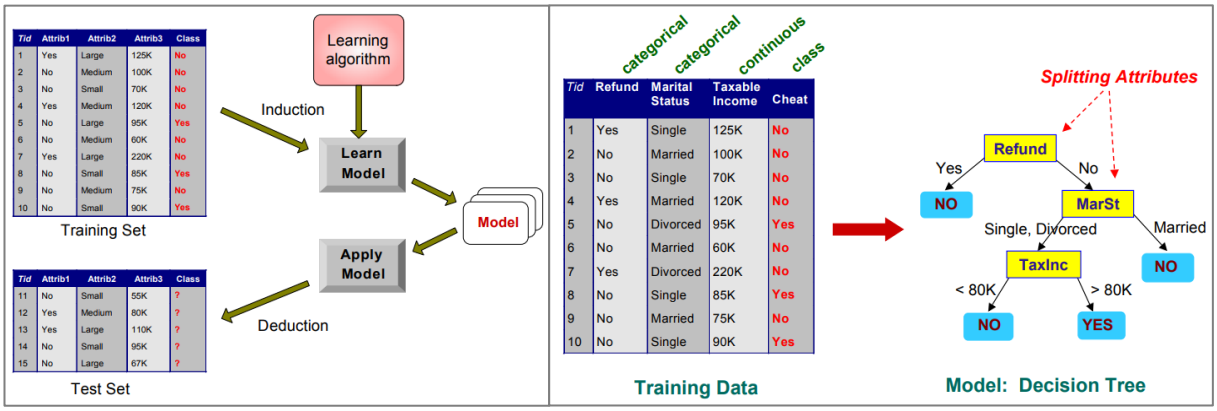


Figure 4 - General approach to the classification learning task and decision tree example
Source: (James *et al.*, 2013)

Advantages of decision trees in machine learning are, according to (James *et al.*, 2013a):

- Trees are easy to understand and explain compared to some other machine learning models. Trees are not black box models.

- DT follow the human decision process more than the regression and classification approach.
- Because of their graphical representation allows them to be easily interpreted by non-experts, making DT popular in various fields.
- Trees can easily handle nominal predictors without the need to create dummy variables.

Disadvantages of DT are:

- Trees do not have the same predictive accuracy level as other regression and classification approaches, so methods such as bagging, random forest, and boosting must be extended (James *et al*., 2013a).
- A small change in the training data can lead to a large change in the DT and, consequently, the final predictions (James *et al*., 2013a)
- The trees are based on a greedy algorithm with significant uncertainty inherent in finding the globally optimal decision tree.

One of the most critical issues in the decision tree concerns the size of the tree and the complexity: the cost of splitting, when to stop splitting, and how reasonable the splitting is. The splitting cost is related to the greedy algorithm to find the most homogeneous branches. There are two approaches to terminating splitting: setting a minimum number of training inputs for each leaf and setting a maximum depth of the model. The Gini score can provide information about the quality of the splitting.

Researchers continue to improve decision trees using ensemble and boosting techniques suitable for targeting datasets with missing values or imbalanced datasets.

Histogram-Based Gradient Boosting Classification:

Light Gradient Boosting model emerged from the need to overcome the limitations of existing GBDT, such as time consumption and accuracy. Previous models checked the information gain of all possible splitting points for each variable, making the calculations more complex and time-consuming. Resources needed in this case are equal to the product of the number of variables and the number of instances, which is not a problem for small datasets but becomes a challenge for large datasets.

Authors of HGBC (Ke *et al.*, 2017) presented a solution to the problem in two directions: reducing the number of instances and decreasing the number of variables, without compromising accuracy, while significantly reducing computation time. On the side of

reducing the number of instances, the authors used Histogram based algorithm (HBA) and Gradient-based One-Side sampling (GOS). HBA were used to find the best split by grouping continuous variable values into discrete bins and selecting those with the greatest contribution to info gain. HBA was chosen because it proved to be less time-consuming than the pre-sorted algorithm, which had been dominantly used by existing models. GOS first sort by absolute value and select top N, while randomly selecting and assigning a constant to the remaining (small) ones. A histogram is used in this random selection process.

Regarding the variable reduction side of problem, the authors use two techniques: Greedy Bundling algorithm and Merge Exclusive Features. Using a histogram, exclusive variables are grouped into a new variable. The model constructs a histogram of variables that is almost identical to the histogram of individual variables, where it is possible to identify the value of each individually, thus not losing accuracy and using significantly fewer resources.

Our model is implemented using the scikit-learn library in Python. Some crucial model parameters that should be noted and can be adjusted are *max_bins, max_iter, max_depth* (of each tree), *l2_regularization, learning_rate*. Those parameters were left as default.

We chose this tree classifier because, among other reasons, it supports missing values, performs well on datasets with more than 10,000 samples, has good document support, and is commonly used for binary predictions but is also suitable for multiclass predictions (Guryanov 2019), (Pedregosa *et al*., 2011). It outperforms the "normal" tree because it is less affected by outliers, has sample weighting and support for categorical variables (no need for one-hot coding), and provides the ability to add prior knowledge to the model about which predictor variable contributes more to the target variable (monotonic constraints), and is faster than the "normal" tree.

Its advantages are comes into focus in the chapter 6 when we explained how our models "decide" among churned and not churned students, using SHAP and PI.

3.3.2 Random Forest

The Random Forest classification model can be used for classification and regression tasks as well as in supervised and unsupervised learning. The forest consists of *n* trees without pruning. Each tree grows to its maximum extent. Random Forest uses a collection of trees to increase the accuracy of the predictive model. "*A Random Forest is a classifier consisting of a collection of tree-structured classifiers, where the classifiers are*

independent, identically distributed random vectors and each tree gives a unit vote for the most popular class given input x" (Breiman, 2001, p. 2).

As the number of trees increases, the model becomes more robust and accurate. It gives the strong and weak predictors in the dataset an equal chance of being considered in the partitioning of the nodes (James *et al*., 2013), i.e., they are normally distributed. In the case of classification, the classifier that receives the most votes in the set of trees will be the highest vote and the result of the RF prediction. RF combines the different weak classifiers of the same type to produce a strong classifier, i.e., ensemble learning. It is possible to draw a parallel between RF and the wisdom of the crowd experts when it comes to classification tasks.

Advantages of RF (Criminisi *et al*., 2011):

- Works for classification and regression, supervised, unsupervised, and semi-supervised learning.
- Fairly good handling of missing values and maintaining accuracy with missing data.
- Random forest will not overfit the model.
- Handles large data sets with high dimensionality.

Disadvantages of RF:

- Comparing results between classification and regression tasks, the model shows better performance in classification due to the continuous nature of regression prediction.
- Black box inherent disadvantage in that there is little control over what the model does by trying different parameters and random seeds.

The main goal in predicting student attrition is to use RF for classification to achieve high model performance for identifying dropouts. We had to exclude some input variables due to request of RF for non-missing values (i.e. the handling of missing values was not added to RF).

3.3.3 Support Vector Machine

Another supervised and nonlinear machine learning classification model is the SVM, Support Vector Machine (Cortes and Vapnik, 1995) and kernel-based methods (Abe,

2010) which was developed in the 1990s as a linear classifier. It is mainly used in the field of predictive classification and provides high accuracy with low computational power. The algorithm finds the boundary hypersurface between separable data. In two-dimensional space, the hyperplane is a line, in three-dimensional space, it is a two-dimensional plane, and in n-dimensional space, it is an (n-1)-dimensional hyperplane. From the infinite number of hyperplanes, the one with the largest distance (margin) between the data points of two classes and the hyperplane is selected, Figure 5. It is called a hyperplane with a maximum margin.

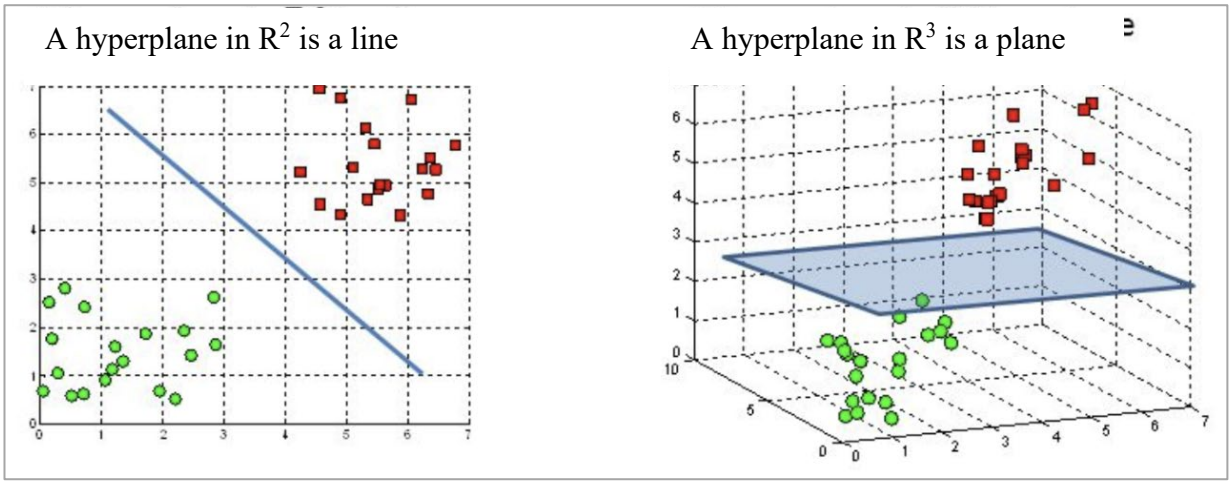


Figure 5 - Hyperplanes in 2D and 3D feature space in the SVM learning algorithm
Source: (Gandhi, 2018)

The position and inclination of the hyperplane are determined by the data points that have the least distance from the hyperplane - the support vectors that are critical in building SVM models. The cost function to maximize the margin is the loss function. One can distinguish between hard and soft margins. The main difference between soft and hard margins is that there are imperfectly separable data points in the soft margin case, implicitly introducing an error that must be minimized to mitigate misclassification in the sample (Auria and Moro, 2008).

The SVM model has evolved from a linear and parametric model to a non-linear and non-parametric one by introducing a kernel. The kernel is a function that can produce *n*-dimensional space and softener margins. In our research, we experimented with the following kernel functions:

– The linear kernel is the most straightforward function and provides a two-dimensional space of decision boundary in binary classification.

- Polynomial kernel where we set kernel degree at 3 and at 8.
- Sigmoid kernel, since most of our predictors, are categorical (binary coded) variables.
- Radial Basis Function kernel, where we set the C, as a regularization parameter at 0.1, and experimented with gamma parameter at 0.1, 0.5, and 1.0.

Like other machine learning classifiers, SVM has its advantages and disadvantages, which are more or less important depending on the dataset to be analyzed. The advantages of SVM are as follows (Auria and Moro, 2008):

- By adding the kernel, the SVM achieves flexibility in choosing the shape of the threshold that separates the binary data and the accuracy level.
- Expert judgment of model results is not required because kernel transformation assumes a robust theoretical foundation: with a kernel, data points are transformed nonlinearly, and data points are linearly separable.
- SVMs provide good out-of-sample generalization because it is possible to rescale outliers. This means that by choosing an appropriate level of abstraction, SVMs can be robust even if the training sample has some bias.
- SVMs produce a unique solution. This is an advantage over neural networks, which have multiple solutions associated with local minima, and for this reason may not be robust across different samples.
- In the case of students dropping out, by choosing appropriate parameters (kernels), it is possible to establish more similarities between students who have dropped out. When a dropout prediction is made for a new student, it is made based on the group with which he has the highest similarity.
- Works well with small data sets (Noble, 2006).

SVM belongs to the group of nonparametric methods and has the inherent disadvantage of nonparametric methods as lacking transparency of results and the inability to represent the score for all records in the data set (in this case students) like a simple parametric function of student churn. In addition, SVMs are less effective on noisy datasets with overlapping classes and are unsuitable for larger datasets because training time can be high (Noble, 2006).

3.3.4 Neural network

The concept of a neural network has been known since 1873 (Bain, 1873) and was introduced by two philosophers Alexander Bain and William James (James, 1890). The neural network concept received its first mathematical algorithm in 1943 (McCulloch, 1943) and continued to evolve until 1969 (Block, 1970), when Marvin Minsky and Seymour Papert pointed out two main obstacles to the development of the NN. The first obstacle was the inability to solve the exclusive-or problem, and the second was low computer power, as computers were unable to handle large processing tasks. However, this changed in 2006 with the pioneering research of Hinton, LeCun, and Bengio in machine learning algorithms (Hinton *et al*., 2006), (Bengio and LeCun, 2007), which dramatically accelerated the implementation of Deep Learning. Deep learning is a subclass of artificial NN. Every NN belongs to machine learning, but not all NN belong to Deep Learning. It is generally accepted that a deep learning NN has to contain at least two hidden layers.

Neural networks are used to solve complex tasks such as pattern recognition, where they receive a series of inputs from neurons (input layers) and produce the output (e.g., classification, pattern recognition). There are one or more hidden layers between the output layer and the input layer (of neurons). These layers are not input or output neurons. Each neuron has a weight and a threshold or bias. The weight is a real number that represents the importance of each input to the output. The output of a neuron can be represented as a number between 0 and 1 or -1 and +1 and depends on whether the sum of the weights is less than or greater than the threshold. The threshold or bias is a parameter of the neuron that indicates how easy it is to get the neuron output to 1 or to get the neuron “to fire". By changing the weights and biases, one can influence the decision process of the model.

The feedforward network is commonly used for classification. It is a NN constructed so that the output of one layer is the input to the next layer, i.e., there are no loops. The forward method is used to train NN (Nielsen, 2015). The weights and biases change slightly as long as the NN is trained on a test data set to achieve high accuracy. When the NN is well trained, it can make very accurate predictions, in some cases more than 95 percent of the time.

Advantages:

- NNs can solve complex nonlinear models without specifying them in advance (Livingstone *et al*., 1997), especially in complex pattern recognition problems (Nielsen, 2015)where they are superior to other machine learning methods.
- Different insight into the data and the ability to have insight into the two-dimensional and multivariate data space (Livingstone *et al*., 1997).
- Breakthroughs in algorithms accelerated the training process and enabled the use of more data than ever before, which is appropriate for the Internet of Things era.

Disadvantages:

- The problem of selecting the most relevant variables for training the model among the different inclusion strategies (Forward Inclusion, Backward Eliminator), each strategy has its own drawbacks. Some data types fit some models better than others (Livingstone *et al*., 1997).
- Problems of overfitting and overtraining in NN can be solved by combining appropriate architecture and training sets, as well as proper selection of hyperparameters, but they are still common (Livingstone *et al*., 1997).
- The inherent black-box problems, such as the explanation or interpretability of the results, are the cause of the error. In some cases where interpretation of results is important (e.g., bank loan applications, user account management, or business decisions), it is not possible to use a NN (Livingstone *et al*., 1997). But this is diminished with the development of the interpretable machine field.
- Cost of building the NN and lack of control of the algorithm.
- The large amount of data needed to feed the NN can be broken down into smaller amounts by understanding the problem, and then it is possible to use a model with stronger theoretical foundations.

It is necessary to analyze and study the problem and the data before starting to build a NN. Another aspect of the NN that should be considered is the need for graphics processing units (GPU) or special processors for NNs (e.g., Intel Movidius, Google TPU, Apple M1 processor) due to their parallel processing capability compared to central processing units (CPU).

NN for deep learning in Keras library:

Learning algorithm backpropagation is a very popular learning algorithm in a supervised learning architecture. It was developed by Paul Werbos (1974) and is based on gradient descent (GD). This method minimizes NN errors by changing the NN error curve gradient. "*Backpropagation can be applied to any system with a well-defined order of calculus, and even if those calculations depend on past calculations within the network itself*" (Werbos, 1990, p. 1554). The name backpropagation comes from the way the method works. It takes the output error and goes back through the network (propagation) to look for paths and their weights that had the most significant impact on the output error. Mathematically, these paths are derivatives of the activation function (which accounts for bias and error) that are evaluated by the algorithm. The algorithm is presented in detail in a book (Nielsen, 2015).

We set the hyperparameters for the NN setting: Activation functions, number of hidden layers and nodes, learning rate, number of epochs, batch size, and optimizers.

Activation function: we used two hidden layers with the Rectified Linear activation function (ReLu) for the hidden layers and the sigmoid activation function for the output layer. ReLu is a linear function and does not require any transformation of the inputs. It only takes positive values of the node outputs further through the network. If the output of the node is less than or equal to zero, the function returns zero. The advantages of this approach reported in the literature are better performance and simpler training.

Sigmoid is a nonlinear (logistic) function, and the inputs to the function are between 0 and 1, giving it the classic S-curve. This is the reason why we use it in our binary prediction. Our dataset contains mostly categorical features that are binary coded, as is our target variable.

We experimented with techniques to automatically tune the hyperparameters using the Adam (adaptive moment estimation) optimizer, which is the most commonly used in practice. Adam's parameters are: Step size or learning rate (to find the local minimum), which slow down the learning process if they are small, and vice versa; the decay rate for the first and second moments of the estimates; Epsilon - to prevent division by 0.

Adam is intended for the training part in the NN, where it replaces the classical gradient descent process by applying a different learning rate for each weight instead of applying the same learning rate to all weight updates. This includes the advantages of two types of gradient descent: adaptive gradient algorithm and root mean square propagation, resulting in fast calculations.

We experiment with different layers and neurons, where size is the number of hidden nodes, width is the number of nodes per hidden layer, and depth is the number of hidden layers, Table 6.

Table 6 – Size, width, and depth of trained neural networks

Modeling data	Description	Scenario 1	Scenario 2	Scenario 3
Pre-enrollment	Size	57	77	87
	Width	27/30	27/30/20	27/30/20/10
	Depth	2	3	4
Enrollment	Size	67	87	97
	Width	37/30	37/30/20	37/30/20/10
	Depth	2	3	4
End of year	Size	75	95	105
	Width	45/30	45/30/20	45/30/20/10
	Depth	2	3	4

Source: Authors' contribution.

Input layer neurons are equal to the number of predictor variables. The output layer has two nodes.

3.4 ML evaluation metrics

After employing the ML models, we need to evaluate the modeling results to find the best model for university churn classification. Except the standardly used evaluation metrics, models are evaluated by the time needed to fit the model, and to calculate SHAP values and PI.

The first metric implemented in evaluation phase is confusion matrix, Table 7. TN represents the number of university students correctly classified as non-dropouts, meaning they actually graduate from university. FP carries the number of undergraduate students misclassified as dropouts but who actually are not (error Type I).

Table 7 – Confusion matrix for binary classification

		Predicted class	
		0 (False)	1 (True)
Actual class	0 (False)	True negative (TN)	False positive (FP)
	1 (True)	False negative (FN)	True positive (TP)

Source: (Kulkarni *et al*., 2020)

FN is the number of university students misclassified as non-dropouts but who actually are (Type II error). TP represents the correctly classified students as dropouts, meaning that they did drop out.

Accuracy tells us that all positive and negative cases were correctly predicted. It is useful measure when both classes (in our case dropouts and non-dropouts) are equally presented in the train and test set. Some authors argue that precision can be used when the main class is less than 80 percent represented (Brownlee, 2021). This matrix yields the following coefficients (Sharma *et al*., 2022):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Precision is the proportion of true positive cases correctly identified among all predicted dropouts by the model.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity or recall (TP rate) is the proportion of actual positive cases that are correctly predicted among all truly dropped out students.

$$TP\ rate = \ Sensitivity\ (recall) = \frac{TP}{TP + FN}$$

Specificity is the proportion of actual negative cases that are correctly identified among all students who are not dropouts.

$$Specificity = \frac{TN}{TN + FP}$$

F1 or F score measures the importance of both: precision and recall. It works well when both of the target classes (dropouts and non-dropouts) are equally presented in the data set.

$$F1 = 2 * \frac{precision * sensitivity\ (recall)}{precision + sensitivity\ (recall)}$$

The evaluation of the above metrics is divided into three groups of metrics (Table 8). Alternative or additional measures are needed in machine learning because some performance measures do not distinguish between the number of correct labels of different classes (like precision) and focus only on the class (Sokolova, *et al*., 2006).

Table 8 – Classical performance measures of HGBC, RF, SVM, and NN in this thesis.

Performance group	Measure	Short description
Threshold/ bias	F1 Score	Harmonic average that takes into account precision and recall (scale between 0 and 1) and does not take true negatives into account from the

Performance group	Measure	Short description
		confusion matrix. The value 1 means perfect precision and recall, while 0 means the opposite.
	Accuracy	The simplest measure: the number of correct predictions divided by the number of total predictions.
Ordering/ rank	ROC Area[10]	The power of the model to make distinctions between the classes (between 0 and 1) is represented as a graph that plots true positive and false positive rates from the confusion matrix. The higher curve, the better classification model.
	Precision	The number of correct positive results divided by the sum of the true and false positive predictions by the classifier.
	Recall	It is the number of correct positive results divided by the sum of the true positive and false negative numbers (all samples that should have been identified as positive).

Source: (Caruana and Niculescu-Mizil, 2006), (Williams *et al*., 2006)

ROC (receiver operating characteristic) AUC (area under the curve) is another way to validate our models, as it represents the plotted values between FP and TP rates, where the threshold is between 0 and 1. Values equal to or less than 0.5 indicate that our model is unusable at the current threshold level. The model performs better when the values are closer to the upper left quadrant, i.e., at 1. The ROC AUC is widely used metrics to evaluate binary classification ML models with balanced data.

3.4.1 Accuracy paradox and additional metrics for imbalanced data sets

The accuracy paradox in binary classification arises when one class (e.g., dropouts) is significantly underrepresented compared to the other class (non-dropouts). When a model is trained on highly imbalanced data, it may exhibit a seemingly high accuracy. However, this high accuracy is misleading because it primarily reflects the model's tendency to classify most instances as belonging to the majority class, rather than its true predictive performance across both classes. Our data are almost in perfect balance, but due to experiment with churn definition change, the imbalance was occurred.

In our data set with definition of churn change, the imbalanced data, many categorical features, and missing data, we cannot rely on traditional accuracy measures. In a situation with imbalanced data, instead of finding the most accurate model, we must make a trade-off between the cost of identifying the student who actually dropped out as a non-dropout (False Negative) and the cost of identifying the student as a True Positive (Table 7), i.e., a trade-off between precision and sensitivity (recall).

[10] Aslo written as AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve, and AUROC (Area Under the Receiver Operating Characteristics).

In the literature we found following metrics for binary imbalanced classification tasks: Cohen's kappa, Matthew's correlation coefficient (MCC), Youden's J statistic, G-mean, PR curve, and Unweighted Average Recall (UAR).

Cohen's Kappa was introduced in 1960 by Jacob Cohen (Cohen, 1960) and is used for imbalanced data sets. The formula is:

$$K = \frac{p_0 - p_e}{1 - p_e}$$

Where p_0 is the overall accuracy of the predictive model and p_e is the probability of agreement between the model predictions and True class.

The advantages of Cohen's kappa are:

- For imbalanced data sets, it is a better-suited measure than overall accuracy.
- It is possible to calculate the maximum achievable Cohen's Kappa value for each confusion matrix and compare it to the observed Cohen's Kappa value.

Some disadvantages of this metric are:

- Cohen's Kappa metric can't be used to explain the accuracy of a single prediction.
- Balanced data sets yield higher Cohen's Kappa values. The values are lower for imbalanced data sets than for balanced data sets.

To compare our binary prediction models using imbalanced datasets, we employed another metric to evaluate models with imbalanced datasets, MCC. This measure is based on all four values of the confusion matrix, as opposed to precision, F1 score, and recall. It treats the True class (0 and 1) and the Predicted class (0 and 1) as two variables and calculates Pearson's correlation between them. MCC is actually a Phi- coefficient (φ) applied to binary classifiers, with the following formula:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Values range from -1 to +1, where 0 means that there is no correlation between the predicted class and the observed class (the model is random). Higher values mean that both classes are accurately predicted. Lower values are indications that True class and Predicted class are poorly correlated (Shmueli, 2019).

Youden's J statistic (also called informedness or balanced accuracy score) was introduced in 1950 by W. J. Youden (Youden, 1950) to measure the performance of accuracy of disease diagnostic tests in medicine. The test values range from 0 to 1. If the value is 0,

the metric is useless. A value of 1 indicates that there are no FN or FP values, i.e., the test is perfect. The simplified formula is:

$$J = sensitivity + specificity - 1$$

If the value of the statistic is negative, it is a sign that we have reversed the positive and negative classes. Youden's J statistic can be visually represented as the value of the vertical line ROC AUC. The Youden's ratio J is calculated for each point on ROC, and the maximum value of the J statistic is a point of the model with the best performance. Given that we conducted only a single experiment using the best performed model with imbalanced data, we chose to evaluate the performance of the imbalanced models using only the Precision-Recall (PR) curve and correlation analysis.

3.4.2 Explainabilty of ML models

The tradeoff between interpretability and accuracy of the model is quite present in the literature.

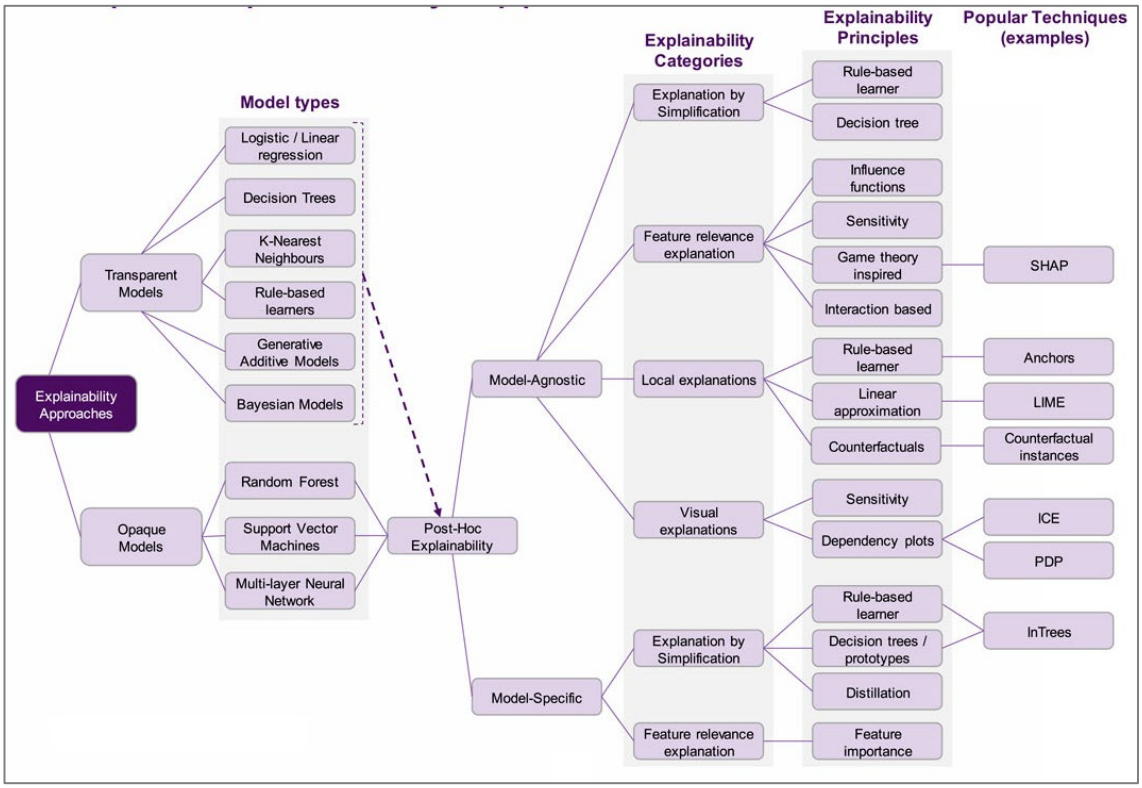


Figure 6– Taxonomy of the approaches to explainability of ML models

Source: (Belle and Papantonis, 2021), p. 5.

The possibilities of interpretation and visualization of ML models depend on whether the model belongs to transparent or opaque models. In this thesis we employed opaque

models: RF, SVM, NN and HGBC that is DT, but as ensemble tree belongs to the opaque models, i.e. black box models (Figure 6). Therefore, visualization and interpretation possibilities are limited to post-hoc methods: feature relevance, and model simplification. The SHapley Additive exPlanations (SHAP) was chosen as industry standard, due to its wide exploitation in practice and by researchers. The SHAP was introduced by (Lundberg and Lee, 2017) in order to offer reliable solution to the accuracy – explainability challenge, and to be model agnostic, i. e. applicable to any ML model. Authors were inspired by work of (Shapley, 1953) and its contribution to the Game Theory, where Shapley values indicate the individual contribution of each player to the game's overall outcome. That approach was transferred in the context of ML models, where features act as "players," and the model's prediction serves as the "outcome" of the game. For a given prediction, SHAP values quantify the contribution of each feature to that prediction. Positive SHAP values indicate that a feature is pushing the prediction higher, while negative SHAP values indicate the opposite. In our case of binary prediction, higher values move up the prediction to the value 1 (dropout), and vice versa.

Pros:

- Model agnostic solution: the possibility of application is not limited to one ML model, or restricted to data set specificity.
- The power of visual interpretability by non-expert users and understanding feature interactions, and building trust with stakeholders by providing transparent explanations for model predictions.
- Model debugging by analyzing the contributions of individual variables to model predictions to identify potential issues or biases in the model's behavior. In that way we gain a deeper understanding of ML model's behavior and identify areas for improvement, ultimately leading to more robust and trustworthy ML systems.

Cons: SHAP alone cannot provide answers to the questions like how much would the student risk of dropping out changes if he/she would have different high school degree, Table 9?

Table 9 – Examples of SHAP limitations in ML interpretation

Potential limitation case	Example in our case	Solution
How does the model's outcome change if one variable changes, and others remained the same?	How does the student risk of dropping out changes if he/she would have different high school degree?	Partial Dependence plots, Individual Conditional Expectation combined with SHAP

Potential limitation case	Example in our case	Solution
How model did make classification regarding the individual instance?	How model decided to classify particular student as dropout?	Anchors combined with SHAP
How much some variable has to be changed to influence the model outcome change?	How many (more) ECTS credits must obtain particular student to be classified as non-risk student?	Counterfactuals
Which instances have the highest impact to the model outcome?	Which (group of) students have the highest impact to the churn classification?	Deletion diagnostics
How in general model made decisions?	Can we set a rule(s) for identification student at risk of dropping out?	InTrees

Source: Authors adaptation and (Belle & Papantonis, 2021)

For some questions it is possible to provide the answers by combining the SHAP with other techniques. This implies that only one technique for interpreting the ML models outcomes is not enough, and the unified approach sill does not exist.

3.5 The research gap

In order to address student dropouts effectively, it is crucial to identify the types of students who are at risk of leaving HE. The previous section of this dissertation highlighted the **information gap** about the number of students who leave higher education and reasons to do so in Bosnia and Herzegovina. This lack of data extends to neighboring countries as well, where there are also no reports or statistics about interruptions in HE, aside from occasional research papers. Bosnia and Herzegovina is not the only case where there is a lack of information about students leaving tertiary education. This issue is not only present in European countries, but also at a global level. Few countries in Europe and around the world have established standardized methods and statistical reports for tracking study dropouts. In addressing this research gap, this thesis provides data on the study interruption trends and reasons for churn for the years 2007 to 2018. The dropout data includes approximately 20 percent of students from Bosnia and Herzegovina (i.e. the UNIBL).

The literature review revealed the second research gap: there are **domestic studies** in the field of educational data analysis and machine learning. However, these studies mainly focus on secondary schools or smaller groups of students, predicting their success in passing specific courses. The time frame for students' cohort in domestic research is a maximum of two school years in one research paper, while it was shorten in others. The

dissertation makes a theoretical contribution by modeling students churn on a longitudinal dataset, where each generation was tracked within minimum 6 years from enrollment.

On the global level, researchers highlighted the need for longitudinal data usage in modeling the dropout, as well as the comprehensive survey of students who leave the tertiary education institutions. Also, (López-Zambrano *et al*., 2021) stressed out a lack of research of churn prediction in the earliest stage of tertiary education. In this research the churn prediction is set in three time points: pre-enrollment stage, in enrollment week, and at the end of first (freshman) year, to enrich the segment of early prediction.

While the existing literature includes instances of the Histogram-Based Gradient Boosting Classifier, HGBC, being employed, its application in the field of EDM has not been documented. This study contributes to the field by introducing HGBC, thereby enhancing model diversity and enabling a comparison with commonly used machine learning models.

With the rise of explainable machine learning, nine papers were identified that tackles the dropout and explainable models by June 2024. They are presented in more details in Chapter 3.2. The dissertation makes a theoretical contribution by addressing the problem of poor dataset and missing data by modeling students churn on mostly binary dataset. The variety of ML models were run and their interpretability and explainability were compared using visual explanation techniques: SHAP and PI. Both techniques are model-agnostic, making them suitable for comparing feature importance in this research, especially when extended to other public universities in B&H.

The practical contribution is by setting the effective model of predicting students churn at earliest stage at UNIBL, to prevent it on time. This dissertation also want to emphasize the need to go beyond “simply” setting up an early warning system that alerts professors when a student is at risk of dropping out. The best possible additional outcome of this research is rise of awareness of students HE churn importance which may lead to establishing a systematic support program to prevent attrition in higher education in the whole country by policymakers.

4 METHODOLOGY AND DATA

The Chapter four contains brief description of chosen methodology, data and steps that were needed to be done prior data preparation. Two types of data are collected: qualitative and quantitative. The purpose of qualitative data was to set and support the dropout definition, that's why they are presented and described first. The second purpose of those data was to find the reasons of leaving the HE, and present it in the systematic manner for the first time in the country, which is summarized in following chapter. The purpose of quantitative data was to prepare dropout report and serve as training and testing set for churn prediction and explanation of the model outputs.

4.1 The methodology framework and research design

The methodological framework used for the research is the Cross Industry Standard Process for Data Mining (CRISP-DM), a complete process that includes feedback support, checks and balances, properly created and transferred, starting with understanding the problem (assuming that the researcher has no prior knowledge in this area) and ending with applying the model (answering questions). There are practitioners who argue that CRISP-DM is not suitable for the Big Data domain when the data is real-time, large-scale, and has other characteristics of Big Data: Variety, Velocity, and Veracity (Stirrup, 2017). Considering the arguments of these authors, the datasets in this research do not have the mentioned characteristics, and in the author's opinion, CRISP-DM is the appropriate framework for the methodology. In the literature review, some authors have used this methodology in machine learning prediction models for student retention and attrition (Delen, 2011).

CRISP-DM is one of the oldest and most popular models in data mining, proposed in 1990 by a European business consortium that included Integral Solution Limited, the original owner of IBM-SPSS, National Cash Register Company, and DaimlerChrysler. The methodology is developed in a six-step process, with each step consisting of several sub-steps, which are shown in Table 10. The order of the steps is not strictly defined, and as seen in Figure 7, one can move forward or backward. The advantage is also the flexibility of the model so that it can be changed depending on the problem area, and then some steps can be emphasized more than others (IBM, 2014).

Table 10 - A brief overview of the six steps of the CRISP-DM methodology with sub-steps.

Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine business objectives	***Collect Initial Data***	Data set	***Select modeling technique***	***Evaluate results***	***Plan deployment***
Background	Initial data collection report	Data set description	Modeling technique	Assessment of data mining results w.r.t. business success criteria	Deployment plan
Business objectives		***Select data***	Modeling assumptions		***Plan monitoring and maintenance***
Business success criteria	***Describe data***	The rationale for inclusion or exclusion		Approved models	Monitoring and maintenance plan
Assess situation	Data description report		***Generate test design***		
Inventory of resources		***Clean data***	Test design	***Review process***	***Produce final report***
requirements, Assumptions and constraints	***Explore data***	Data cleaning report	***Build model***	Review of process	Final report
Risks and contingencies	Data exploration report		Parameter settings		Final presentation
terminology		***Construct data***	Models	***Determine next steps***	
Costs and benefits	***Verify data quality***	Derived attributes	Model description	List of possible actions decision	***Review project***
Determine data mining goals	Data quality report	generated records	***Assess model***		Experience documentation
Data mining goals		***Integrate data***	Model assessment		
Data mining success criteria		Merged data	Revised parameter settings		
Produce Project Plan		***Format data***			
Project plan		Reformatted data			
Initial assessment of tools and techniques					

Source: (Wirth and Hipp, 2000)

To have a clear idea of the methodology used, it is presented in Figure 7. The arrows indicate the most important and frequent dependencies between the steps.

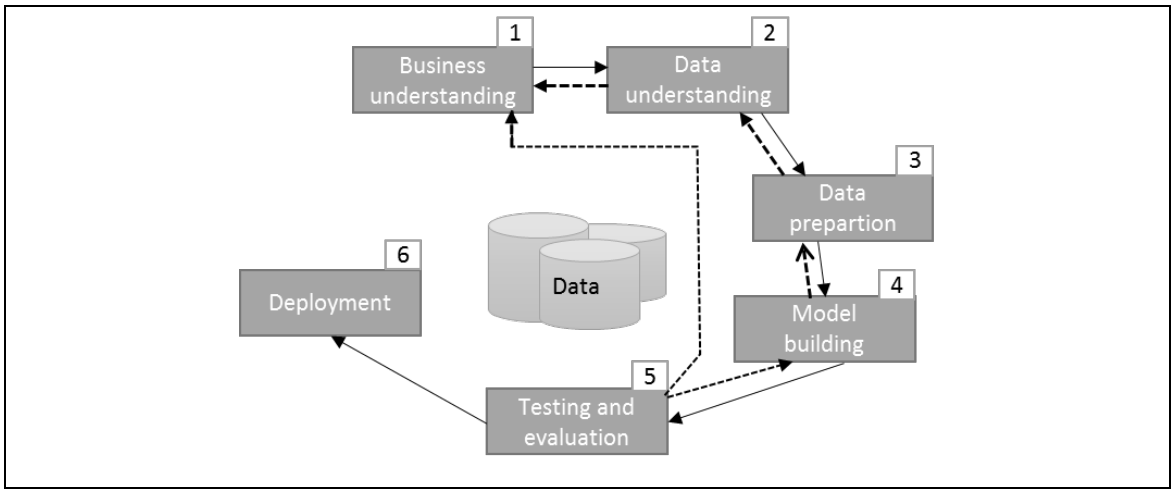


Figure 7 - Dependencies between the steps of the methodology CRISP-DM

Source: (IBM, 2014)

The research is data-driven and is combined with additional survey-based research to derive benefits from both and to reduce the negative consequences that authors have argued against survey-based dropout research because they are not generalizable and have high costs when conducted on a large scale (Cabrera *et al*., 1993).

This research uses data mining techniques to gain insights and better understand the phenomenon of university dropouts in Bosnia and Herzegovina for the first time. Historical institutional data from the University Information System database are collected and processed using ML models. The models' performances are measured and validated in practice through predefined metrics. Qualitative research is conducted at the level of university dropouts, with the sample of 96 + 10 students.

4.1.1 CRISP-DM at UNIBL case and its deployment

The methodology of data-driven research part of this study, that has a goal to identify and deploy the ML model which has the highest performance regarding dropout classification at UNIBL, is shortly presented here, together with research design (Figure 8).

Business understanding phase in our case was identification of the relevant law and University rules, and their changes and annexes between 2007/08 and 2018/19 academic year. We consider only law and rules which may affect retention and dropout determinants. For a better understanding of the UNIBL case, certain concepts from the mentioned documents are presented in more detail in the section 4.4. Also, several short interviews with administration and University data center staff were done to understand the everyday practice and how those rules are indicated in the data set obtained from the University data base.

Data understanding phase requested more information from university data base center stuff regarding when, and which faculties started to use data base, capacity of use, types of student's enrollment, typical student's path, unique identification of student, representation of variables, meaning of the empty data in specific cases, and mistakes / specific cases which occurred due to data migration in the past. In this phase, missing and exploratory data analysis (EDA) were employed, as base for data description (chapter 4.4)

In *data preparation phase*: Filtering (excluding master, PhD, and Erasmus students), removing repetitive variables, and coding the whole data set were the next steps. To distinguish between dropouts and non-dropouts, the transformation of date/time and other variables, and adding new (dummy) variables is done. In this phase data were prepared

for dropout reports by gender, the science disciplines, and study duration, but not for ML modeling. To prepare data for ML, more transformation of variables had to be done, together with data imputation, imbalance analysis, and feature selection. Due to large number of categorical variables and their coding, several variable groupings were done and their inter correlation, normalization, and impact to the model (high school type, place of residence, municipality level of development, high school degree name, ECTS credits clustered in ranks, and distance of place of habitation from UNIBL) has measured. The result of this phase was prepared data set for ML, and our understanding why, and which models, and evaluation metrics we can apply to current data set.

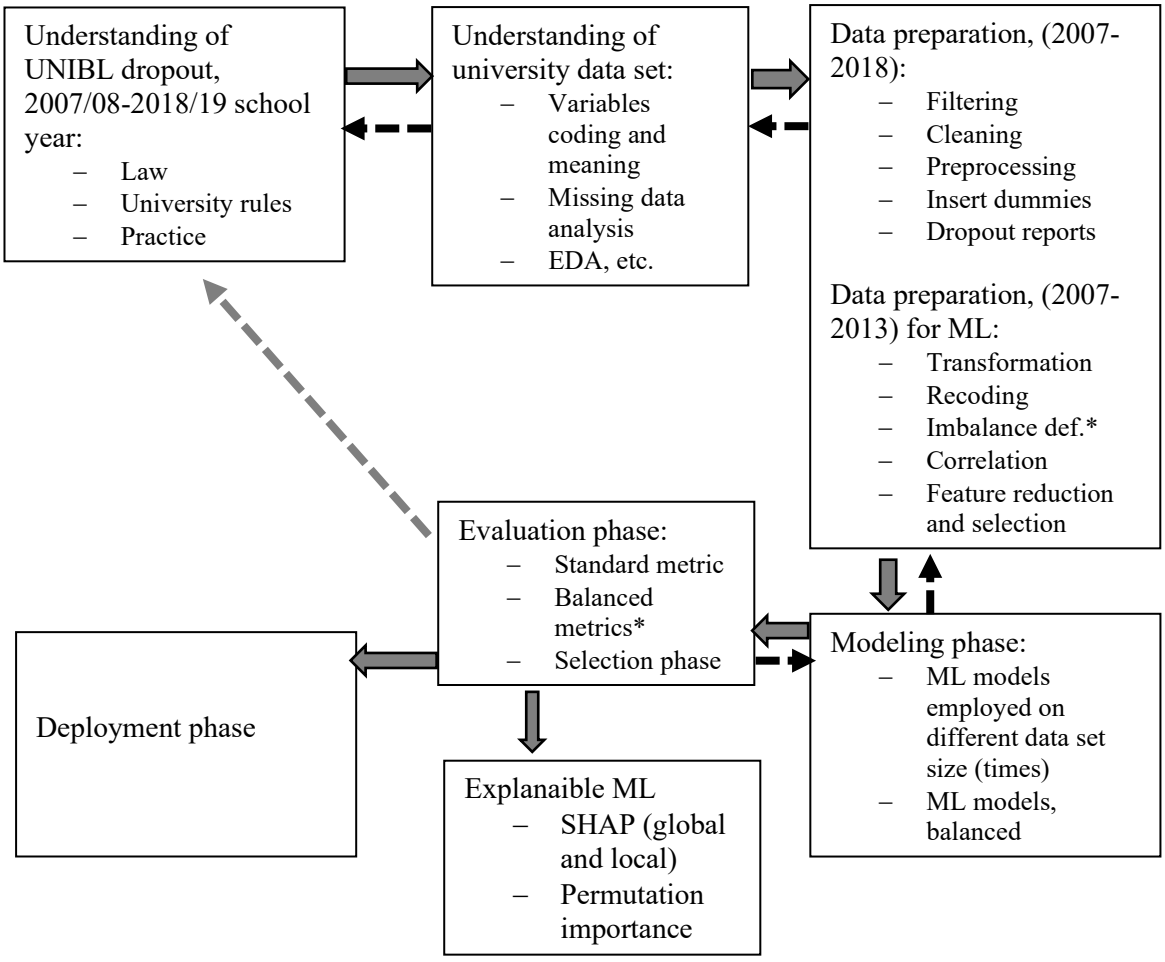


Figure 8 - Research design of the study

*As experiment using the best performed model
Source: Authors contribution.

The *modeling phase* was implemented in Python, together with split and train of models. *Evaluation* of each model by pre-defined standard metrics, PI and Python library SHAP to explain variable importance at global and local level. Models are compared regarding

the best dropout performance, ranged, and the best one is selected for deployment phase. The best model is “tested” regarding the slightly change of definition of churn, where imbalance data occurs. Also, the best model was "tested" at the end of the first year of enrollment data, at the beginning, and prior to enrollment to measure the impact of additional data added over time.

Deployment phase may began in the fall semester of 2024/25 school year. The first step is to report the UNIBL management of dropout rates in detail using current definition and data which are going to be obtained from the University data Center for 2019/20-2023/24 school year in the end of October 2024, together with ML model performance on the current dataset. The same report will be sent to the Ministry of Higher education of RS entity. This report can serve as ground base for request of additional students’ data from the Statistical office of RS entity, which are stored in electronical form for each student from 2007 until today. Those data have attributes which are not recorded in university data base, like educational level of parents, student’s residence while study, place of residence, rural/urban, detail information of pre-enroll GPA, education, finance conditions, and working status of students and parents. Those data have students ID that is the same as in the university data base, which allow enrichment of data base with highly useful variables. We asked the same university data from all eight public universities in B&H in June 2018 to conduct country representative dropout report.[11] The purpose of this doctoral dissertation is to serve as a basis for making:

1) The country representative dropout report, as base for direction of action for the universities and ministries of HE in B&H.
2) Establish the annual dropout, retention, time to degree, and completion rates reports aligned with EU suggestions.
3) Development and deployment of ML models at public funded universities and
4) Contribute to the Open Data initiative by making the university data available for the researchers around the world.

[11] We received data sets from University of Sarajevo (March 2022), University of East Sarajevo (October 2022), University of Tuzla (request granted but still not fulfill), University of Zenica (July 2022), University of Džemal Bijedić (April 2023), University of Mostar (no integrated university data base), University of Bihać (no fully electronical data base).

4.2 Dropout estimation approaches

The dropout rate calculation follows the retrospective longitudinal study approach, since all students were tracked between one (the last 2018/19) and 12 years (cohort of students enrolled into 2007/08 academic year). Then was introduced the dropout variable which contained zero value for non-dropout, and value 1 for dropout. The conditions of introducing the dropout target variable in detail are placed in chapter five.

Since the database has been in place from 2007 and data were collected from 2007 to 2018, we calculate the dropout definition as follows:

$$Dot_{ij} = \sum \frac{Do_{ij}}{E_i}$$

Dot_{ij} – total dropout rate, Do_{ij} – number of students who dropped out of generation *i* in the school year *j*. E_i – number of enrolled students by generation *i*.

Total dropout rate by gender:

$$Dot_{ij,m/f} = \sum \frac{Do_{ij,m/f}}{E_{i,m/f}}$$

$Dot_{ij,m/f}$ represents the total dropout rate of generation *i* in the school year *y*, for the male (*m*), or female (*f*) population, respectively. $Do_{ij,m/f}$ is the number of dropped male (*m*), or female (*f*) students in generation *i* in the school year *y*. $E_{i,\ m/f}$ is the number of enrolled male or female students for generation *i*.

Similar to the above, we calculated the attrition rate by science domain and gender. The faculties are divided into three science areas, as follows:

- Social Science: Faculty of Economics, Faculty of Law, Faculty of Security Sciences, Faculty of Political Science, Faculty of Physical Education and Sport, Faculty of Philology, Faculty of Philosophy.
- STEAM disciplines: Academy of Arts, Faculty of Architecture, Civil Engineering and Geodesy, Faculty of Electrical Engineering, Faculty of Mechanical Engineering, Faculty of Agriculture, Faculty of Natural Sciences and Mathematics, Faculty of Mining, Faculty of Technology, Faculty of Forestry.
- Medical Science: Faculty of Medicine.

In order to minimize the misclassification of dropout students, we calculated the average time of graduation at UNIBL, which was 5 years and two months. Due to database

architecture, we set at least 6 years wide range in the database to determine is a student advanced or at least enrolled in the last study year (for 2007-2013 generations, data set for ML). Those students are not considered dropout candidates, their status is "study". Since our database ranged from 2007 until 2018, to get the most precise and accurate student dropout classification, we trained and tested our models on the generation of students enrolled between 2007 and 2013. Generations who enrolled in 2014 and after are presented for 5 years long (and shorter) in the database, which raises the possibility of misclassifying those students who drop out reluctantly.

4.3 Data

This section describes obtained data in detail and provide insight into most relevant University rules regarding the student's attrition. For better understanding the purpose of collected data, the Figure 9 is introduced here.

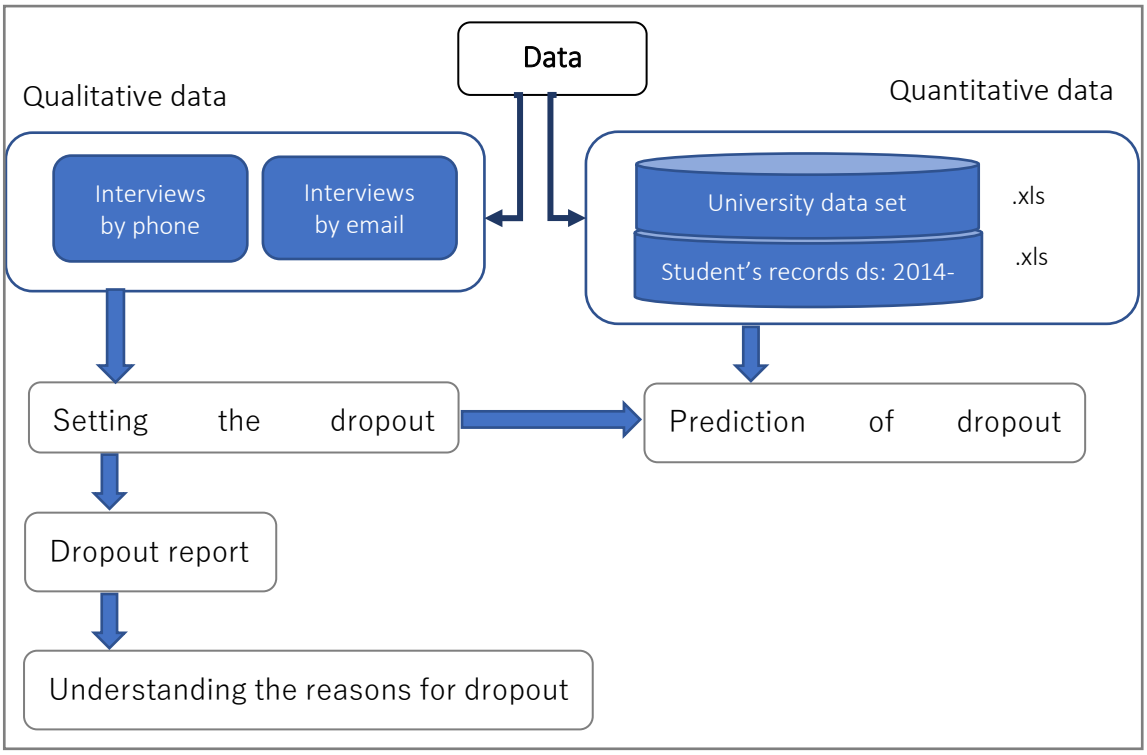


Figure 9 – Data by their type, source and purpose, used in this research

Source: Author

The current Higher Education Act and its annexes for the period 2007 and 2018, as well as the University Regulations for the period 2007 - 2018, were analyzed to understand the data. A student may lose student status if: a) he/she completes his/her studies, b)

he/she discontinues them at his/her own request (request to discontinue), c) he/she has not enrolled for the current academic year and does not have passive status, d) he/she is de-registered due to a disciplinary action, e) he/she has not completed his/her studies for 12 years. The student has the right to carry student status for 12 years and to change to part-time or full-time study ones for 12 years. Due to the vagueness of the law and university regulations, the student may retain student status beyond these 12 years and enjoy the benefits of student status until 30 years old (health care, student discount, and transportation discount). The new Study rules are applied from the school year 2022/23, and since data set in this research is ranged 2007-2018, those new Study rules are not consider.

4.3.1 The qualitative data: questionnaire and interviews

This research tend to analyze the reasons behind student attrition at UNIBL. Prior to this, there are no available data or reports why students leaving the HEI in B&H. This section comprehends the structure of the online questionnaire distributed to students who churned by own request, and the interview done by phone or email with students who have empty status in the database and by the date of interview did not graduated. Since we used those interviews data to set and support the definition of churn at the quantitative data set, we introduced here the quality data first.

Due to the inability to split students into those who quit HE permanently and those who quit but continued HE in the university database, we performed the questionnaire, which was distributed to the students who dropped out by own request and interviewed 10 students who did not enroll at all in some of the school years within their education. The following section provides insight into the data we collected while performing quantitative research.

Students who leave the UNIBL by their own request - the on line survey:

The questionnaire was distributed by email on March 6, 2022, to students who dropped out of their studies at their request. The purpose of the survey was to determine how many students who leave the university at their own request continue their education. This share is used latter to set up the dropout definition.

We also wanted to determine the reasons for dropping out and find out what happened after students left the faculty which was their first choice. The survey structure, with questions and answers is shown in Table A3, in Appendix.

Only 20 percent of the email addresses of all students who dropped out at their request were available in a database, Table 11. Responses were collected over 90 days.

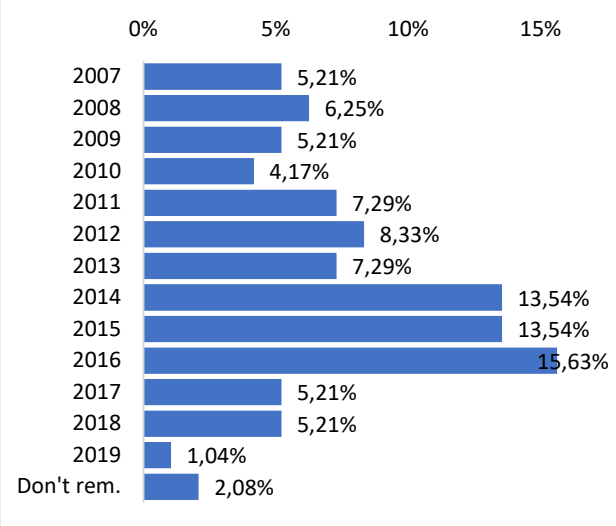
Table 11 – Available email addresses of all quitters and the number of dropouts by own request, by first enrolled year.

Year of the first enroll	No. of email addresses recorded in a database of students who drop out by own request and those who never enrolled (drop out by law)	No. of students who left UNIBL by own request
2007	0	44
2008	0	97
2008	0	253
2010	1	209
2011	2	220
2012	32	231
2013	53	418
2014	166	399
2015	227	429
2016	411	538
2017	361	347
2018	459	381
Total:	1,712	3,566

Source: Author.

A total of 321 e-mail addresses could not be delivered for various reasons. After 90 days it was collected 96 respondents.

Data description:



Faculty	Size
Faculty of Law	17.7%
Faculty of Economics	15.6%
Faculty of Architecture, Civil Engineering and Geodesy	13.5%
Faculty of Philology	12.5%
Faculty of Natural Sciences and Mathematics	11.5%
Faculty of Political Science	9.4%
Faculty of Electrical Engineering	8.3%
Faculty of Philosophy	5.2%
Academy of Arts	3.1%
Faculty of Medicine	1.0%
Faculty of Agriculture	1.0%
Faculty of Technology	1.0%

Figure 10 - Distribution of the dropouts by year of enroll (left), and by faculty (right) in the survey at UNIBL. Sample size 96.

Source: Author's contribution

The gender structure of the respondents shows that 23 percent of the respondents are men and 77 percent of the respondents are women, the sample size of 96. Most of the respondents dropped out of their studies between 2011 and 2016, as shown in Figure 12. The university has 17 faculties and in the survey sample were 12. The most numerous are students of Law, Economics and Faculty of Architecture, Civil Engineering and Geodesy, while Medicine, Agriculture, and Technology are the list represented in the sample.

Not enrolled students - the interviews:

Semi-structured interviews were conducted via telephone and email to better understand the instances of dropout that occur when students have not graduated and had a blank status in a database. The goal was to gather information about whether students re-enroll (come back) to the university, the reasons for having a blank status in the database (a situation that occurs when a student has not enrolled for the next year of school and does not have passive status), and their plans regarding their education. Respondents were selected from 10 different faculties between the 2010 and 2014 generations. Four students were surveyed by email, and 6 students were interviewed by phone during the second half of March 2022. A clear pattern in responses was found: Regardless of the reasons for non-enrollment, students who succeeded in enrolling in their final year and had few exams left to pass expressed a strong willingness to continue their studies. This survey pattern and the average length of the study were the basis for determining the rule of which students are considered as churn and which are not.

The rule which classify student as not churn if he/she has only few unpassed exam and is on the last study year or advanced (absolvent) student may underestimate the number of churned students.

4.3.2 The quantitative data: University database

The following section explains in detail the quantitative data, i.e., dataset obtained from UNIBL, regarding its collection, description, exploration, missing data, and preprocessing stages according to CRISP-DM methodology.

4.3.2.1 Initial data collection and description

Data were provided by the University Computing Centre of the University of Banja Luka in two Excel files in Cyrillic. The first file contained demographic and enrollment data of students since 2007-08 up to the 2018-19 school year (Table 12, Table 13).

Table 12 – Student demographic and enrollment data, 2007-2018

Database content	Demographic and enrollment data of students
Time frame	2007-08 – 2018-19 school year
Format	.xls
Received	Via university email
Date of receiving	3rd October 2019
Size	61379 KB
Database size	48.072 rows, 116 columns, 5.576.352 cells
Features	Total 116

Source: Author

Table 13 - Variable description: Demographic, pre-enrollment and enrollment data

Variable	Description
faculty:	Name of the faculty, text, 17 unique values.
sprogram:	Study program (or module), text, 230 unique values.
indexs:	Id of the student, system-generated number, numerical, 8 digits. Representation YYYYBNNN, where the first 4 digits are the year of study entering, the second digit represents bachelor (1), master (2), and Ph.D. (3), part-time students (5) last three digits are the number of students at the faculty evidence in the year of entry.
altid:	Alternative ID of the student, string (numerical data, divided with '/').
birth:	Date of birth, string, format DD.MM.YYYY.
gender:	Text, 2 unique values.
email:	Email of dropped students, text.
birth_p:	Place of birth, text, 569 unique values.
birth_c:	Country of birth, text, 51 unique values.
citizenship:	Citizenship, text, 37 unique values.
habit_p:	Place of habitation, text, 297 unique values.
habit_co:	County of habitation, text, 125 unique values.
habit_c:	Country of habitation, text, 20 unique values.
hs:	Name of the high school, text, 391 unique values.
hs_degree:	High School degree, name of profession which student obtain finishing secondary education, text, 158 unique values.
score_t:	Total entry score, total score at entry exam plus high school score (based on average high school GPA), numerical, ranged 0-100.
score_e:	Entry exam score, numerical, ranged 0-50.
date_g:	Graduation date, the official date of public defense of a thesis. String: DD.MM.YYYY.
duration:	Study duration, string, representation YY-total number of years of study, MM- total number of months, and DD, the total number of days of study, 00 hours, 00 minutes, 00 seconds from the 1st September of the enrolled year until the date of graduation.
The following variables are study data by school year and are repeated in each school year from 2007-08 to 2018-19:	
time_2007:	How many times did the student enroll in his/her current academic year during the 2007-08 school year. Range 1-10, numeric. (Example: In the 2007-08 school year, the student enrolled for the first time in his/her sophomore year. The value is blank if the student was not enrolled in the 2007-08 school year).

Variable	Description
sy_2007:	Year of education in the school year 2007-08. Value range: 1-5 (undergraduate programs last three and four years and undergraduate programs in medicine lasting 4, 5, and 6 years). Value 10 - Advanced, student who has completed the exams of the last academic year. These students have resumed their final year of study. Numeric, 6 unique values. The value might be blank if the student was not enrolled in the 2007-08 academic year.
status_2007:	The student may have one of five statuses with respect to their financial obligations to the university, defined by low: 1-Part-time student (pays the total amount of tuition), 2-Self-funded (pays less than part-time), 3-Student foreigner, 4-Co-funded (partially funded by the state), 5-Scholarship student. Text, 5 unique values. Value is blank if a student was not enrolled in the academic year 2007/08.
type_2007:	The student can have one of 5 statuses regarding his/her activity in the current academic year: 1-dropout, 2-discontinuation (passive year), 3-transfer from another faculty (validated), 4-normal, 5-study program change. Text, 5 unique values. Value is blank if the student was not enrolled in the 2007/08 academic year.
profile_ac_2007:	The abbreviated name (code) of the program and study group for which the student enrolled in the 2007/08 school year. String, 319 unique values.
profile_fn_2007:	Full name of the study program and module the student enrolled in during the 2007-08 school year. Text. Unique values, 219.
ects_2007:	Total number of ECTS credits earned in the 2007/08 school year. Numeric.
npe_2007:	Number of courses passed in the school year 2007/08. Numeric.

Source: Author.

The second file contains the examination records of the Faculty of Economics and the Faculty of Law during 2007/08 – 2018/19, in Cyrillic. Other faculties are not included because they do not have continuous data entries in the database. Since exam records were unavailable for the data used in ML modeling from 2007 to 2013 — due to the fact that exam records only began in the 2014/15 academic year — these records are presented in the Appendix, Table A4 and A5. Although they were not included in the main analysis, they were utilized in a presentation at the REDETE 2021 conference, which zoom in at faculty level of prediction, and hold potential for future research applications.

Summary of categorical variables is provided in the Table 13, through the count of unique values in description.

4.3.2.2 Data exploration report

To explore our data, we started at the dataset of 37,667 students, i.e., that was used for introducing the dropout variable, and encompasses all students who enrolled into their freshman year since 2007 until 2018, for categorical and numerical variables. This dataset had outliers included, due to fact that variables with outliers are not part of dropout definition.

The total gender share of enrolled freshmen was 60:40 in favor of women (more detailed table is in Appendix, Table A6). The share of graduated students in the total amount of enrolled students is 17.3 percent. The 78.9 percent of students are B&H residences, but this data is underestimated due to significant amount of missing data of students' origin, Table 16.

Table 14 – Share of students by the country of origin, 2007/08-2018/19 school year at UNIBL

Country of origin	Number of students	Percent
B&H	29,727	78.9%
Serbia	112	0.3%
Montenegro	16	0.04%
Other	3	0.01%
Missing (lack of label)	7,809	20.7%
Total:	37,667	100.00%

Source: Authors contribution.

Geographical analysis shows that 1/3 of enrolled students come from Banja Luka, where the University is located. Also, the Banja Luka's bordered municipalities in RS entity, generated the highest number of enrolled students, comparing to the rest of the country, Figure 11.

Based on geographical data, it was possible to create two new variables: the one who shows how far is the UNIBL of student's place of habitation and the second that tells how much is the each municipality developed.

The first new variable introduced by geographical data was distance from Banja Luka (*dist*) and analysis by distance from the UNIBL. The distance range was checked for each municipality in the country, in Google Maps, selecting the shortest route. The distance range was 0-347 kilometers (km), Figure 14.

The second variable derived from geographical data was level (degree) of development of the municipalities (*mld*). Since B&H is divided into two entities and Brcko District, each entity has its own categorization of development of the municipalities. Federation entity has methodology and ranged municipalities in 5 categories of development. RS entity ranged the municipalities classified in 4 development categories by their budget, every year.

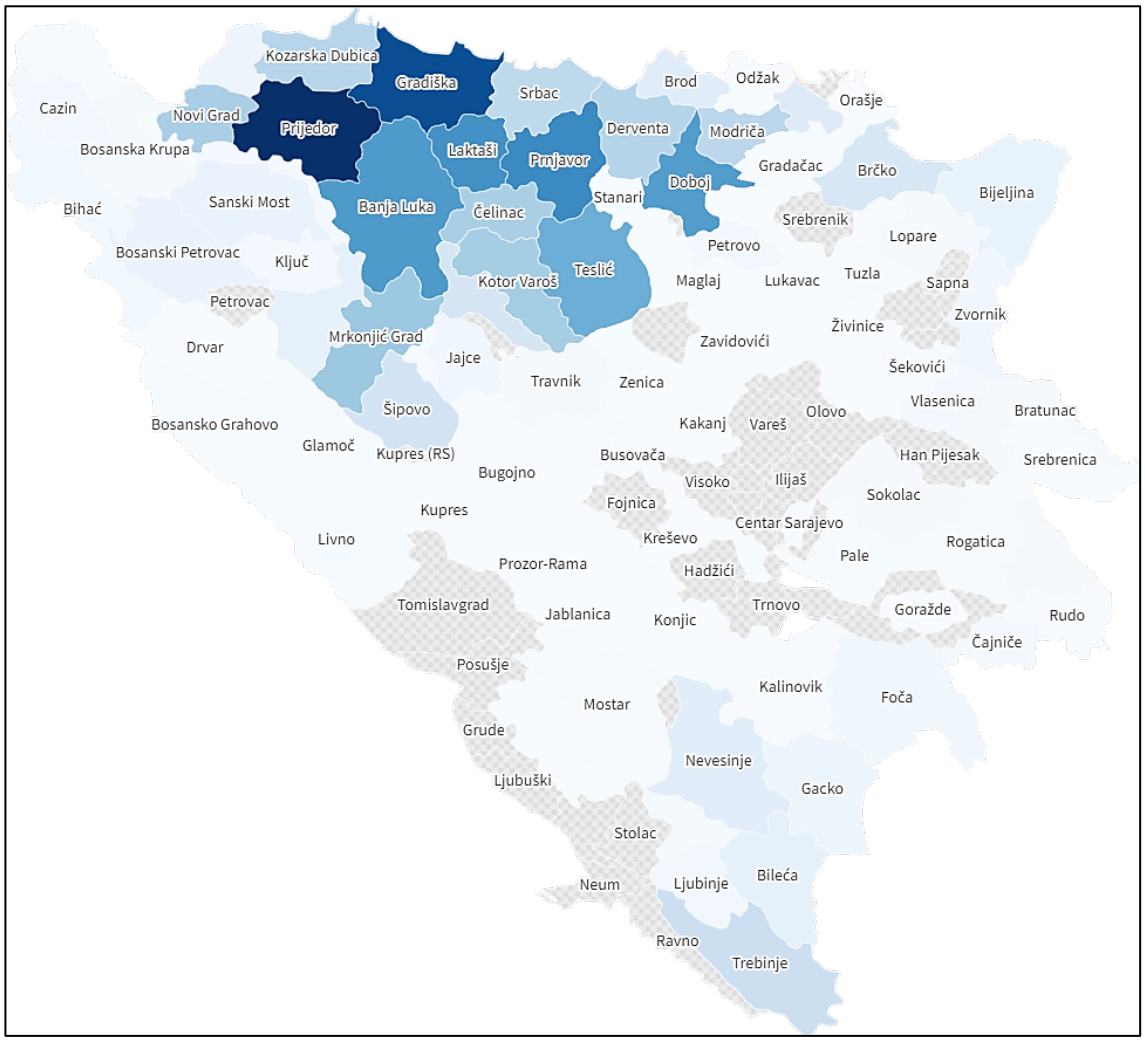


Figure 11 – Spread of enrolled UNIBL students in the country, 2007/08-2018/19.

Source: Author's contribution.

Combining those two approaches, the analysis by municipality development is presented in the Figure 12. The large portion of students from Banja Luka (that is categorized as developed municipality), creates impact to the data representation in the Figure 12. The detail data of number of students by municipalities are in Appendix, Table A7.

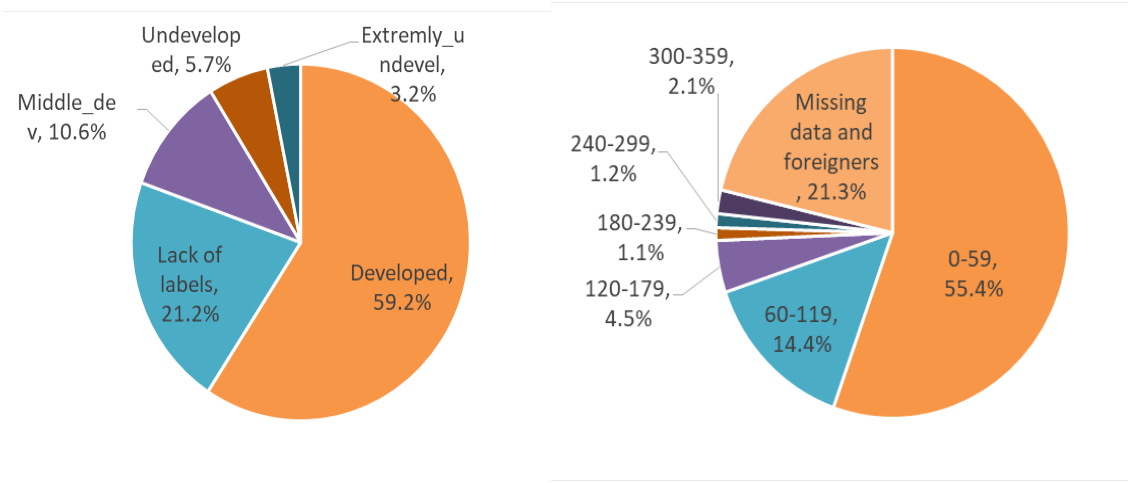


Figure 12 – Share of enrolled students into freshmen year, by municipality development (left), and by distance from the UNIBL in kilometers (right), 2007/08-2018/19

Source: Author's contribution.

The majority of students come from Gymnasium, STEM, and Economics high school, Figure 13. Almost ¾ of enrolled students attends the 4 year (8 semesters) bachelor duration study. The bachelor of 5 and 6 years long belong to the Faculty of Medical.

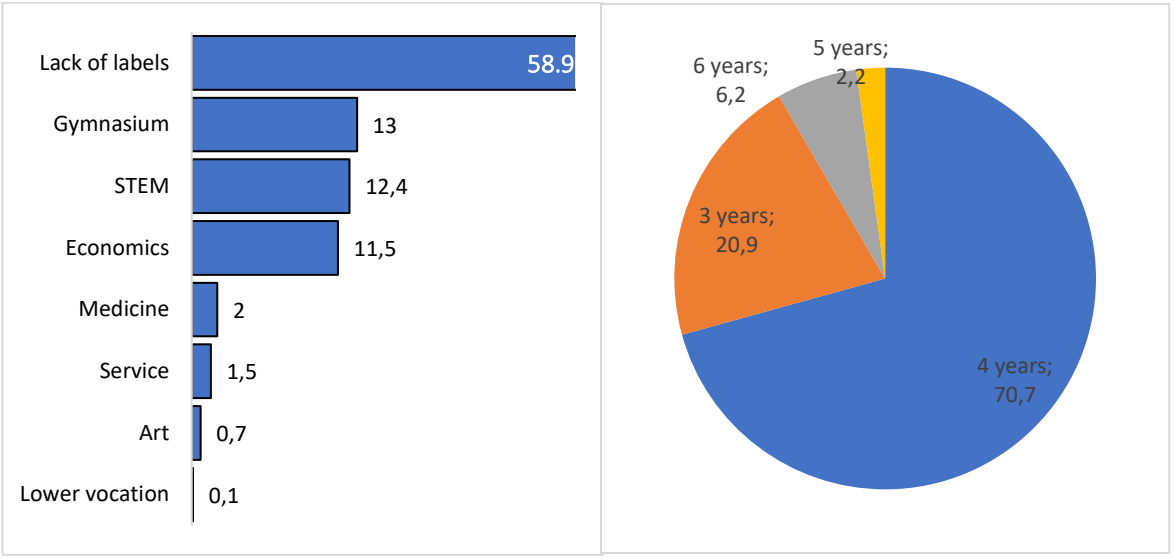


Figure 13 – The share of enrolled students into freshmen year at UNIBL by high school degree (left), and study duration (right), both in percentages, sample size 37,667.

Source: Author's contribution.

Within those 12 years of collected observation, the STEAM and social students carry almost the same portion of enroll (Figure 14). Top 5 most popular faculties are: Philosophy, Natural Sciences and Mathematics, Medicine, Economics and Law.

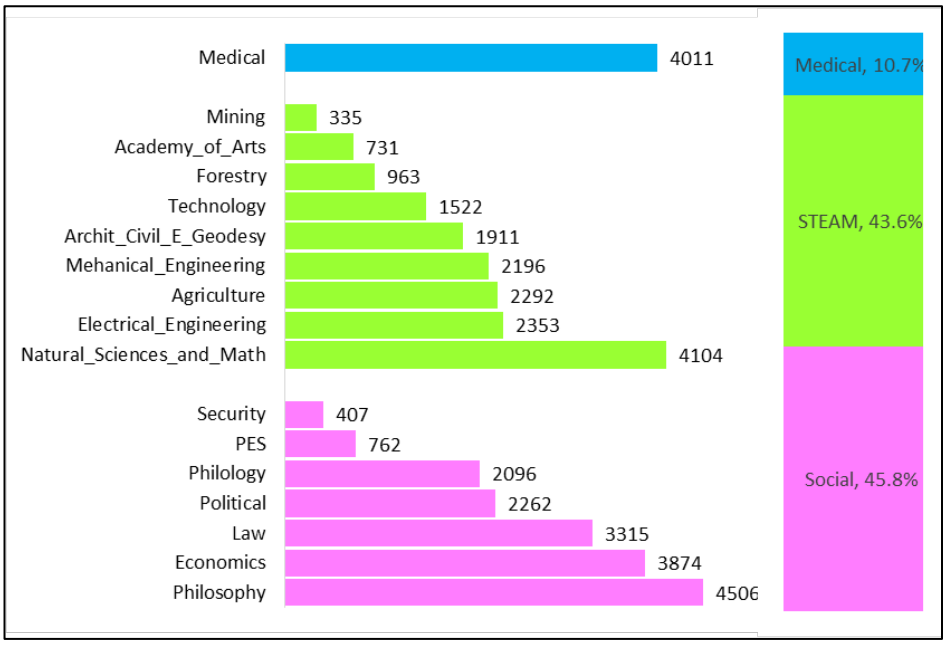


Figure 14 – UNIBL, enrolled students by faculty, and science area, 2007/08-2018/19

Source: Author's contribution.

The majority of students have "normal" type of enroll into their freshmen year, which means they don't belong to other entry categories (Table 17). The share of transferred students is very low, as well as those who make their freshmen semester/year passive, and transferred as type "Acknowledged from another faculty". The total dropout by students request in their freshmen year is 9.5 percent within 12 years of observation. This number had strong incline up to 2016-17 school year. After that the dropout by own request in freshman years decline. Also, students did not change the study program on the freshman year up to 2017, according to university data base.

Table 15– UNIBL, share of enrolled students by type and status at enrollment (in percent), sample 37,667, since 2007/08-2018/19

Type of enroll	Share in total enroll
Normal	89.0
Dropout	9.5
Acknowledged from another faculty	1.1
Passive	0.3
Transferred from another faculty	0.2
Status of enroll	**Share in total enroll**
Co-financing	56.1
Scholarship holder	40.1
Part-time	2.7
Foreigner	0.7
Self-financing	0.4

Source: Author's contribution. (More detailed tables are in Appendix, Table A8 and A9).

The dominant entry status for freshmen is co-financing, that means the government pays part of the admission costs, while there is 40.1 percent of scholarship holders which pay 42 Euros per year. University data shows interesting fact: between 2011 and 2014 it was three times more students enrolled as part-time students than before and after those years. Also, the number of foreigner students grows since 2014. (Appendix, Table A8 and A9). Examination of the numerical variables in the dataset of 37,667 students reveals outliers (due to human error in data entry) and valuable information about the scope of those variables through summary statistics (Figure 15).

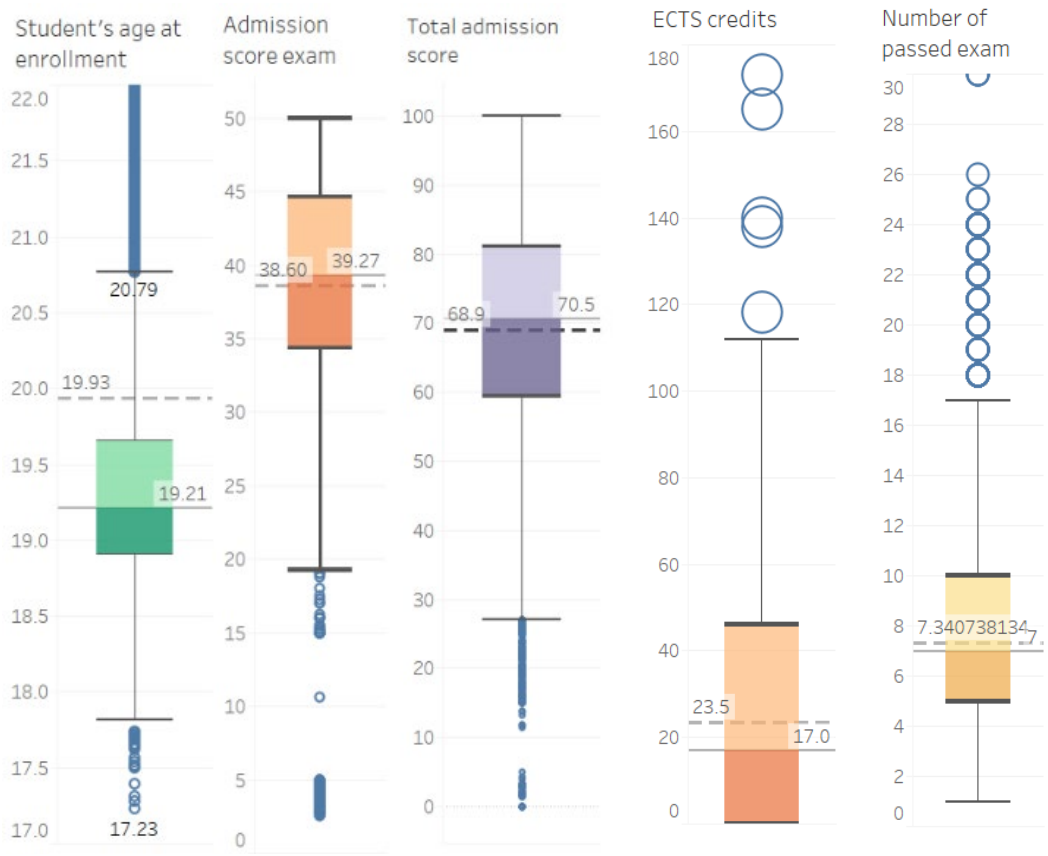


Figure 15 – Box plots of numerical variables in the university dataset that is used to estimate dropout, 2007/08-2018/19, at UNIBL

Source: Author's contribution.

The Figure 15 shows box plots, together with (dashed line) average, median, 1st and 3rd quartile, and boundaries of 1.5 of standard deviation. The outliers are presented by blue circles outside the border of 1.5 of standard deviation. In this stage the outliers are not removed. The average freshman at UNIBL is around 20 years old, has admission exam

score around of 39 points, and total entry score of 69 (of 100) points, while at the end of its first university year collects around 23 (of 60) ECTS credits, and passed in average 7 exams.

It is not possible to trace the student's success at the end of semester, since there is only available variable of total ECTS successfully collected at the end of each school year. Based on available data, Table 18 shows the ranged ECTS credits at the end of freshmen year at UNIBL by number of students who obtained those credits. Around one half of students succeed to obtain by 20 ECTS credits, and almost 40 percent had zero credits. Part of explanation is that not all faculties recorded those ECTS credits in the university data base, and there is zero value by system default for every student.

Table 16 – Ranged ECTS successfully collected at the end of freshmen year, 2007-2018, (in a percent).

Freshmen year	0 ECTS	1-20 ECTS	21-40 ECTS	41-60 ECTS	60+ ECTS
2007	52.9	59.2	10.1	26.8	3.9
2008	52.3	63.5	9.9	25.9	0.6
2009	43.1	54.0	15.4	29.7	0.9
2010	40.4	49.5	12.8	36.4	1.2
2011	40.6	51.0	12.0	36.3	0.7
2012	39.1	50.2	13.5	35.3	1.0
2013	36.3	50.6	13.5	35.3	0.6
2014	35.6	52.7	14.5	32.6	0.2
2015	37.0	53.0	13.5	33.3	0.2
2016	38.4	55.8	13.8	29.8	0.5
2017	25.0	40.6	15.0	44.1	0.3
2018	28.4	45.4	24.0	30.4	0.2
Total number of students:	14,719	19,630	5,227	12,497	310
Share (%):	**39.1**	**52.1**	**13.9**	**33.2**	**0.8**

Source: Author, based on university data set.

The summary table of all numerical variables with summary statistics is provided in Appendix, Table A10 on the dataset of 37,667 students.

4.3.2.3 Data preprocessing report

In the university dataset preprocessing steps, one might perceived three different turning points. The first one is reserved for dataset obtained in its row (original) format from the UNIBL, which contained 48,071 students (Table 19). After removing students who began their studies before the 2007/08 school year, as well as masters, and doctoral students, a

total of 78 percent of the original records remained (37,667). The dropout variable is introduced on this dataset. The dataset ready for ML modeling contains 43 percent of students recorded in the raw (original) dataset.

Table 17 – Preprocessing and transformation steps done at university dataset.

No.	Description of the preprocessing steps	Type of added variable	Dataset size, number of rows and columns (missing values in percentages)
I	**Raw data set obtained from UNIBL**		**48,071x120 (56 percent of missing values)**
1	Removed master, part-time and PhD students		43,420 x 120
2	Added variable *MAX_value*. Last year of study enrolled by a student (values ranged 1-5; 10 for absolvent year)	Numeric, float64	
3	Added variable *ECTS_total.* Total number of ECTS credits collected during the bachelor study	Numeric, float64	
4	Added variable *science*: Social science faculties, value 1: Faculty of Economics, Law, Security Science, Political Science, Physical education and sport, Philology, Philosophy. STEAM science, value 2: Academy of Arts, Architecture, Electrical Engineering, Mechanical engineering, Agriculture, Natural Science, Mining, Technology, Forestry. Medicine, value 3: Faculty of Medicine.	Numeric, int64	
5	Removed students enrolled in their freshmen year before 2007/08 school year		37,667x125
II	**Dataset for estimation of churn students**		**37,667x125 (54 percent of missing values)**
6	Added variable *enr_1st_year*: Year of first enrolling. Year of first-time student entered the university. Value range: 2007-2018.	Numeric, float	
6	Removed rows for students which *duration* variable was higher than 12 years		37,663x125
7	Added variable *study_duration*: Bachelor full-time study duration at University (by multiple conditions: by faculty, study program, enrolled year): 3 and 4 years, 5 and 6 (some study programs at Faculty of Medicine).	Numeric, float	
8	Added variable *absolvent*: Advanced student: a student is advanced if the subtraction of *study duration* and *MAX_value* is less or equal to zero.	Boolean	
9	Added variable *graduated*: True if a student has recorded the date of graduation in a database.	Boolean	
10	Added variable *dropout_by_request*: Dropout by student's request. True if a student has "Dropout" in the *type_* of enroll variable during the study.	Boolean	
11	Added dummy variables for each generation and year after enrollment year which represents dropout by law. True if a student does not graduate, does not drop out by own request, is not an advanced student, has empty status in a database in n+1 year after year of observation (n).	Boolean	
12	Added variable *dropout_final*: True if a student is dropped by request or without request (by law).	Boolean	
13	Removed abnormal student's paths.		37,640x207

No.	Description of the preprocessing steps	Type of added variable	Dataset size, number of rows and columns (missing values in percentages)
14	Added variable age1: Freshmen's age at the entrance of the university, as subtracted of 1st September of enrolled year and date of birth.	Numeric	
15	Added variable age2: Student's age at the date of graduation: as subtract of date of graduation and date of birth.	Numeric	
15	Recode the municipality of student's origin. The number of unique municipalities in the B&H before recode was 210. After recode it decreased to 111 unique municipalities.		
17	Added clustered variable *mld*: municipality level of development for places in the B&H. Since B&H is divided into two entities and Brcko District, each entity has its own categorization of development of the municipalities. B&H Federation entity has clear methodology and ranged municipalities in 5 categories of development. RS entity each year publishes the municipalities classified in 4 development categories by their budget. Values of the *mld*: 1 – developed, 2 – middle developed, 3 – undeveloped, 4 – extremely undeveloped, None – for students from Serbia, Croatia and lack of labels in the variable *habit_p*.	Object	
18	Added variable *dist*: the distance in kilometers from the town Banja Luka where the UNIBL is placed to each municipality in the country, based on shortest way on Google Maps. Variable has None values for foreign students and for lack of labels in the variable which represents the municipality of student's origin.	Object	
19	Added variable *hsd*: high school degree type clustered from the *hs_degree* variable. The 152 unique values of the *hs_degree* are categorized in following values of *hsd* variable: Gymnasium, STEM, Economics, Service, Lower vocation, Medicine, Art and None for lack of labels.	Object	
20	Added variable *ECTS_categories*: clustered variable generated from *ects 1* (total number of ECTS credits collected by student in his freshmen year). Values: ECTS_<=20, ECTS_20_40, ECTS_40_60, ECTS_>60.	Object	
21	Transformation of the columns in the data set 2007-2013, where each enroll year has data of the next 6 years after enrollment		22,728x169
22	Removed all unnecessary dummy variables		
23	Removed 5 and 6 years in *study_duration* variable.		20,814x81 (25 percent of missing values)
24	Categorical variables attributes in Cyrillic are renamed in English		
25	One hot encoding for all categorical variables		
26	Outliers check for numerical variables *age1, npe 1, ects_1, score_t, socre_e* due to their contribution importance to the model.		
27	Removed outliers		20,754x75
III	**Dataset ready for ML modeling**		**20,754x75 (3.3 percent of missing values for numerical variables)**

Source: Author's contribution

During the preprocessing stages we added dummy and auxiliary variables which helped us in managing of data transformation and served as control check within coding, which is evident in the table ubove.

4.3.2.4 Missing data report

To follow the CRISP-DM methodlogy, the missing data reports were created. Missing data report was done three times, following the three datasets presented in the Table 20 (above).

I Raw data set obtained from UNIBL:

It was purples to run in detail a missing data analysis for received dataset in its original state due to the structure of the data, which requred the data transformation. To portrayed the challenge of data transformation, here is presented the missing data report on raw data (Figure 16, left).

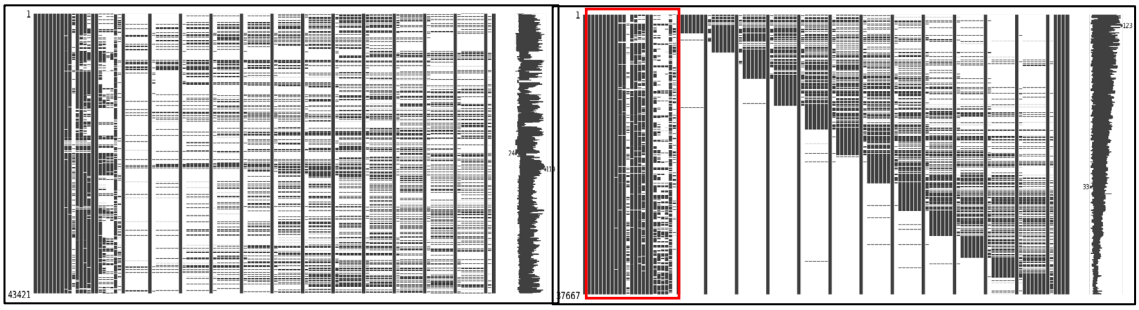


Figure 16 - Missing data and its presentation on raw data (left). Missing data and its presentation on data set for churn estimation (right).

Source: Author.

We reported only total share of missing data at this stage (raw data) – 56 percent of missing data. The missing data analysis was done in detail for the second and third (II and III) datasets size, shown in Table 19. Some variables in the dataset were set in a way to do not have the empty cells if the administration stuff do not fill the student's data. By database system defolt in the case of „ECTS collected at the end of study year“, all empty fileds had value 0, that is defolt by system.

II Data set for churn estimation:

The total share of missing values in this dataset was 54 percent of all data. However, having in mind that our dataset currently is partialy transformed (Figure 16, right) and

still not ready for ML modeling, those high percentage of missing values are far overestimated. In this stage, for the variables (2007-2018, Figure 16, right) was not possible to calucalte the missing data due to its current structure.

The Figure 16, right, shows more realistic missing data report for the numerical and categorical features of students, that are recorded for each student and do not depend on study year. (Those features are highlighted by red coloured rectangle on Figure 16, right).

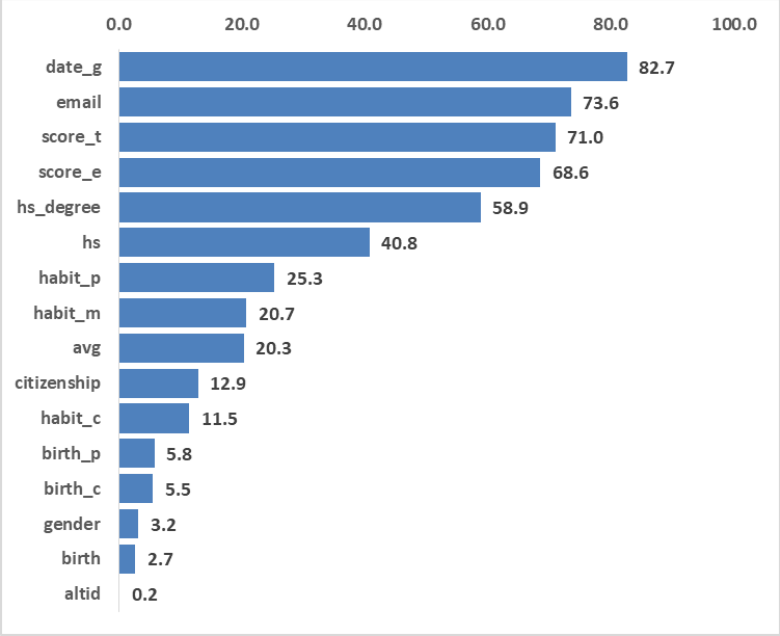


Figure 17 – Percentage of missing data per variable in the dataset for student's churn estimation (sample size 37,667)

Source: Author

Still there is a challenge to calculate how many missing variables and data we have for ECTS credit amount collected at the end of each year, due to fact that system „puts" zero value by defolt for each record (student) if administration stuff do not change it.

III The data set redy for ML modeling:

Due to One Hot coding of categorical variables, it looks like we remained without huge amount of missing data. However, some „new" variables like *hs_missing, dist_0, mld_missing* represent the lack of data for variables they were derived from. The reason why we left both gender variables in modeling set is the lak of labels for 5.2 percent of gender. Also, there is a lot of missing data which „disapeard" due to One Hot categorical variable coding for the high school degree name, Figure 18.

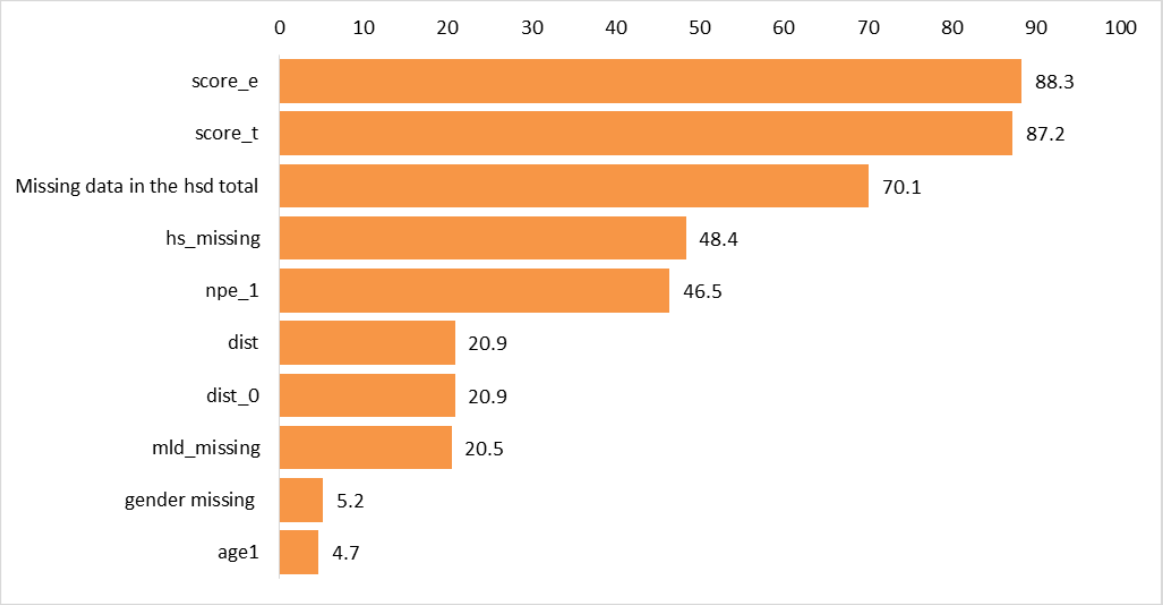


Figure 18 – Percentage of missing data per variable in the dataset ready for modeling (sample size 20,754)

Source: Author

Still, our final modeling ready dataset contains only 3.3 percent of only numerical missing data.

CRISP-DM is a document heave methodology, which is affirmed also in this thesis, through all its phases. From understanding the UNIBL rules, laws, and cases, data understanding where was provided a help of the UNIBL Computing Center stuff and Head of the Student Administration Service at Faculty of Economics, through all the reports of data collection, preparation, description, modeling and evaluation. More than 80 percent of time was invested into data understanding and transformation, i.e. preparation for modeling.

4.4 The feature engineering

For the feature engineering phase we employed several techniques, relying on the literature review: correlation by Pearson, Spearman, Kendall and Phi coefficient, Logit model, Permutation importance, and importance by SHAP. The purpose of using all those pre-hoc, and post-hoc techniques was to compare their results, to ensure that only the most important variables are used in the final dataset, to compare the importance of variables through the time by different models (from pre enrollment data up to end of freshman year), and to find the new knowledge of how model decides who will dropout and why, in three different time intervals.

Correlation: Feature or dimensionality reduction is the process of removing variables in data set from several reasons. The first one is multicollinearity: the case when there is a very high or perfect correlation among independent variables (predictors), where we remove one of two such kinds of variables to be able to perform modeling. Otherwise, we won't be able to generate the singular matrix of the model. Each time we had multicollinearity among predictors, i.e., we removed the one which showed a less strong correlation with the target variable. Another reason for the feature reduction process is to simplify our ML model, so it uses the minimum resources regarding the computational power and the minimum resources regarding the number of independent variables for optimal performance.

Pearson correlation coefficient was used since we have independent continuous numerical variables in the data set (age at the time of enroll, admission score, GPA from high school). Since our final data set used for ML modeling does not follow Gaussian distribution, i.e., it is not normally distributed, which is one of the assumptions of Pearson correlation, we added *Spearman* and *Kendall's* correlation coefficients, as a robustness check of Pearson correlation, due to large data set size. The literature supports the usage of *Phi correlation* when we have a dichotomous variable, i.e., Boolean, so it is added as an additional dimensionality reduction tool.

Logistic regression (Logit) model: Since our model has a binary target variable, and the most of the predictors are binary, we used a logistic regression model to inspect the variable importance prior to ML modeling phase and support results gettering by correlation in two ways. First, we considered the influence of each variable on the logit model as model coefficient, then we check which variables were significant to the model, and finally, we checked for the importance of each variable by logit model and its power/contribution to predict class 0 (non-dropouts) or class 1 (dropouts).

Permutation importance: The permutation importance (PI) is a model agnostic method to measure the impact (ramifications) to the model output by each single predictor variable by permuting its values (i.e. by shuffling it, while keeping all other variables order the same). One variable is important to the model if its remove or shuffle their values decrease the model performance, and opposite: if the PI of some variables shows negative value, that variable values shuffling does not affect the model outcome in any way and it is considered as not important. The PI shows the importance of the variable for a given model, and not its predictor's power.

The PI was employed by using Python library scikit learn after the models were feed with predictor variables and measured their performance. All models were checked for Permutation importance and based on those results, selected top N variables for final model at the end of freshmen year data (Table 40). The PI was calculated at three different time moments, each time by adding the more variables which were known in particular time (Table 20):

1) Before enrollment, i.e., pre HE data (27 variables, in total). The 26 variables were binary, while there is just one, ent_1st_y contained the years from 2007 up to 2013.
2) At the enrollment (added additional 12 variables which were known at the time of enrollment into freshman year in enrollment week). There is 10 binary variables and two more (*score_t, score_e*) were float. Due to many missing labels, they did not included into input features for RF, SVM and NN.
3) At the end of first (freshmen) year is added 8 variables more to the model, where six were binary and two float. The number of passed exam, *npe_1*, has 60+ percent of missing values and was not included into input features for RF, SVM and NN. Additional: Using the final set of all predictors, the only top N were selected in order to improve the models performance by cutting less important features.

The results were compared on both: train and test sets, where the average importance obtained by calculating the 10 times (repetitions) of results was presented on the box plot charts and in the tables. In the shuffling process was used the accuracy scorer, the same measure as for the ML model performance.

Table 18 – Number of predictor variables added with the time in each phase of prediction, and interpretation technique

Model	Pre-enroll	Enroll	End of 1st year	PI	SHAP
HGBC	27	27+12*	39+9**	Yes	Yes
RF	27	27+10	37+8	Yes	Yes
SVM	27	27+10	37+8	Yes	Yes
NN	27	27+10	37+8	Yes	Yes

*In enroll set of predictor's variables there are two variables with 80+ percent of missing data: *score_e*, and *score_t* which were not applied to RF, SVM and NN since there was not specified the algorithm of dealing with missing data. **At the end of first year there is a variable number of passed exam, *npe_1*, which also contains more than 60 percent of missing values.
Source: Author.

The PI was calculated for all 78+11 ML models in three stages of prediction, for train and test set, and all results are saved in internal document, while in Appendix, in Tables A24-A29 are reported for all models in three stages of time for test set.

SHAP importance: SHAP importance was evaluated at three different time points for each model, resulting in a total of 267 SHAP feature importance charts, which are stored in an internal document. In the Appendix, Tables A24-A29 are presented summarized results for the test set only. The SHAP charts, both at the global (model) level and individual level, are included in the text exclusively for the best-performing model.

All predictor variables are listed and described in Table 21, which is separated in three segments, according to three times of prediction.

Table 19 – List of variables used for ML modeling with the time: in thee time intervals (pre enrollment, enrollment, and end of first study year).

No.	Variables	Short description	Time interval
1	ent_1st_y	Year of (potential) enrollment into freshmen year. Values: 2007-2013.	Pre enroll data
2	age1	Age at the time of enroll into freshmen year.	
3	gender_1.0	Gender: True if it is a male. (Bool)	
4	gender_2.0	Gender: True if it is a female. Here are present both genders, due to 5 percent of lack of labels in the data. (Bool)	
5	dist_0	Missing data in variable distance (dist). True if the student's place of origin is not recorded. Boolean.	
6	dist_up to 80	Distance in kilometers between Banja Luka and municipality of student's origin. (Bool)	
7	dist_between 81 and 160		
8	dist_between 161 and 240		
9	dist_more than 241		
10	mld_1	Municipality level development, ranged in four categories (1 – developed, 2 – middle developed, 3 – undeveloped, 4 – extremely undeveloped), and variable that carries missing data. (Boolean)	
11	mld_2		
12	mld_3		
13	mld_4		
14	mld_missing		
15	hs_gym	High school name (secondary education institution) is derived from the variable high school name, were was 48 percent of missing values. High school names are Gymnasium, Economics, or contains the name of the STEM area, Medical, has a lack of name label (hs_missing), or has name like: Mixed high school, or the name of some famous person (hs_o). (Bool)	
16	hs_econ		
17	hs_stem		
18	hs_medic		
19	hs_missing		
20	hs_o		
21	hsd_Art	High school degree: clustered into seven categories, to overcome the possible misleading due to high school name variable. We have here 71 percent of missing data but they are not separated into another variable. (Bool)	
22	hsd_Economics		
23	hsd_Gymnasium		
24	hsd_Lower vocation		
25	hsd_Medicine		
265	hsd_STEM		

No.	Variables	Short description	Time interval
27	hsd_Service		
28	score_t	Total score at enroll: admission score for faculties which have entry exam + GPA score from high school. Ranged: 0-100.	Enroll data (in enroll week)
29	score_e	Entry (admission) exam score. Ranged: 0-50.	
30	dur_3.0	Official study duration of the study program, that student is enrolled in. If it is True – it is 3 years long (has 6 semesters). (Bool)	
31	s_co-financing	Status at the enrollment into freshmen year. Student may have one of the following statuses at enroll into freshmen year: scholarship holder, partial scholarship holder (co-financing), pays full tuition fee for domestic citizens (self-financing), or have foreigner status (pays full tuition fee for foreigners). At the beginning of the year student may choose to have part-time study and then pays fee according to University rules for part-time study. (Bool)	
32	s_foreigner		
33	s_part-time		
34	s_scholarship		
35	s_self-financing		
36	t_enroll_from_a_sp	Type of enrollment into freshmen year has five choices. Enrolled from another study program; Normal enroll or Passive year. The next two types of enroll are not appear in the enrollment week. Due that reason we excluded them from the enrollment data and moved them below, to the data which appear during the school year and are known at the end of first year of study. (Bool).	
37	t_normal		
38	t_passive_year		
39	t_acknowledged_from_a_f	Type of enroll which occurs during the school year: Enroll from another faculty, and dropout. (Bool)	End of first (freshman) year data
40	t_dropout		
41	ects_1	Total amount of ECTS credits collected by student at the end of freshmen year.	
42	npe_1	Total number of passed exams by student at the end of freshmen year.	
43	ECTS_less than 20	ECTS score at the end of freshmen year ranged in four categories to overpass the huge number of zero values which was found in a data set. The variable which carries the zero values (ECTS_0) is not included. (Bool).	
44	ECTS_between 21 and 40		
45	ECTS_between 41 and 60		
46	ECTS_more than 60		
47	***drop_final***	***Target variable: 0 – non dropout, 1 – dropout. (Bool)***	

Source: Author.

4.4.1 The feature importance and reduction

To exclude variables that do not influence the classification of students as dropouts or non-dropouts, we performed the feature reduction. The feature reduction process had several steps due to the types of variables in the obtained dataset.

First, the qualitative, non-numerical variables were selected individually, One Hot encoded, and then their multicollinearity and collinearity with target variable was inspected. This process was done to have some insights into what one may expect of model evaluation using post-hoc techniques: PI and SHAP.

We selected municipality, i.e., place of habit, high school degree (vocation or profession), high school name, distance from the Banja Luka, and municipality development level as the variable which were One Hot encoded, and then the Phi's correlation was applied.
Due to weak correlation, the municipality variable was removed from further research. Still, we wanted to cluster them in some way to check the possibility of impact to the dropout. Thus why we proceed to variable *mld* – municipality level of development which is clustered representation of municipality variable[12]. Then we used Phi correlation test and results of correlation between degree of municipality development and dropout are in the Table 22. Results are showing the weak or non-correlation.

Table 20 – Phi correlation between municipality level of development (left) and distance from UNIBL (right) and target variable.

Municipality level of development	Target variable: dropout	Distance from UNIBL	Target variable: dropout
mld_1	0.129092	dist_0 (<= 80 km)	0.122809
mld_2	0.013155	dist_1 (81-160 km)	0.133488
mld_3	0.010155	dist_2 (160-240 km)	0.010995
mld_4	0.022112	dist_3 (>=241 km)	0.039878

Source: Author.

The second way to inspect impact of municipality to the dropout at UNIBL, was to analyses the distance variable (*dist*), Table 22, right. There was 21 percent of missing labels for *dist* variable. Those results again have weak correlation.
A new variable was created based on the high school vocation/degree, *hsd*, which exhibits a lower correlation with the target variable compared to high school name variable, *hs*, derived from the name of the school. The *hs* shows a higher correlation with the target variable

Table 21 - Phi correlation matrix between high school vocation (degree) variable and target variable.

Secondary (high school) vocation (degree):	Target variable: dropout
hsd_Art	0.01826
hsd_Economics	0.07928
hsd_Gymnasium	0.06307
hsd_Lower vocation	0
hsd_Medicine	0.02404

[12] The distribution was tested using Shapiro test[12] due to binary variables correlation. Statistical significance was 0.487 = probably not Gaussian distribution.

hsd_STEM	0.08813
hsd_Service	0.02056

Source: Author.

The results of correlation among predictors and target variable by different correlation coefficients are presented in the Figure 19. The correlation by Phi's coefficient is sorted by descending order and only values greater than 0.10 (by Phi's) are showed on Figure. The detail data of all variables are in Appendix, Table A11.

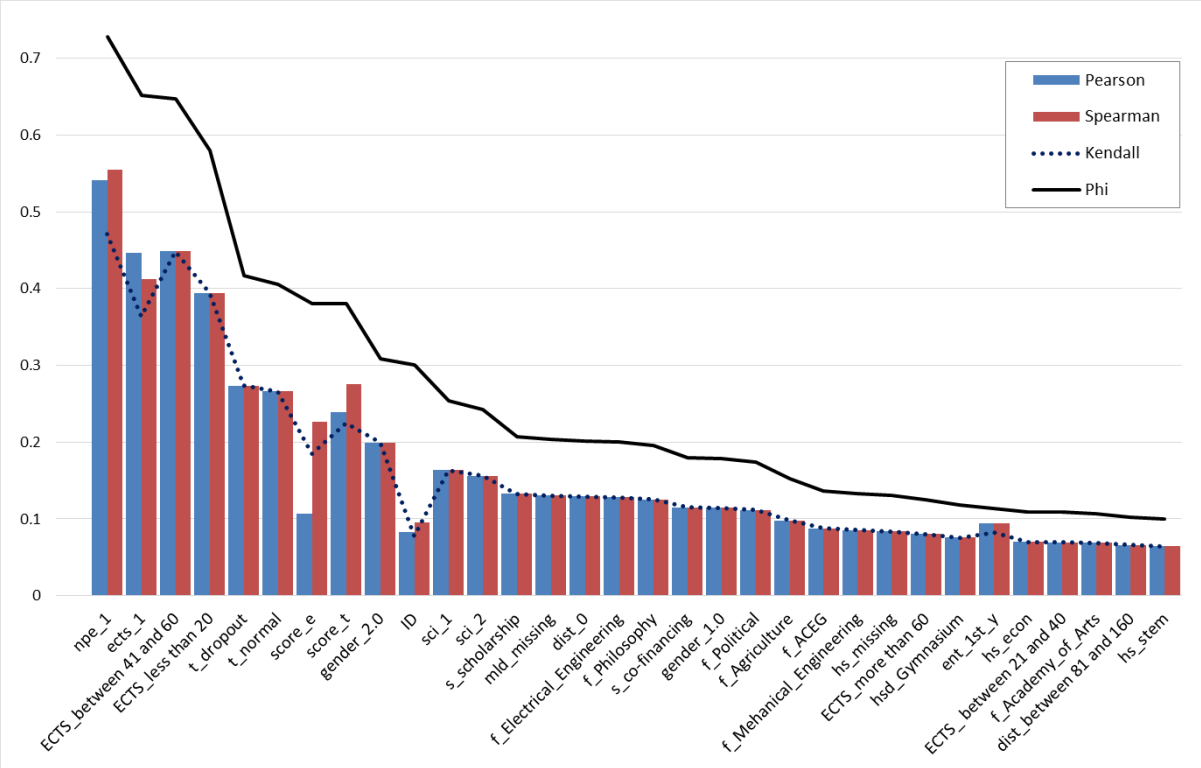


Figure 19 – Correlation with target variable by Pearson, Spearman, Kendall, and Phi correlation (Phi >=0.10, values are presented in absolute numbers).

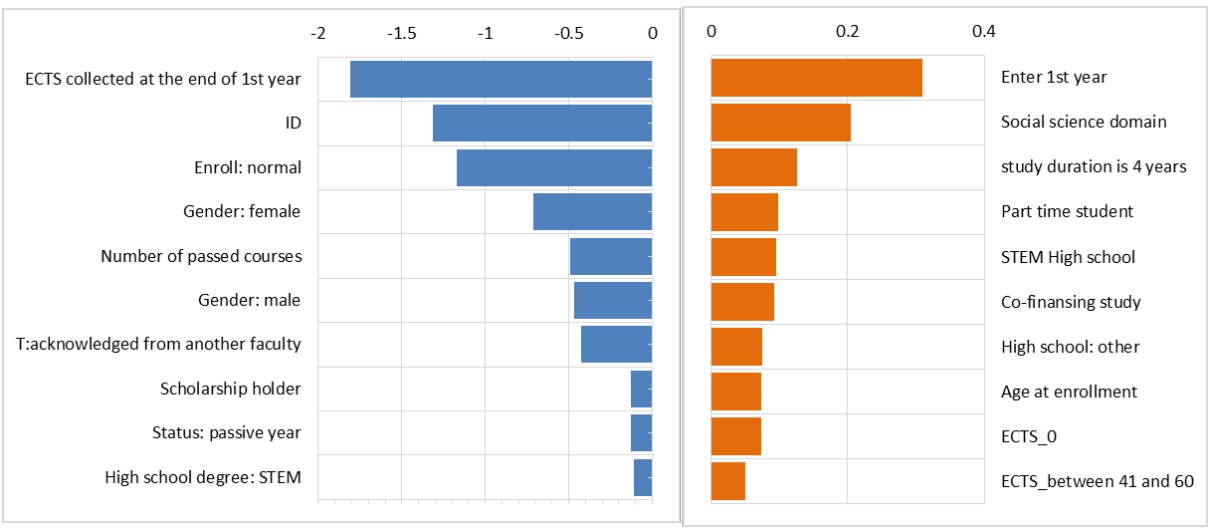
Source: Author's based on correlation matrix.

Pearson's and Spearman's coefficients showing very similar values, except for variable the entrance exam score (*score_e*), which has very low score by Pearson's coefficient. Generally speaking, all four coefficient yielded similar rank of variables, except for *ID* variable that is showing high score by Phi's coefficient. This variable is added at the beginning of the transformation process started at 1 and incrementally grows by the last record in the data base. Since the freshman students in the data base are listed by success of entrance exam and high school GPA in freshmen year by most of the faculties (although there is a lot of missing data in the variables *score_e* and *score_t* that encompass those values) it may be the case that this variable carry the part of information of student's

rank at the enrollment stage by study program and faculty, and that is why is important. Also, it is indicative that Phi's correlation bring the highest values than other three. Phi correlation is used as a feature selection technique because the target variable is binary, and the dataset contains many categorical features.

According the correlation stage, the variables that showed very weak or no correlation are listed in the Appendix, Table A11. The ones which tend to attract the more attention are ECTS score that is equal to zero (ECTS_0), because that variable was zero by default and for students who graduated, very often is zero, due to no-updated data in a database by university stuff especially between 2007 and 2011 school years. The self-financing or foreigner student status also do not showed correlation with dropout.

In order to support our feature inspection, we employed a logistic regression model to inspect the variable importance in two ways: as coefficients in a logistic regression model and using an odds ratio. The first approach results are presented in the Figure 20. The positive coefficient scores indicate a feature that predicts class 1, whereas the negative scores indicate a feature that predicts class 0. On the Figure 20 are presented top 10 variables (without faculties). The student has higher chances to be classified as dropout if he/she is enrolled in the last few years of university data for ML (2011-2013)[13], at some of social science faculties, is part-time student, comes from STEM or other (uncategorized) high school, is paying part of tuition fee (co-financing), and has zero or between 41 and 60 ECTS credits at the end of freshmen year.



[13] Enter 1st year (ent_1st_y) that shows differences across generations which entered the UNIBL since 2007 afterward.

Figure 20 - Feature importance according to a logistic regression model. Negative score (left) tend to predict non-dropouts, while positive score (right) tend to predict dropouts.
Source: Author's contribution.

In contrast, the student has higher chances to be classified as persistent or non-dropouts if he/she has some credits collected at the end of freshmen year and some courses passed, had lower ID at faculty level, is female and had standard (normal) status of enrollment, is scholarship holder, possibly is switching between faculties and has some of the STEM secondary vocation. The complete list of feature importance by logit model is in Appendix, Table A12.

The partial contribution of given variables where *p*-value is less than 0.05 is statistically significant to the model, but only in a current combination of those predictors. It does not necessarily mean we must drop variables that are not significant. Significant variables by logistic regression are: *score_t, gender_1, gender_2, score_e, ects_1, age1* and *npe_1*.

5 RESULTS: UNIBL CHURN CASE AND REASONS FOR DROPOUT

In this chapter we contribute to HE dropout report by presenting the estimated dropout rates at one public HEI in B&H, for the first time, on the sample of 37,667 bachelor students, according to definition of churn presented billow. Together with dropout report, for the first time in the country is done a survey of those who dropout, with the goal to understand reasons for leaving and what happened after leave.

The short introduction of UNIBL:

The UNIBL was established on 7 November 1975 in ex-Yugoslavia, in province Bosnia and Herzegovina, included faculties of Electrical Engineering, Technology, Mechanical Engineering, Law, Economics and three colleges. The rest of today's faculties are established as follow: The Faculty of Medicine 1978, Agriculture and Forestry 1992, Philosophy 1994, Architecture and Civil Engineering 1996, Natural Sciences and Mathematics 1996, Academy of Arts 1999, Faculty of Physical Education and Sport 2001, Philology, Political Science and Mining Engineering 2009, and the Faculty of Security Science in 2017. UNIBL today comprises 16 faculties, the Academy of Arts and the Institute of Genetic Resources, and offers 66 bachelors, 64 master and 13 PhD study programs. The University's offers two student's campuses with student's dormitories, cafeterias, sport space and University Computer Centre. The classroom area has around 16 thousand m^2 and 10 thousand m^2 of laboratories. Currently has around 15 thousand students from both entities and from abroad (University of Banja Luka, 2023).

5.1 Identified dropout types at UNIBL

According to UNIBL study rules from 23 October 2009[14], article 34, the student's status terminates according to the Law on HE in Republic of Srpska entity (RS), the Statute of UNIBL, and the statutes of faculties at UNIBL:

1) By graduation
2) By withdrawal (before graduation)
3) When student do not enroll into next year of study, and he/she did not ask for passive year/semester.

[14] Since October 2022 the new Study roule document is in effect where the graduation deadline is set as two times of bachelor study duration (absolvent years are not mentioned).

4) When student did not renew (re-enroll) the study year, and he/she did not ask for passive year/semester.
5) When student was expelled through a disciplinary procedure.
6) When student does not finish study within the official study duration time: the study duration time is equal to double time of bachelor study duration. This deadline includes the absolvent time.

Article 40 defines two years of absolvent time for bachelor study. The absolvent student's status gains the student who attended all the classes at the last year of study but did not pass all exams at the last study year. It means that for bachelor study of four years, student has two additional years to obtain the degree with absolvent status, but maximum (4+2)*2 years of time to finish its study. After 12 years student may switch to the part-time study if he/she did not finish its study. In every-day practice at UNIBL, which is in favor of students, a student who fell into (3) or (4) category above, may come back to study and continue without the standard enrollment procedure they had have when they entering the study for the first time, and without paying additional fee.

Dropout definition in our research is dependent on data source we obtain from UNIBL, current law, study rules, and permanency of decision to leave the HE. The goal is to distinguish between different types of churns, explain them and focuses only on students who have definitively terminated their schooling for good (Figure 21). The following attrition HE types that we identified in given data set are:

1) Total dropout at UNIBL is 47.1 percent within the 12 years of obtained data, as ratio of total dropout cases and total enrollment. It contains all following types of dropouts: permanent and temporary breaks of study due to study program change, passive school year status, etc. We are aware of total dropout undervalue, especially in the first few years of database usage (2007-2010) due to a lot of missing data.
2) Dropout requested by student is conditioned by database attributes. In a university database, until the 2019/20 school year, faculties changed the student's status to dropout only at the student's request. The percentage of students who dropout at their request is 22.5 percent from all enrolled students. This includes voluntary dropout like transfer to another faculty, or university, or leaving the study for

good. Involuntary dropout may occur due to financial or other reasons which are, together with voluntary dropout, presented in detail under chapter 5.4.

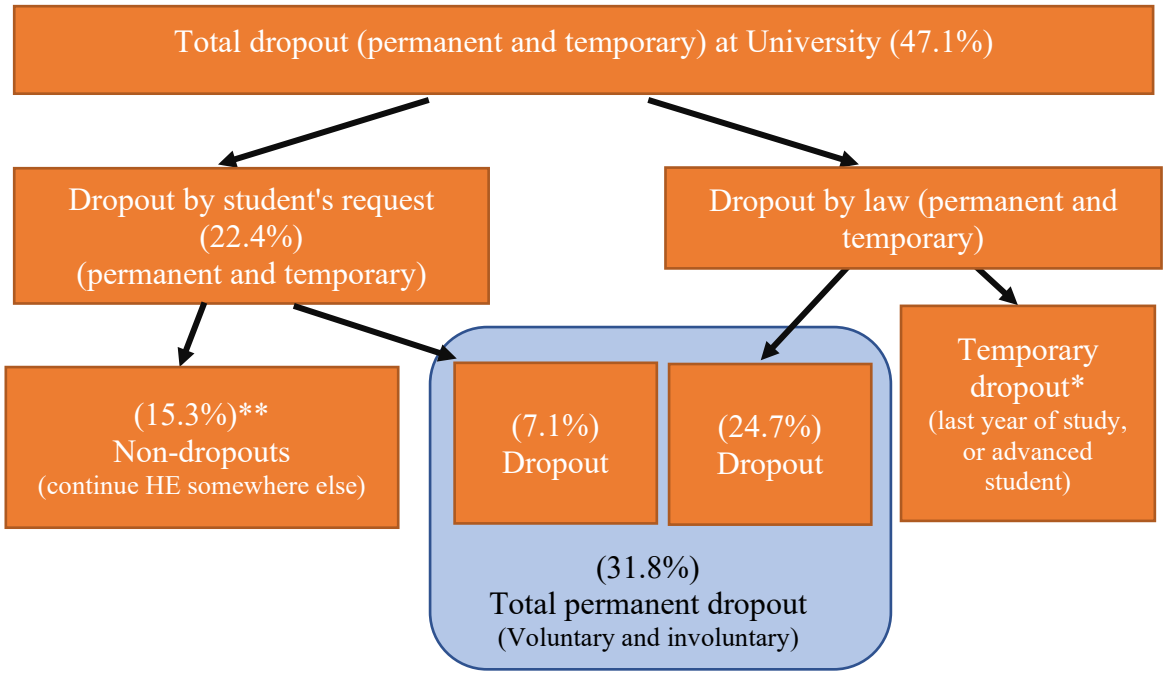


Figure 21 - Dropout definition in our research

*Source for potential underestimation of dropouts.
**Second source of potential underestimation of dropouts: Some of students who continue HE somewhere else also leave the HE.
Source: Author's contribution.

2.1) *Dropout (requested) by student as **permanent** HE leaves, i.e., dropout without graduation in HE*. We estimated that around 1/3 of all dropouts by student's request is pulled in this category. The estimation in more detail is presented in the chapter 5.3.

2.2) *Dropout (requested) by student who continue education at another faculty, or university in the country or abroad*. We estimated that around 2/3 of all dropouts indicated by student's request belongs to this category.

3) Dropout by university study rules (dropout by law), are the students from categories (3) and (4) of article 34. In the University's database the blank data field of student's status occurs every time when student does not change their enroll status at the beginning of the school year in accordance to the University's regulation, i.e., there is a blank (empty, missing) data field. In reality, it means that the student did not show up at all during the enrollment week in the fall or

spring semester, and the university staff did not fill any semester data for this particular student. This study break can be temporary or permanent.

3.1) *Temporary dropout by law*. To distinguish dropout by law which is temporary we did interview of ten students from this category which had "a break" for more than one school year in the database, and students who were at their last study year claimed the strong willingness to come back to study. This type of dropout is not calculated as the percentage of total enrollment, or numerically estimated. This part of dropout is *the source of underestimation* of the student's churn at UNIBL.

3.2) *Permanent dropout by law* are students which don't satisfy the condition above i.e., they have an empty data for any school year after the freshmen year and they have not enrolled into their last study year, 24.7 percent of all enrolled students.

5.2 The magnitude of dropout at UNIBL, 2007-2018

The total percentage of all dropout types (temporary and permanent) at UNIBL within 12 years of available data is 47.1 percent, and it includes students who dropped out by own request 22.4 percent (8,427) and those who dropped out for other reasons, 24.7 percent (9,302), Figure 22.

Of those who dropped out by own request, by our estimation, around 2/3 are transfers, i.e., continued HE at another faculty or another university in the country or abroad. The rest of the students who dropped out by their own request (the remaining 1/3rd), together with those who dropped out by law (24.7 percent), carry the 31.8 percent of quitters who left the HE for good, Figure 22. Those 31.8 percent of attrition at UNIBL within 12 years of available data is the subject of our research in the following section.

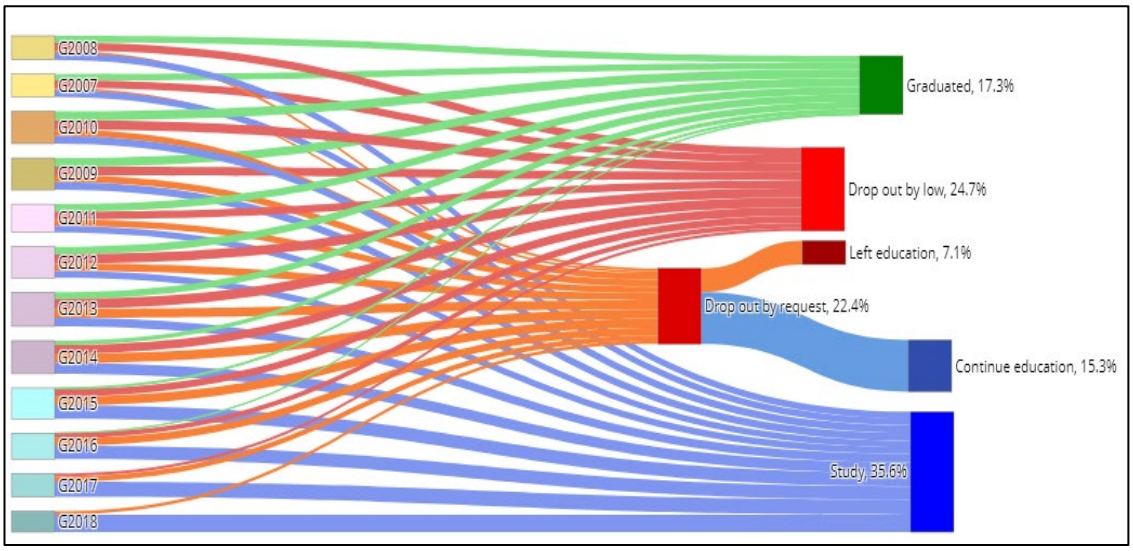


Figure 22 – Sankey chart: what happens with students after enrollment? Types of dropouts at UNIBL, 2007-2018, sample size 37,672.

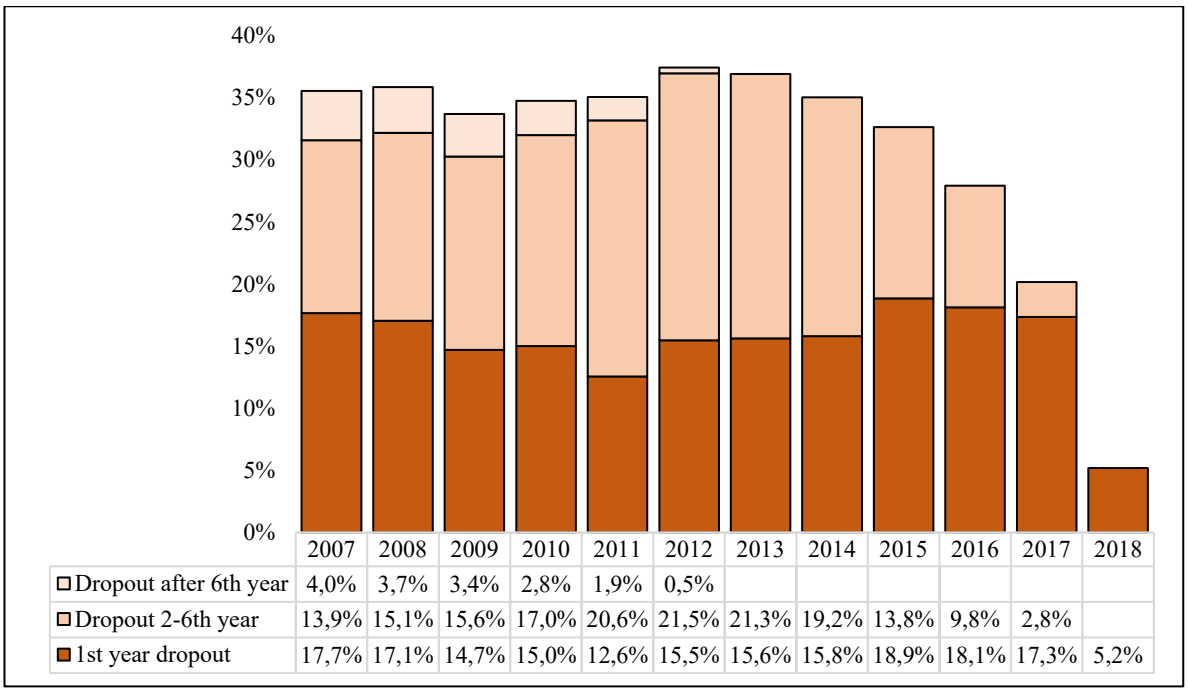
Source: Authors contribution. For details see Appendix, Table A13.

To illustrate the complexity of HE dropout data comparability challenge, we calculated attrition rates at UNIBL using the OECD definition, which was explained in chapter 2. We simply summarize dropout by law and by students request for the students who left HE for good, which is aligned with OECD dropout methodology, and for generations from 2007-2013, we got 85 percent of dropout instead of 36, by our definition choice.

Dropout by freshmen and sophomore years: One of the contribution of the research are churn rates calculated on the sample of more than 37 thousands of bachelor students, within 12 years. Sometimes this sample is a bit smaller due to missing data in the dataset (like for gender share of dropout, due to lack of 5.2 percent of gender labels).

This section presents dropout rates at UNIBL by year of university entry (generation, cohort) for students who permanently quit their HE (for more detail see Appendix, Table A14). The highest student's outflow occurs at 1st study year, Figure 23.

For the students enrolled from 2007-2010, the 1st year attrition made half of total dropouts, but in generations 2011-2014, more than half of students quit after the freshmen year, and as advance forward, due to shorter observed time frame, total churn rates decrease, Figure 34. Dropout in 2018, our last observed year are students who dropout by own request for good. Since 2011, the freshmen dropout slightly grows with each new generation, as well as the dropout at second, third and fourth year after enroll (more detail data in Appendix, Table A14).



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Dropout after 6th year	4,0%	3,7%	3,4%	2,8%	1,9%	0,5%						
Dropout 2-6th year	13,9%	15,1%	15,6%	17,0%	20,6%	21,5%	21,3%	19,2%	13,8%	9,8%	2,8%	
1st year dropout	17,7%	17,1%	14,7%	15,0%	12,6%	15,5%	15,6%	15,8%	18,9%	18,1%	17,3%	5,2%

Figure 23 – Dropout structure in freshmen year and following years at UNIBL, 2007/08-2018/19.

Source: Authors contribution.

Churn within time:

To achieve better understanding of the size of HE churn by time, we introduce the survival rates curve, named Kaplan-Meier curve. This curve shows the border between percent of students who enrolled into next school year and who churned compared to the total number of enrolled students. On the Figure 24 are presented survival rates within 6 years since enrollment in the UNIBL, by generations.

Kaplan-Meier curve shows that oldest generations of enrolled students had higher survival rates within the six years from enrollment. One of the possible explanation for such trend among oldest generations in the dataset is that those students did not have many job opportunities after graduation, the broad palette of scholarships abroad, "work and travel" while studying, online jobs and part-time jobs. Another possible reason is the local municipalities of student's origin politics due to student's scholarships, and available number of university scholarships. Some municipalities provided scholarships for all their students, while others were restricted to far a smaller number of scholarships. Difference in amount of scholarships by municipalities should be consider to. The further analysis of those financial sources for university students by municipality within 12 years period of time is necessary to better understand their impact to the retention rates.

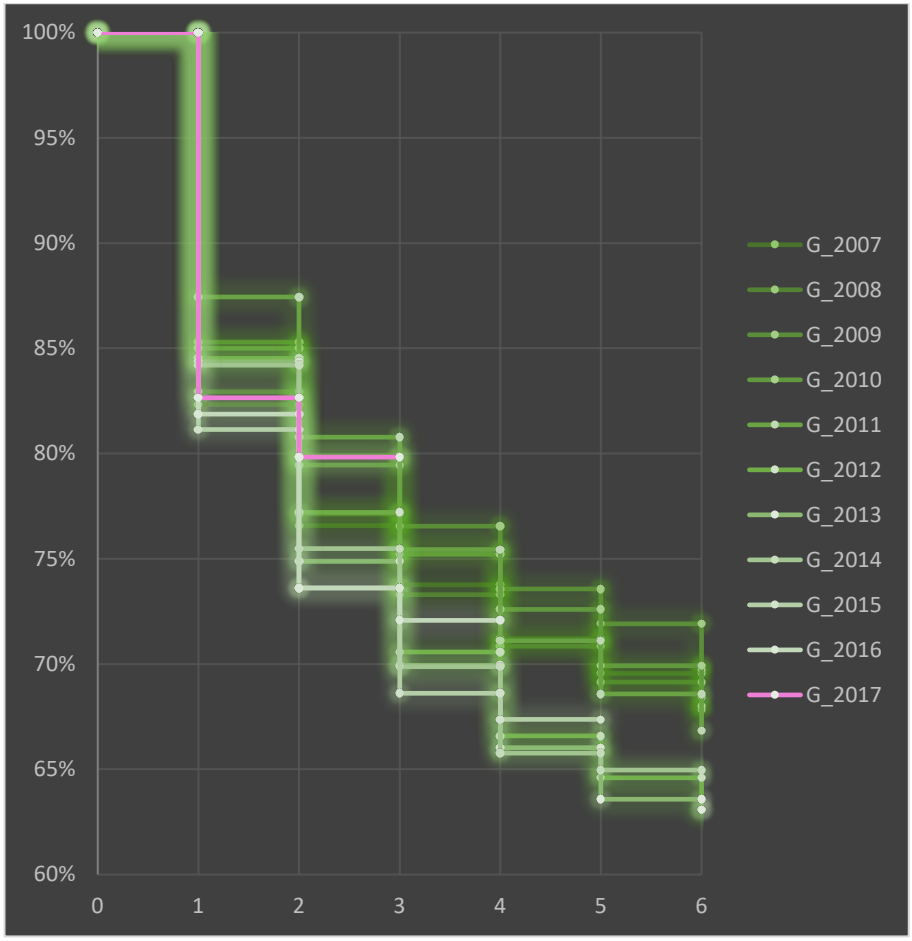


Figure 24 – Kaplan-Meier curve of dropout at UNIBL; 2007/08-2018/19.

Source: Author's contribution.

Churn by gender: The gender structure at UNIBL carries out 60 percent of female and 40 percent of male students, considering all enrolled students between 2007/08 and 2018/19 academic year. The breakdown of quitters by gender structure, including missing gender's label (Figure 25) shows that women drop out less frequently than men. Women are less prone to churn across all science areas, and by each observed time period (in more detail, see Appendix, Table A15, Figure A1). The total churn of men is 55 percent, of all enrolled men, while the women churned 39 percent.

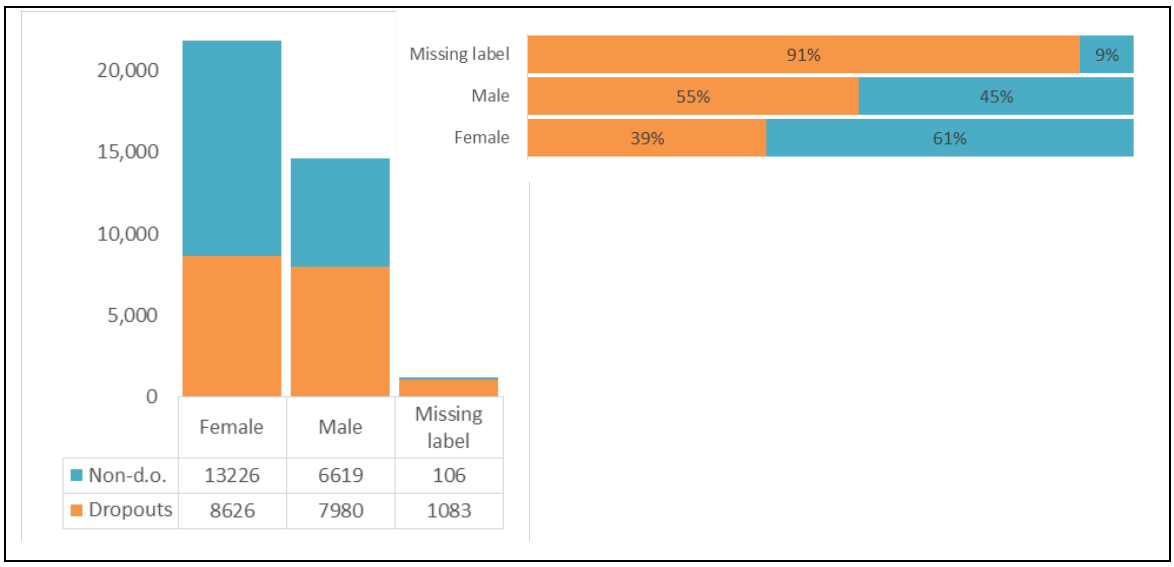


Figure 25 – HE permanent churn by gender, 2007/08-2018/19 at UNIBL.

Source: Author's contribution.

The relative difference between men and women continues to grow, and possible explanation is that women in B&H have less opportunities if they dropout, unlike male peers. The gender structure persistence in favor of women is interesting from social point of view, since psychological studies shows that higher educated women tend to find the partners who have the same or higher level of education, or earnings, while this is less important for the higher educated men, (Qian, 2017).

Churn by faculties and year of enrollment: Until the 2014 the churn rates had increasing trend. Possible reason is the broad time for churn estimation, which students enrolled since 2015 did not have in this research (Figure 26).

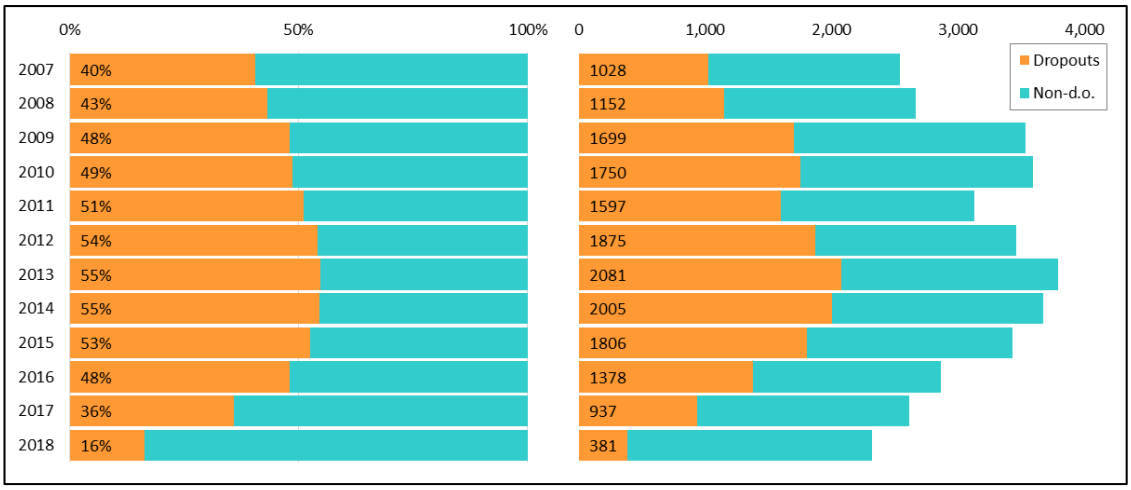


Figure 26 – HE churn at UNIBL, by school years, as the share of enroll students.

Source: Author's contribution.

Figure 27 shows total amount of churn (right) and enrolled students, and percent of churn by university unit (left) within 12 observed years. The highest dropout rates are among STEM group of faculties, and the top 5 are: Faculty of Mining, Mechanical, Electrical Engineering, Agriculture and Technology. The lowest dropout rates have Faculty of Security, that is the youngest established faculty (in 2017) and Academy of Arts due its small number of enrolled students and high admission criteria.

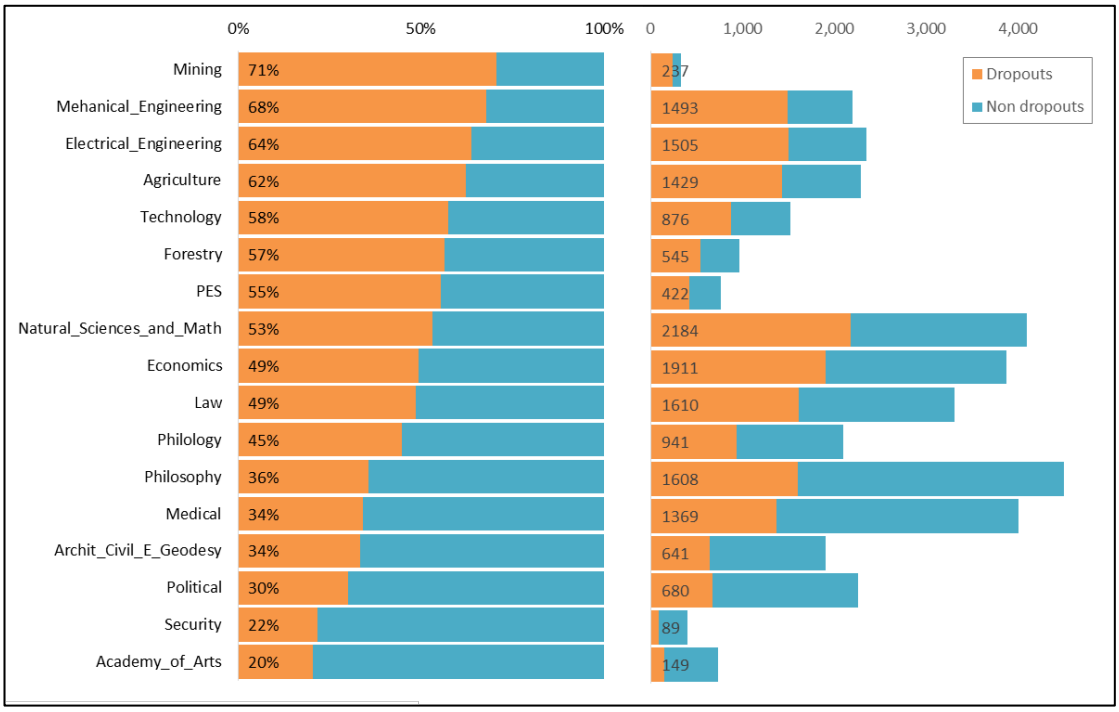


Figure 27 – HE churn, by faculties, within 12 years, as the share of enrolled students.
Source: Author's contribution.

Geographical churn structure: Having in mind that more than 11 thousand of students live in Banja Luka (according to 2007-2018 data set), that is classified into developed municipalities, we found following: a) the more distanced the municipality of student's origin is, the lower are dropout rates, b) with increase of municipality development level, churn rates increase too (Figure 28).

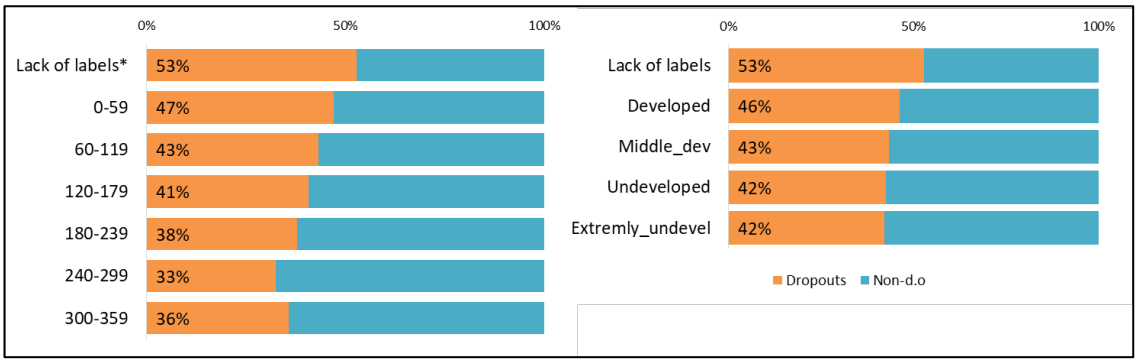


Figure 28 – Distance in kilometers from UNIBL (left), and municipality of student's origin development level (right) for domestic students at UNIBL, 2007/08-2018/19.

Source: Author's contribution.

Churn structure by high school degree and academic performance: The typical dropouts at UNIBL, within 12 observed years are a bit older than their non-dropouts peers (Figure 29), their admission exam score (36.6 at those who dropped out, and 40.1 for those who did not), and total enroll score (63.6 for quitters and 73 for non-quitters) is a bit lower than for non-dropouts.

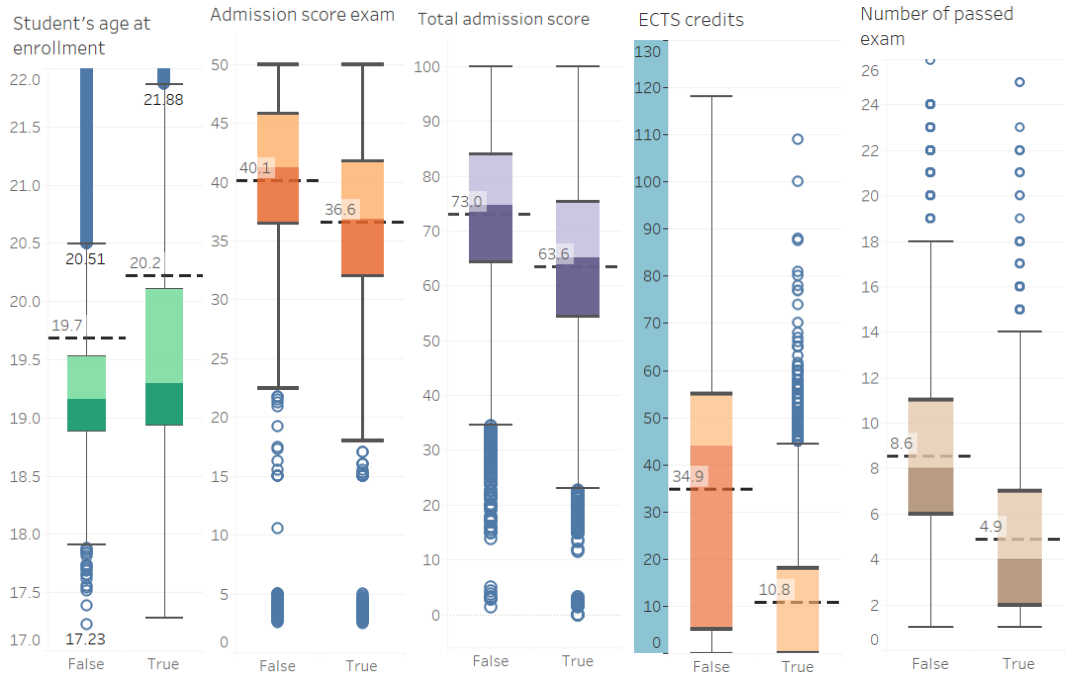


Figure 29 – Description of student's dropout (True and False) by numerical variables for freshmen, 2007/08-2018/19, UNIBL.

Source: Author's contribution.

The highest differentiation between churned and not churned freshmen is their ECTS credits collected at the end of first study year, what is average 35 for non-dropouts and

11 for dropouts. The variable "Number of passed exam" follows the established rule: the ones who churn at the end of first year have completed almost three courses less than they persistent colleagues.

UNIBL, dropout by science domain: The classification of UNIBL faculties by mayor science areas: social, STEAM, and medical is given in Appendix, Table A16.

The highest attrition is in STEAM, followed by medical and social sciences (Figure 30). The absolute size of dropouts by year of enrollment highlights the differences in science domain attrition. The absolute number of dropouts declines after 2014 (a gray area on the Figure 30) as a consequence of narrowed time frame of churn calculation (for example, the generation 2015/16 was represented within 4 years in a given university dataset, and the average time of degree at UNIBL is 5.2 years).

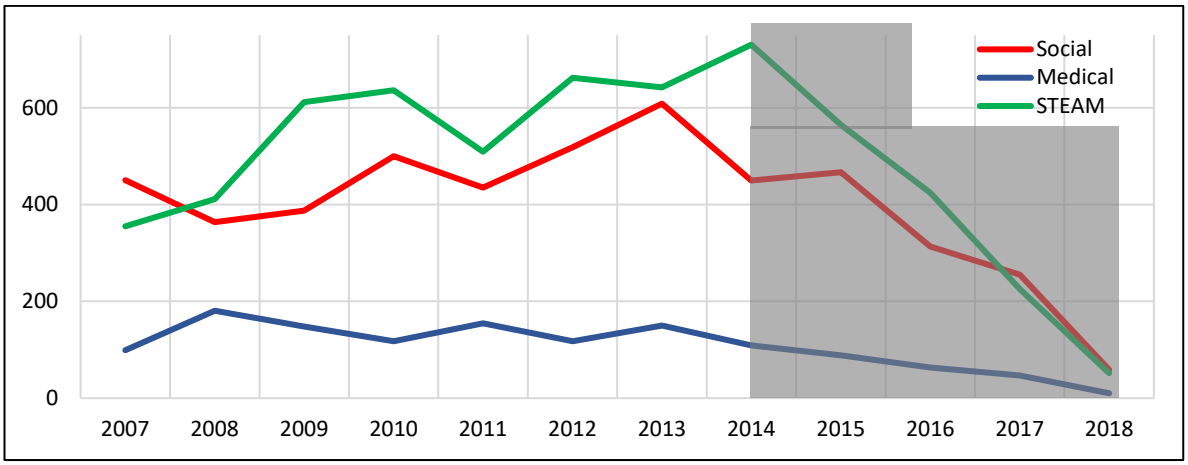


Figure 30 – The absolute number of students who dropped out at UNIBL by science area and generation, between 2007/08 and 2018/19 school year

Source: Authors calculation.

The average attrition for 16 thousands STEAM students is 35.5 percent, for 4 thousands medical students is 32 percent, while in social science for 17 thousands students is 27.9 percent (Appendix, Tables A17, A19, A21).

The STEAM students are the highest quitters, during the 1st and 2nd university year, since 2014. Possible explanation: the opportunities to work part-time, and full-time jobs while studying are higher for STEAM students, and researches confirmed that part- and full-time work has negative correlation with persistence in HE. Investing time in self-studying and unofficial self-education may lead to high earnings and push the decision to leave.

After second year spent at university, the leading quitters are social and medical students. The highest dropout amplitude occurs in the social science sphere. Yet those students tend

to have the lowest dropout rate during the whole study period. Dropouts by STEAM students, from one generation to another, ranged between 20.3 and 42.0 percent within six years from enroll (Appendix, Table A19).

There was relative more quitters among women within the first study year before 2013 (Figure 31). That trend changed from 2013. More and more women have degree in STEAM, since men leave the STEAM faculties more than women in total and by relative indicators. The attrition at STEM faculties grows more and more during the time, while the number of enrolled students decreases. This is an indicator for further analyses considers determinants before and during the university enroll.

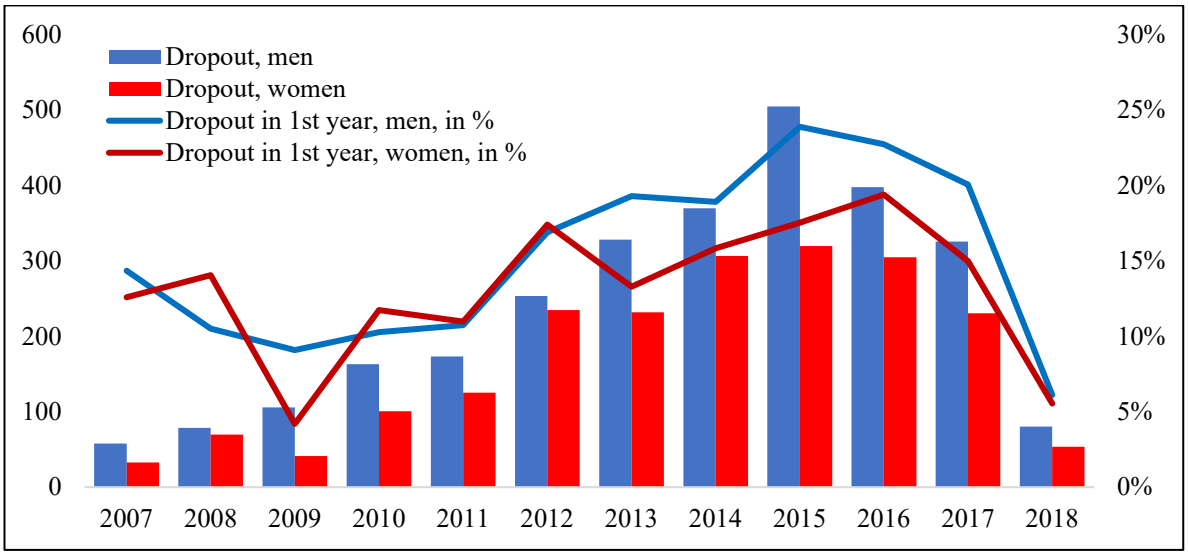


Figure 31 – STEAM students' dropout at UNIBL, 2007/08-2018/19, by gender, the total number (left axis, data by columns) and in percentages for freshmen year (right axis).

Source: Author's contribution (for detail see Appendix, Table A20)

The total share of dropout at STEAM faculties among gender: 34.0 percent of male and 29.2 percent of female students (Figure A2, Appendix). The average students' outflow during the four years from enrollment shows that half of churn happened within the freshmen year.

Students at the Faculty of Medical have the highest gender attrition among all other sciences. Dropout for Medical students shows that share of all male student dropout was 41.7 and for female students, 28.2 percent, within six years from enroll at UNIBL. The dropout rate was the highest for students starting 2011/12 and declines after (Figure 32). Unlike to STEAM faculties, the dropout rates tend to be more stable and even declines after 2011. The popularity of IT sector can explain part of the differences among dropout

trends in STEAM and medical science. When medical students drop out, he/she can't find job in similar areas, like STEAM and social students can. Due to specificity of medical labor market and medical regulations in the country (2007-2018), the major of medical students is employed in the public medical institutions. Also, the study of medicine is a kind of prestige.

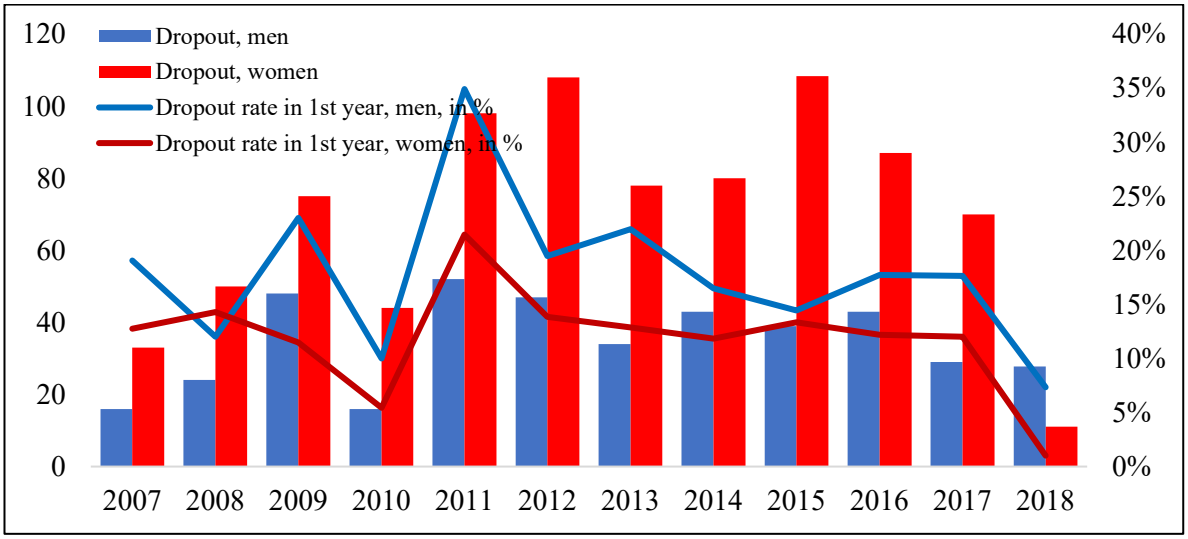


Figure 32 – Medical students' dropout at UNIBL, 2007/08-2018/19, by gender, the total number (left axis, data by columns) and in percentages for freshmen year (right axis).

Source: Author's contribution, (in detail see Appendix, Table A19)

The dropout trend in social faculties has more common with STEAM students, than medical (Figure 33). A far more men (as share of total enrolled male students) than women leave the social science faculties. Growing trend of leave is evident in all generations. Average men outflow from 2007-2018 is 36.8 percent, while women churn less, 23.6 percent. From one generation to another, dropout rate among social science students is in a range of 25-36 percent. From all male students enrolled in 2013/14, the 47.1 percent of male and 29.7 percent of women left HE for good.

Comparing the outflow of social science students by gender, it is evident that women dropped out more than men in absolute figures because the gender ratio is 70:30 in the advantage of women at UNIBL for social faculties. However, the relative indicators (Male and Female dropout, in 1st year after enrolling) show that men churn more than women.

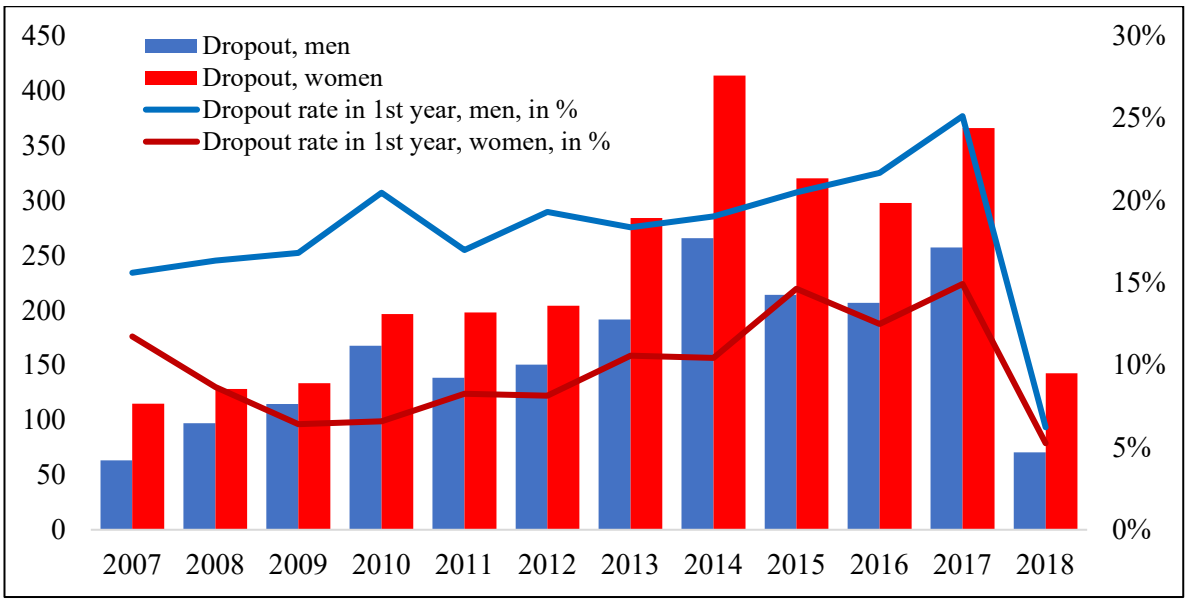


Figure 33 – Dropout of social science students at UNIBL, 2007/08-2018/19, by gender, the total number in each generation (left axis, data by columns), and dropout rate in 1st year (right axis, data by lines).

Source: Appendix, Table A18.

During the time, we witnessed the incline of the absolute number of students who leave the HE in each academic year. A possible reason is the increasing number of students with the intention of continuing HE abroad, prior thus, they often need to enroll at some university in the country.

UNIBL, dropout by bachelor study duration: At 16 faculties at UNIBL, there is 66 bachelor study programs, among them some have 3 (6 semesters) or 4 years (8 semesters) official study duration. The Faculty of Medicine has integrated programs which have 5- and 6-years duration. Due to small share of those students in total enrollment (8 percent), here we present only dropout by 3 and 4 years of bachelor study duration. The information of study duration is obtained from faculties' websites and phone interviews of administration staff. Some of study programs during 12 obtained years changed the study duration ones or two times. We coded each study program and presented the summary analysis on the Figure 34.

The highest quitters' rates come from 3 years study duration programs (Figure 34). The differences ware small in 2011-2014, but from 2015 there is a strong growing trend of dropout among 3 years study duration, while 4 years programs continued a bit slowly to grow. Since 2011 until 2017 the dropout growth 4 times at 3 years bachelor programs, and almost doubled at 4 years programs.

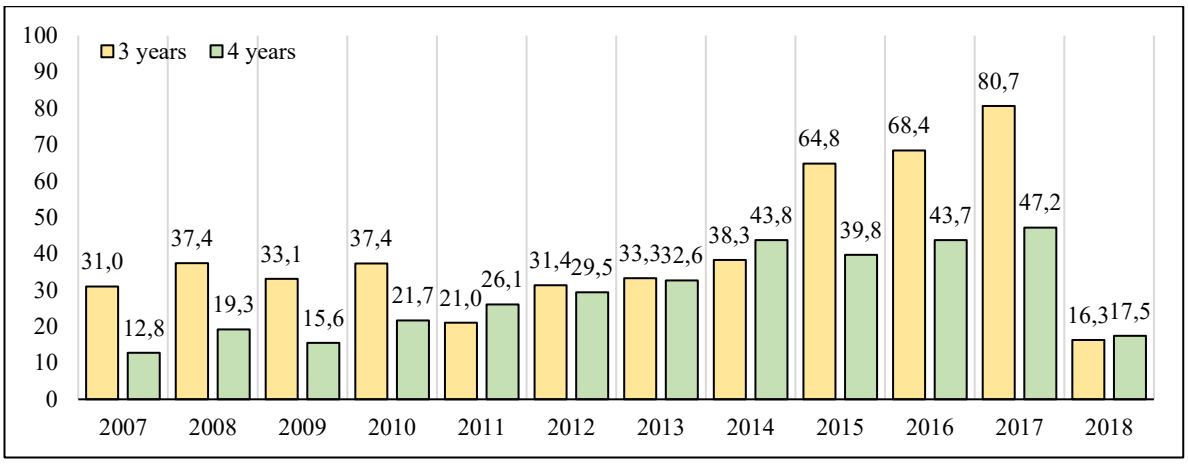


Figure 34 – Dropout at UNIBL, by bachelor study duration, 207/08-2018/19, in percentages

Source: Authors calculation.

Possible reasons may be teachers approach to students. Often happened that from few hundreds of students, only few of them passed an exam. The answer of the University to these situations was to introduce fee for taking exam after three times of not passing. Someone could think of teaching skills, abilities, and the way of interacting with students, as well. Online certification trends which are available to the most of the students make easier dropout decision. The broader research in this direction is needed to understand the phenomenon.

5.3 Reasons for leaving the UNIBL

According to answers of the respondents to our online survey, the UNIBL can retain at least 1/3 of students who churned by own request due to reasons which are directly influenced by university like:

- Dispute or conflict with the lecturer/professor/assistant
- Corruption, unprofessionalism, lack of objectiveness
- Lack of perspective and outdated, poorly organized study program
- Harassment by professor.

All reasons from above are stated in online survey of students who quit HE at UNIBL at own request.

In the university database, the field for reason of dropout does not exist. This brought us to the hard copy documents of dropouts. Each faculty unit has its own records (hard copies) of dropouts which consists of student's ID and student's name, date of birth, place of birth, date of dropout, number of exams passed, and reason for dropout. The field of

the reason for dropping out is usually blank or filled in with "personal reasons." Due to a lack of any research and data on churn reasons at University, we had to start our own source: online survey. We defined a survey distributed by email, in which we collected 96 samples between March 6, 2022, and June 6, 2022.

According to respondents' answers, the main reason for dropping out was dispute or conflict with the lecturer/professor (15.6 percent), followed by financial reasons (13.5 percent) and inability to work and study (13.5 percent), Table 24. Corruption, unprofessionalism, lack of objectivity (6.3 percent), and a study program without perspective, outdated and poorly organized (6.3 percent) were reasons for dropping out. According to one respondent's answer, there is even a case of harassment by a professor. Reasons contributing of 2 to 3 percent to the total dropout are health reasons, lack of motivation or modest previous education, pregnancy or starting a family, shiftees (the ones who changed faculty or study program), and not seeing oneself as a student.

Table 22 – Dominant reasons for bachelor drop out at UNIBL, 2007-2018. Sample size 96.

Reason	Number of respondents	In percentage
Dispute or conflict with lecturer	15	15.6%
Financial	13	13.5%
I worked, and due to work, I did not manage to fulfill my obligations at the faculty	13	13.5%
Continue education abroad	12	12.5%
Went abroad (not for study purposes)	11	11.5%
Corruption, unprofessionalism, lack of objectiveness	6	6.3%
Unperspective, outdated, poorly organized study program	6	6.3%
Wrong choice	5	5.2%
Health reasons	3	3.1%
I did not found myself as be a student.	2	2.1%
Lack of motivation	2	2.1%
Lack of previous education	2	2.1%
Pregnancy/starting the family	2	2.1%
Shifted to another faculty/study program	2	2.1%
Answer not related to dropout reason	1	1.0%
Harassment by professor	1	1.0%

Source: Author’s contribution.

The option of the second reason for leaving the university was completed by 62 respondents. The summary is presented in Table 25. The first place among the reasons for leaving the college is again disputes or conflicts with the lecturer (23.4 percent), followed by the imbalance between work and study (20.3 percent) and moving abroad

(14.1 percent). Financial reasons are equally distributed among respondents as the lack of perspective, outdated, poorly organized study program (9.4 percent).

Table 23 – Summary of the second reason for dropping out at UNIBL, 2007-2018. Sample size 64.

The second reason for leaving the university	Sum of Quantity	In perc.
Dispute or conflict with lecturer	*15*	*23.4%*
I worked, and due to work, I did not manage to fulfill my obligations at the faculty	13	20.3%
Moved abroad	9	14.1%
Financial	6	9.4%
Unperspective, outdated, poorly organized study program	*6*	*9.4%*
Pregnancy / starting family	4	6.3%
Health reasons	4	6.3%
Continue education abroad	2	3.1%
Corruption, unprofessionalism, unobjectivnes	*2*	*3.1%*
Family emergency	1	1.6%
Lack of motivation	1	1.6%
Shifted to another faculty or program	1	1.6%

Source: Survey distributed via email. Author's contribution.

Further analysis asked for the personal, institutional, pedagogical and financial reasons, which were systemized and offered, as major reasons identified in the literature.

Looking at the answers about personal reasons for dropping out of studies (Table 26), almost two-thirds of the respondents cited issues related to mental overwhelm, psychological unpreparedness, or lack of occupation and motivation. Only a small number of students (3.5 percent) interrupt their studies due to "pregnancy and marriage". This is important information because it allows the university management to find ways to help this group return to or continue their studies. Research shows that some universities have organized childcare facilities for students who are also parents.

Table 24 - Summary of the personal reasons for dropping out at UNIBL, 2007-2018.

Personal reasons	No of answers	In percent
It was an exhausting study, mentally, for me	27	31.8%
Not applicable to me	23	27.1%
At that moment, I was not ready for such kind of commitment	13	15.3%
Blank	11	12.9%
As I got familiar with the study program, I felt that this career path was not for me and that I would not do a job well	9	10.6%
Other (unsatisfied with curriculum, staff and professors, lack of interest, mobing (3), health, went to study abroad (3), ask for original documents)	9	10.6%
I had very good revenues and I was not motivated to study	8	9.4%
Pregnancy and marriage	3	3.5%
It was difficult to study because my family was not close to me	1	1.2%

Personal reasons	No of answers	In percent
Total no. Of students answer the question	85	

Source: Survey distributed via email. Author's contribution.

Almost one-third of the students who dropped out of their studies did so because of financial reasons, as indicated in Table 27. While tuition fees for domestic students are low (a scholarship holder pays 84 EUR per year), other associated costs such as housing, food, transportation, healthcare, etc. are very high, particularly for students who do not live on the campus and do not use the student restaurant.

Table 25 - Summary of the financial reasons for dropping out at UNIBL, 2007-2018.

Financial reasons for dropout	No of answers	In percent
Not applicable	63	71.6%
Parents could not afford to pay for my study, and I left university	11	12.5%
Blank	8	9.1%
I could not afford to study anymore	8	9.1%
I had to find a job to support my family	7	8.0%
Other (I needed a job, left)	2	2.3%
I stayed without my scholarship	0	0.0%
Total no. Of students who answered to the question	88	

Source: Survey distributed via email. Author's contribution.

Almost half of the survey respondents discontinued their studies due to institutional reasons, including insufficient internships or practical experience, programs not meeting the needs of the labor market, and boring teaching methods (Table 28).

Table 26 - Summary of the institutional and pedagogical reasons for dropping out at UNIBL, 2007-2018

Institutional and pedagogical reasons for dropout	No.of answers	In Percent
I lost motivation and interest to study during the school year	32	35.6%
Not applicable to me	31	34.4%
My expectations were unmet since there was not enough internship	26	28.9%
Professors' classes are boring	24	26.7%
Other (4 study or move abroad, 1 harassment due to physical look (long hair), 1 late at the classes, 2 poor learning environment (faculty physical resources), 1 dean refused to extend the deadline for paying semester's fee (100 Euros), 3 organization of the class, exam, poor curriculum, 1 unable to attend the classes, 1 corruption)	9	10.0%
Blank	6	6.7%
Total no. Of students who answered to the question	90	

Source: Survey distributed via email. Author's contribution.

We were unable to add the "reason for dropping out" variable to our model due to the following reasons:

- The interruption's reasons are not recorded in the database.
- When we examined the Dropout Student Book (hard copy) at one of the faculties, we found either an empty field or "personal reasons" listed in the section of "Reasons for dropping out".
- The online questionnaire distributed had a low number of respondents.

What happens after students leave: At the literature review stage we did not identified any research that follows the path of churned students after quitting their HE in B&H. According to survey data, almost one third of quitters leave HE permanently (Figure 35). That number increased at a bit more than 40 percent, by counting those who leave their second HE institution.

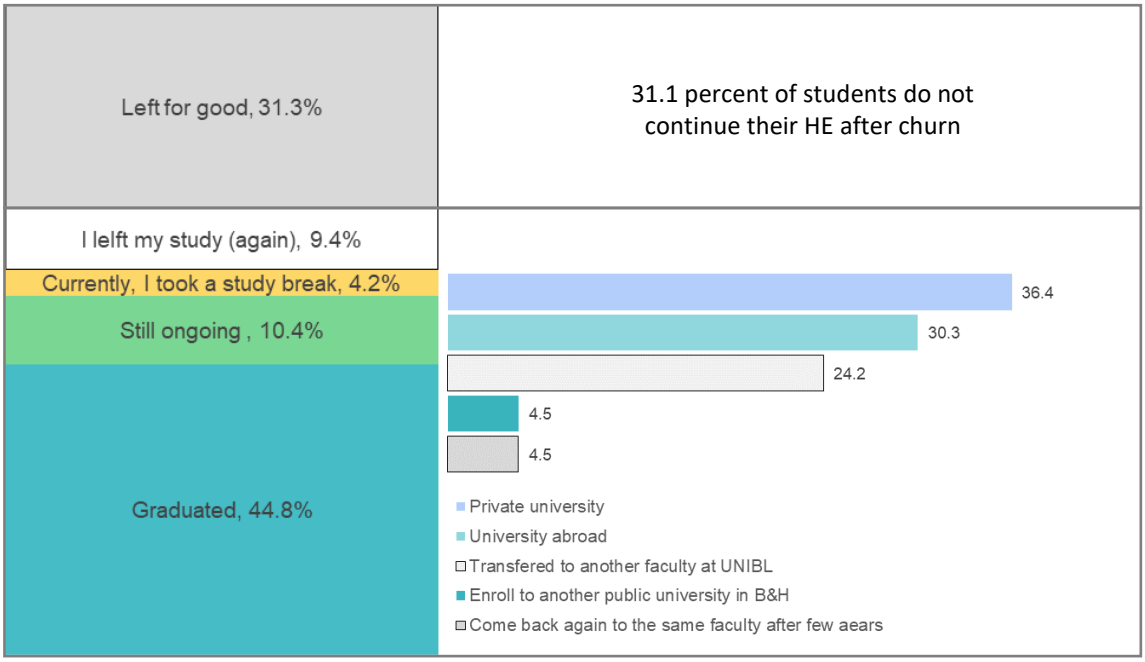


Figure 35 – What happened after dropping out at HEI? Share of students among 96 respondents of those who quit by own request.

Source: Survey distributed via email. Author's contribution.

The majority of students who quit at UNIBL went to some of the private universities in the country (36.4 percent), or abroad (30.3 percent). The rest of them, around one-third stayed at UNIBL, or another public university in the country.

The 28 respondents answered to question “Why you did not continue HE (somewhere else)?” with variety of reasons: health reasons (of the students or of the family member), lack of money, lack of motivation, fear of harassment, not having enough time due to full-time work, family or other reasons, high income at current job, career change, change of habitation place, outdated study program, enrolled but did not passed admission requirements, lack of flexibility from the faculty side due to specific medical condition of students.

The quitters are satisfied with their decision to leave HE for good (77 percent) and majority of them are employed (93 percent) (Figure 36). Still, 13 percent is not happy about their decision to leave UNIBL, and 19 percent think that they would have higher income now if they would not leave the HE.

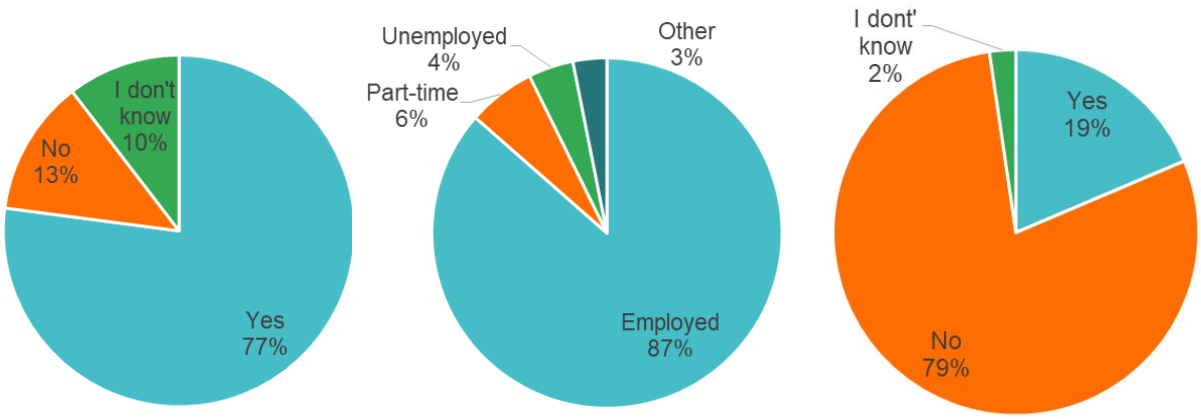


Figure 36 – Satisfaction and employment after HE leave. Answers to the questions: Are you satisfied with your decision to terminate the first enrolled study? (Left) Status of employment (Middle) Do you think that your income would be higher now if you had finished your studies? (Right)

Source: Author’s contribution.

Half of the employed leavers work in administration, business or IT. The rest of them is in service (accommodation, restaurants, transport, wellness, etc.), retail, wholesale, construction, health care, or manufacture.

6 RESULTS: EVALUATION OF THE EMPLOYED ML MODELS

This chapter presents the results of modeling churn at the University of Banja Luka, Bosnia and Herzegovina using data from students enrolled in bachelor's programs across all fields except Medicine. Data from students between 2007 and 2013 have been cleaned and preprocessed to be used in the modeling. The dataset mainly consists of binary data and has the following percentage of missing information:

- Total enrollment score of students, *score_t*: 87.3 percent of missing data
- Admission exam score, *score_e*: 88.3 percent of missing data
- Total number of passed courses at the end of the freshman year, *npe_1*: 46.4 percent of missing data.

All models are fed with the same number of predictors at three different times for churn prediction: before enrollment, at the beginning of the school year (enrollment week), and at the end of the first year (before enrolling in the second year). The only exception is the best performing model, which is HGBC. Several experiments were conducted on this model, including:

- Testing whether faculty variables contribute to the model's performance
- Running the model without highly ranked artificial variable (ID)
- Running the model with three variables that contain a lot of missing data (*score_t, score_e, npe_1*)
- Narrowing the churn definition and testing it on HGBC, on imbalance data.

PI and SHAP importance are documented for each model at four different times (pre-enrollment, enrollment, end of year, and using the top N variables). Detailed discussion of all metrics and feature importance are provided only for the best-ranked model, along with corresponding tables and figures. Tables and figures for the other models are included in the Appendix due to the document being data-intensive.

6.1 Feature importance evaluation over time

After training and test phase were over, for each model was inspected PI and calculated SHAP importance at global level. This was repeated four times for each model: with pre-enrollment data set, with data set at enrollment week, and at the end of freshmen year. The fourth time was inspected only with top N the most important features at the end of freshmen year in order to ameliorate the models performance. The top N is different for every model due to values of PI and thus that some models ranged only a few variables as important.

The performance improvement by feature reduction, using only top N features, was recorded in two cases: with HGBC and with SVM, polynomial kernel, 8-degree. This section describes results obtained using three data sets at three points in time (by PI and SHAP), summarized by models (12 models * 2 feature importance * 3 times of prediction).

In the analysis of the **pre-enrollment predictor** variables before the enrollment week, several variables showed significant trends (Figure 37). The variables with the highest number of appearances in the top 5 among the models were:

- Female gender (*gender_2.0*),
- Generation/cohort (in this dataset interpreted as year of expected enrollment - *ent_1st_y*),
- Male gender (*gender_1.0),*
- Variable indicating missing data of municipality (*mld_missing*),
- Distance up to 80 kilometers from UNIBL and *dist_0*,
- Attending a STEM, the secondary education (*hsd_STEM*),
- Student's age (*age1*),
- Variable representing the lack of labels for secondary education institution (*hs_missing*).

Over time, the number of appearances in the top 5 for all these variables decreased by two times or more, except for female gender. Female gender remained the dominant variable in the top 5 leading up to the enrollment week (24 times), but its frequency decreased only a bit, and fell to the second place according to the number of appearances by the end of first year data. The leading variables from the top 5 of pre-enrollment remain in the top 10 at enrollment week. However, new variables, such as secondary schools and

level of municipality development, *mld*, were expected to appear there. The Economics high school is also in the top 10 at the beginning and at the end of the year, alongside Gymnasium and STEM schools, unlike in top 5. The high school predictors (name of the secondary education institution, with prefix *hs,* and secondary education vocation or occupation, with prefix *hsd*) in the top 10 features are present in the prediction with pre-enroll variables more often than in the prediction using enrollment variables. In the pre-enrollment set, almost all secondary schools, except artistic ones, appear in the top 5 or top 10 variables, while in the enrollment time of prediction, only STEM secondary schools (or vocation) and Gymnasium appear in the first top 5/10. The importance of high schools/vocations in the time at the end of the year is reduced to only one model in the top 10 for Economics and Gymnasium (Figure 37).

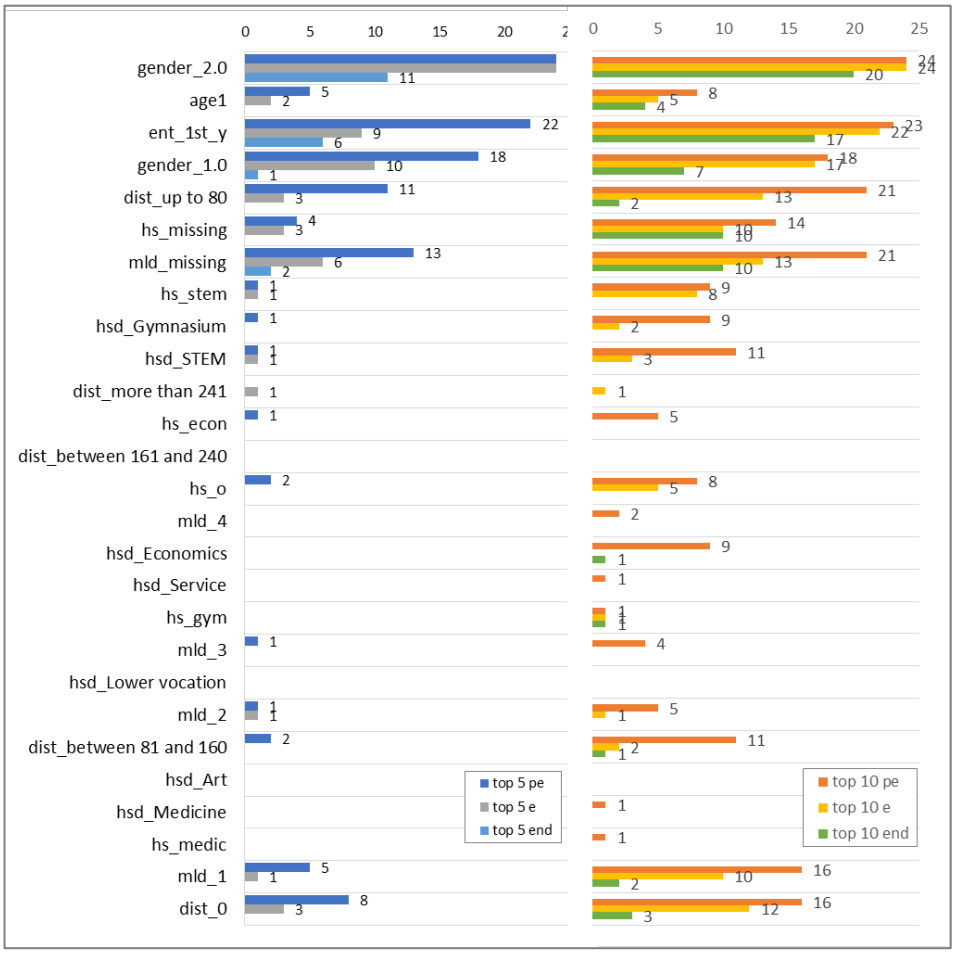


Figure 37 – Number of pre-enroll feature's occurrences in top 5 (left) and top 10 rank (right), by PI and SHAP for each model at three points in time (pre-enrollment, enrollment, and end of freshmen year).

(Abbreviation: end – end of freshmen year; e – enrollment; pe – pre-enrollment).
Source: Author.

For the **enrollment set of predictors**, the most important features for the most of the ML feature importance reports are: female gender (*gender_2.0*), calendar year of enrollment (*ent_1st_y*), type of enrollment: normal (*t_normal*), whether the student is a scholarship holder (*s_scholarship*), ID variable, whether the student co-finances their studies (*s_co-financing*), and whether the student is male (*gender_1.0*), (Figure 38 and 39). In the sixth to tenth positions, the remaining statuses *(s_part-time, s_self-financing, s_foreigner*) and enrollment types (*t_acknowledged_from_a_f* and *t_passive_year*) are included in a single model. Additionally, two variables from the beginning of the year, namely enrollment week variables, *score_t* and *score_e*, appear in the top 10, with each variable appearing only in one model (HGBC, because the other can't handle the missing data by default). The effect of missing data is present in almost half less models as in the pre-enrollment data, and this effect continues to decrease in even fewer models at the end of the first study year dataset.

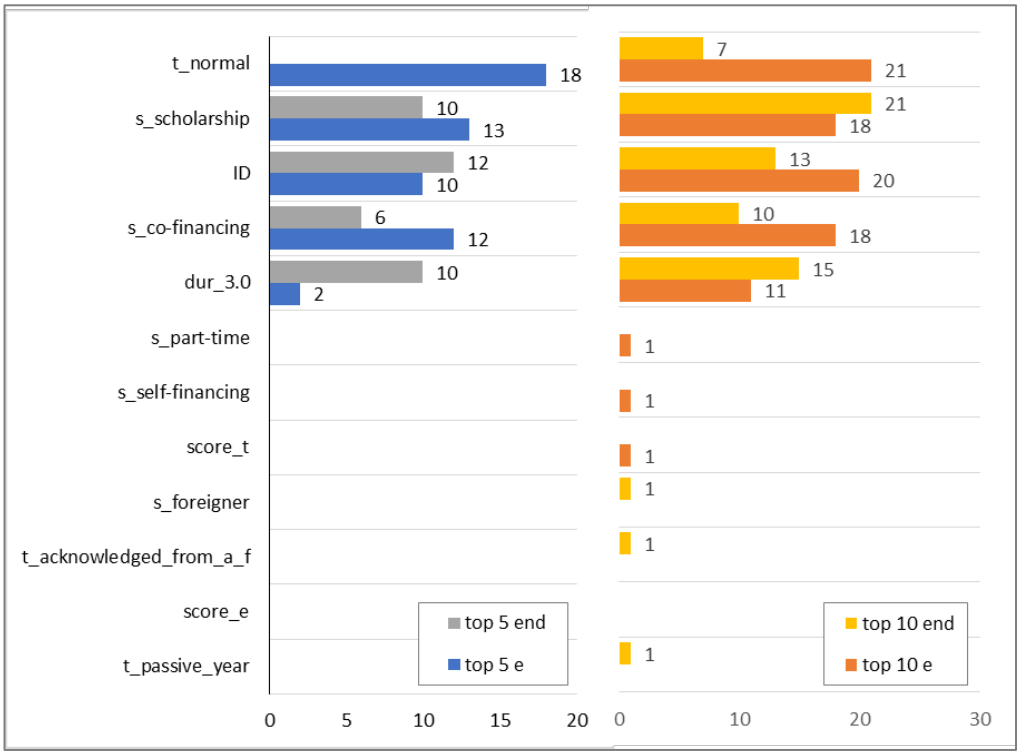


Figure 38 - Number of enroll feature's occurrences in top 5 and top 10 rank, by PI and SHAP for each model at two points in time (enroll, and end of freshmen year)

(Abbreviation: end – end of first year, e – enrollment)
Source: Author.

Out of the 8 variables added to the prediction model **at the end of the year**, 7 of them are among the top 10 (39). The variable with the total number of ECTS points collected at

the end of the year, *ects_1*, is among the top 10, but the ECTS ranks exist more often. To predict study interruptions at UNIBL after the first year, 18 and 15 different feature importance of models identified that ECTS ranks of less than 20 points and 41-60 points were among the top 5 most influential variables based on PI/SHAP. The variable *t_dropout* was among the top 5 for 8 of the models.

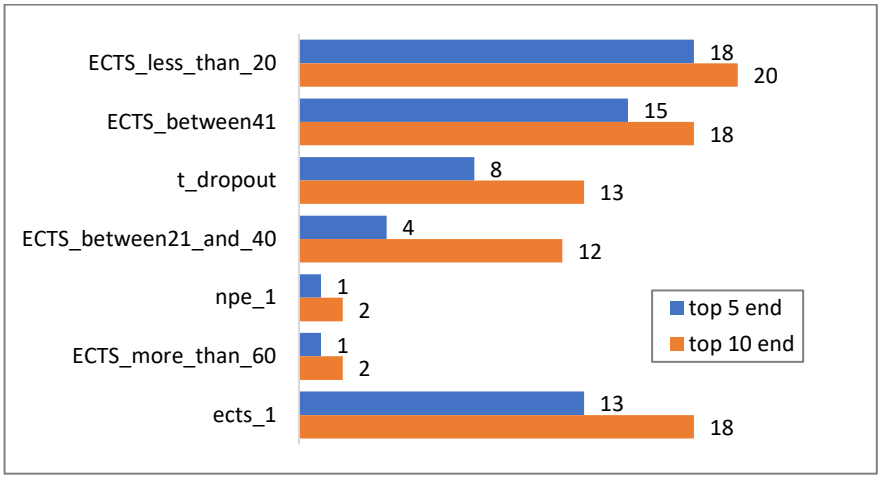


Figure 39 - Number of end of freshmen year feature's occurrences in top 5 and top 10 rank, by PI and SHAP for each model at the end of freshmen year

Source: Author.

It's important to note that even at the beginning of the year, only three variables from the pre-enrollment set remain dominant. However, over time, their frequency in the top 5 and top 10 decreases. These variables are female gender, *ent_1st_y*, and male gender. The is no single variable from pre-enrollment data set which showed increase in number of occurrences in top 5 or top 10 with passing of time, unlike is the case with some of variables added in enrollment week.

By the end of the first year, female gender (11 times) was the most frequently appearing variable in the top 5 from the pre-enrollment set. Out of the 12 enrollment variables that were tracked from the beginning (enrollment week) to the end of the year, only 4 of them consistently appear in the top 5 (*s_scholarship, ID, s_co-financing, dur_3.0*, Figure 39). It's interesting to note that the occurrences of two variables (*ID* and *dur_3.0*) increase over time, while the frequency of occurrence of the other enroll variables in the top 5 decreases or disappears completely (for example, type of enroll, *t_normal*).

In summary, at the end of the first year of study, the predictor variables that appeared most frequently in the top 5 by PI and SHAP for all models are ECTS ranks (less than 20, and between 41 and 60) with 18 and 15 repetitions respectively, followed by *ects_1*.

ECTS ranks were brought in to mitigate the large number of zero values in *ects_1* variable due to data base default features, were was impossible to determine are the zeros represent lack of labels or zero scores by students.

6.2 HGBC performance evaluation

The best performed model, taking into account evaluation metrics in three time intervals plus time needed for training and evaluation was ensemble tree, i.e. HGBC. While the model showed only a modest test accuracy in the pre-enrollment data set, still it had the highest score among all other models on the pre-enrollment data. The not impressive, but still the highest overall accuracy on pre-enrollment data set with all 27 predictors (0.63171) was improved by adding more data (enroll variables) and jumped to the 0.75 (Table 29).

The model performance at the end of the freshmen study year were further improved by feature reduction to 13 of 48 the most important variables. The model is able to correctly classify 82 of 100 true dropout students at the end of the academic year, while that number on the beginning of academic year (enroll week data) was quite high: 72, and at the same time far better than the most of the models. The highest ROC AUC (between 0.63 and 0.83) and specificity (between 0.66 and 0.83) among all other models, observed on all three set of predictors, together with high speed of calculating the PI and SHAP values, puts the HGBC on the first place as the best performed model.

Table 27 – Summary of HGBC model

HGBC	Pre enroll	Enroll	Enroll*	End of 1st year (all variables)	End of 1st year (top N)**
Accuracy	0.63171	0.75223	0.75511	0.82415	0.82535
True Negative	1375	1636	1634	1735	1728
False Positive	716	455	457	356	363
False Negative	815	575	561	375	363
True Positive	1251	1491	1505	1691	1703
True Negative	33.08%	39.36%	39.31%	41.74%	41.57%
False Positive	17.22%	10.95%	10.99%	8.56%	8.73%
False Negative	19.61%	13.83%	13.50%	9.02%	8.73%
True Positive	30.09%	35.87%	36.20%	40.68%	40.97%
F1	0.62038	0.74327	0.74727	0.82227	0.82430
Precision	0.63599	0.76619	0.76707	0.82609	0.82430
Recall	0.60552	0.72168	0.72846	0.81849	0.82430
ROCAUC	0.63155	0.75204	0.75495	0.82412	0.82535
Matthews corr.	0.26347	0.50510	0.51070	0.64830	0.65070

HGBC	Pre enroll	Enroll	Enroll*	End of 1st year (all variables)	End of 1st year (top N)**
PR score				0.92061	0.92010
Specificity	0.65758	0.78240	0.78144	0.82975	0.82640
Train accuracy	0.67573	0.80589	0.80752	0.86280	0.85955

*Enroll data set with faculty variables included was experiment to show that those variables do not contribute to the model performance in a significant amount. ** Thirteen the most important variables according to PI and SHAP.
Source: Author.

Regarding the most important metrics in dropout modeling, the recall, as indicator of number of correctly classified dropouts among 100 those who really left HE, the HGBC had good-enough performance respecting the rest of models. This is important metrics because the cost of losing one student, i.e. classification of in-risk as non-risk student (False Negative), is higher than predicting non-risk student as dropout (False Positive).

There are models with higher recall than HGBC in pre-enrollment prediction of churn, like 4-layers NN, with averaged recall of 0.6411[15], while HGBC has 0.6055 (Figure 40). Still they are scored lower by overall accuracy, ROC AUC, and specificity, and need far more time for PI and SHAP values calculation.

The highest end of year recall is by SVM, linear kernel with 0.8693, while the HGBC had 0.8243 which is still high. Also, the SVM, linear kernel (with 4-layer NN) had lower recall with adding more variables from pre-enrollment to enrollment phase of prediction.

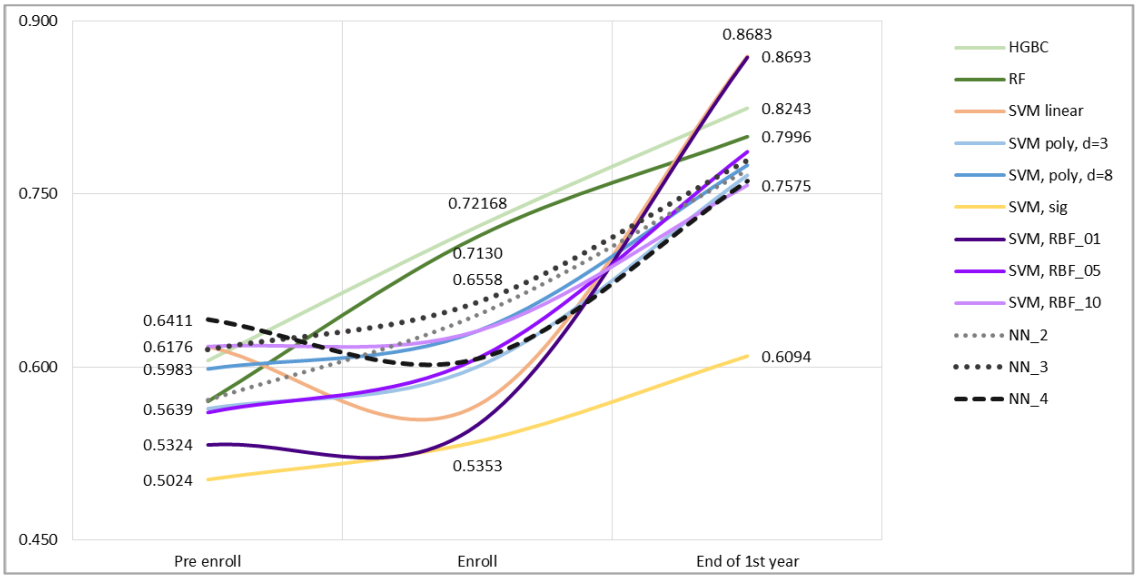


Figure 40 – Summary of recall in three times of prediction by each model

[15] The highest pre-enrollment reported recall is 0.7032, on 4-layer NN, with overall test accuracy of 0.6109

Source: Author.

Further experiment done only with HGBC showed that adding variable of faculty in the enrollment set of predictors does not improve model metric in significant amount (for instance, the accuracy test score was increased from 0.75223 to 0.75511, Table 29). The next experiment was to remove ID variable from end of year set of predictors, while keeping everything else unchanged. The accuracy score goes down to 0.75622 from 0.82415. The third and the last experiment with HGBC was to feed the model by excluding the variables with missing data (*score_t, score_e*, and *npe_1*) in order to give the model the same chance as to those who were not fed with those three variables. The model performance were inconsiderable changed. For instance the overall accuracy before cutting the features was 0.82415 and after 0.82223, and recall was 0.81849 and after 0.81897 (Table A23 in Appendix).

The confusion matrix through all three time points shows improvement in the model performance, which is quite rare at the rest of models. The inclusion of additional variables in the model over time resulted in all elements of the confusion matrix moving in the right direction. The number of True Positive and True Negative values increased, while the number of False Positive and False Negative values decreased.

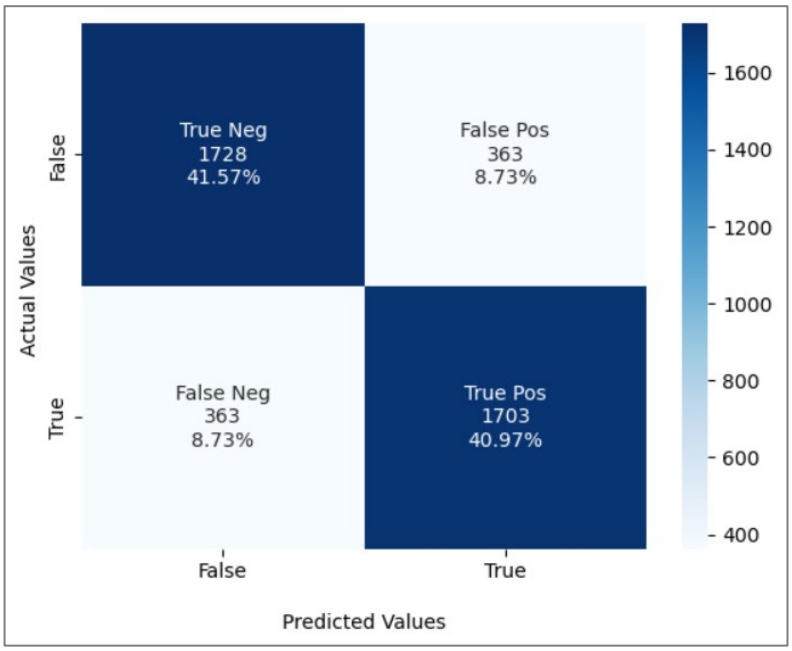


Figure 41 – Seaborn confusion matrix with labels for HGBC at end of first year (top N) data set.

Source: Author based on Python seaborn library results.

Figure 41 shows test set performance at the end of first (freshman) year data with 13 the most important features. This demonstrated that by incrementing the number of variables over time in the ensemble tree classifier model, the model continuously improves.

The only two models that showed improvement in test accuracy after feature engineering, specifically the removal of unimportant variables at the end of the first year, are HGBC and SVM with a polynomial kernel and 8-degree. However, the only model unaffected by large dataset and needed resources for its evaluation is HGBC. The model demonstrates that the highest feature impact on the model, as indicated by PI and SHAP, comes from almost the same variables which are ordered in a similar manner.

Global feature importance before enrollment:

On the pre-enroll data set, using the PI, the model is influenced the most by variable female gender, which is in line with SHAP global importance, too (Figure 42). It means if student is a female, that feature impacts the model accuracy score (decreasing its value by shuffling that variable, while keeping all others untouched), far more than shuffling the other variables in PI.

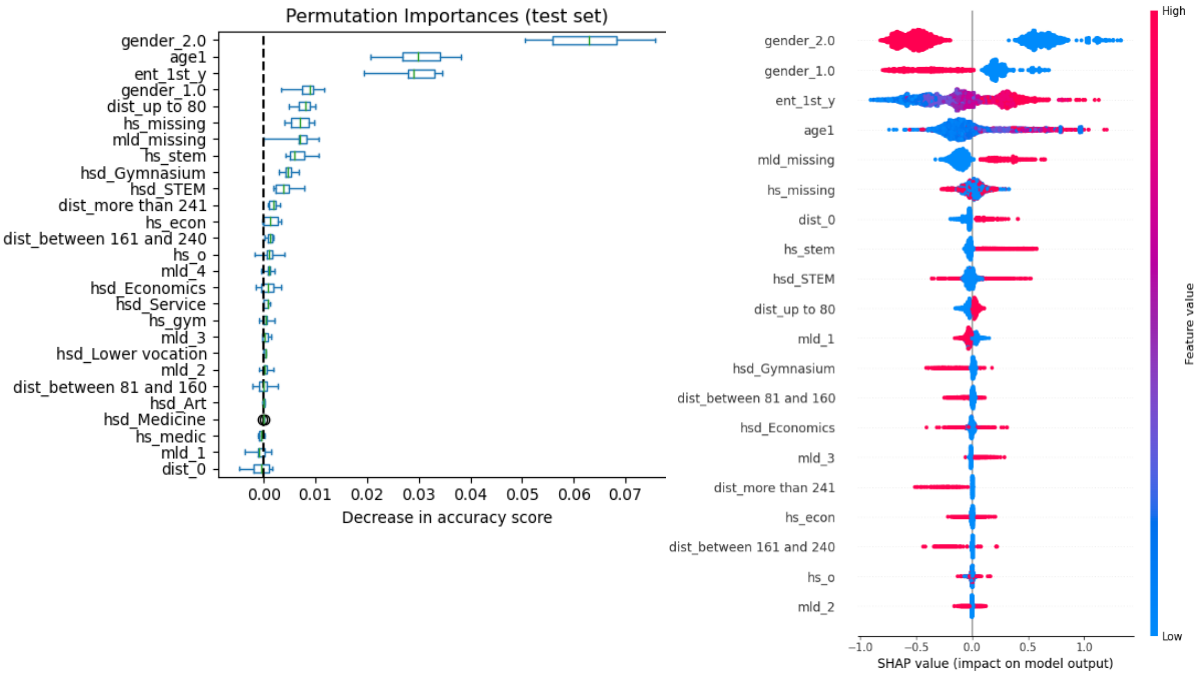


Figure 42 – Pre-enrollment data set feature importance. Left: PI. Right: SHAP global feature importance.

Source: Author, based on models output in Python.

The following are: age of student, year of expected (potential) enrollment, gender male, and distance up to 80 km from the University place (Banja Luka). The two variables that

carry the missing data are included in this step, to check up their impact size to the model performance and importance. They are excluded at the end of school year data in the data set of top N variables, because they don't have explanation importance for the UNIBL, as mayor stakeholder of this research.

On the Figure 42 – left: PI, the variables that do not contribute to the model performance in any way are at the bottom (high school degree in Medicine - *hsd_Medicine*, high school of medical domain – *hs_medic*, municipality level of development 1 – *mld_1*, and variable that carry the missing data in the place of student's origin – *dist_0*). Permutation importance is checked also for train set, to check the possibility of overfitting, but is not recorded in the Appendix, due to its size.

The right side of Figure 42 shows SHAP variable importance at the global (model's) level and their prediction "orientation" to the class 1 (dropout). The female gender stands out again as the most dominant feature. The blue dots (each dot represents one student, i.e. record in a data set) are values 1 (or True) of the variable, and when variable takes value 1 – it means that student is a female – it has higher chance to be classified as a non-dropout. On the second place is gender male.[16] Its dispersion is a bit smaller than variable above, and thus its less impact the model. Feature *end_1st_year* contains data of years of (expected) enrollment (2007-2013), and shows that the oldest the generation is, the more persistent was, i.e. if the students enroll in the recent years, it has higher chances to be classified as dropout. The same explanation follow the age of student – *age1*. The municipality level of development, *mld_missing,* the name of high school, *hs_missing,* and *dist_0*, are three variables which carry the missing data and their importance was quite high almost by each model, due to its large number of lack of labels. If student do not have recorded the place of origin, it has a higher chance to be classified as dropout. If student comes from some of STEM high school (*hs_stem*), it has a higher chance to be classified as dropout (due to fact that majority of those students enroll into STEM faculties which have higher dropout rates), unlike those students who come from Economics high schools, *hsd_Economics, hs_econ*, were it does not necessary means classification in both direction (as dropout, or non-dropout). Another variable which has some recorded impact to the model performance and act in a same way as two previously mention variables is municipality level of development, *mld_2*, which spreads in both direction – to the classification as dropout, and non-dropout. Similar explanation as was

[16] Both genders are kept due to missing gender labels in a data set. Explanation provided in Chapter 5.

for *hs_stem* goes for high school vocation (or degree), *hsd_STEM*, although dispersion of this variable goes and to the left side of SHAP value (in favor of non-dropouts). If student comes from municipalities that are up to 80 km away of UNIBL, it is less persistent, i.e. has higher chance to be dropout, and opposite: the far is place of origin of the student, it has a higher chance to be classified as non-dropout (variables *dist_between 141 and 240*, and *dist_more than 241* km). If student comes from developed municipality, *mld_1*, it has a higher chance to be classified as non-dropout, unlike the students who are from undeveloped towns, *mld_3*. Some high school's degree go in favor of students, like the vocation of Gymnasium, *hsd_Gymnasium,* increases their chance to finish their tertiary education. This is also due to fact that those students enroll into variety of faculties.

Comparing the feature importance by PI and SHAP, it is evident that the rank of features is not the same, still among top 10 in SHAP, there is 9 features from top 10 in PI.

Global feature importance at the beginning of the school year:

When comparing the importance of PI and SHAP at the beginning of the school year, in the top 5 places, they are essentially the same variables, just in a different order. Furthermore, in top 10 for SHAP importance, only two variables from the top 10 of PI are not present (Figure 43).

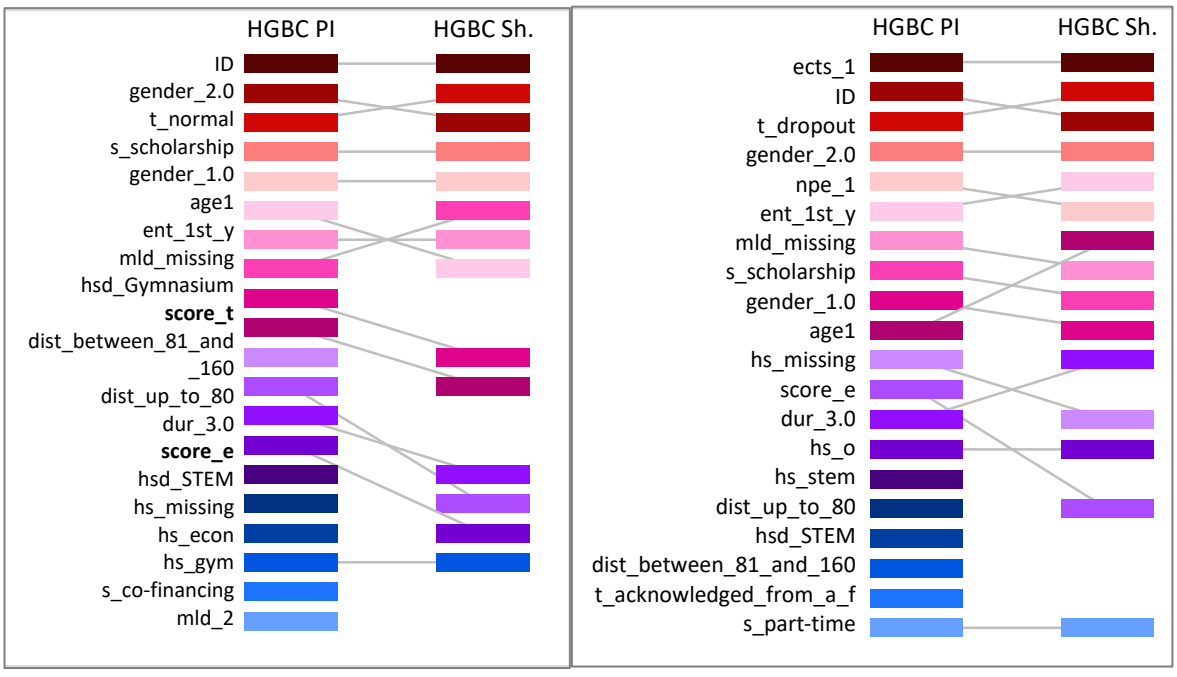


Figure 43 – HGBC: Importance by PI and by SHAP at the beginning of the school year (enrollment week), (left), and at the end of school year (right)

Source: Author.

The figure displays the top 20 features sorted by PI and SHAP for the HGBC at enrollment and end of first year dataset. Empty spaces on the SHAP side indicate that variables ranked 20 or higher according to PI fill those spaces. In total, there is a lack of 6 of the top 20 variables in the SHAP top 20 importance, which are also in the top 20 of PI. At the beginning of the school year (in enroll week) the top 5 variables are *ID*, female gender, type of enroll *t_normal*, *s_scholarship*, and male gender.

The ID variable comes to first place only by two models: HGBC and RF. And is in top 5 by SVM, polynomial, with 8-degree and 4-layer NN. This variable is an artificial variable added before data transformation when students in a raw data set were listed by year of enrollment, by faculty, and by institution ID, which corresponds to the rank at admission (enrollment) score.[17] In that way for ML model is possible to identify the students who are highly ranked even have lack of data for enrollment score (*score_e, score_t*). This is the reason why the ID variable is kept. Female gender variable shows high impact to the model prediction of non-dropout. The following highly important variables at the beginning of school year are type of enrollment, *t_normal*, scholarship holder, *s_scholarship*, and male gender, *gender_1.0* (Figure 44).

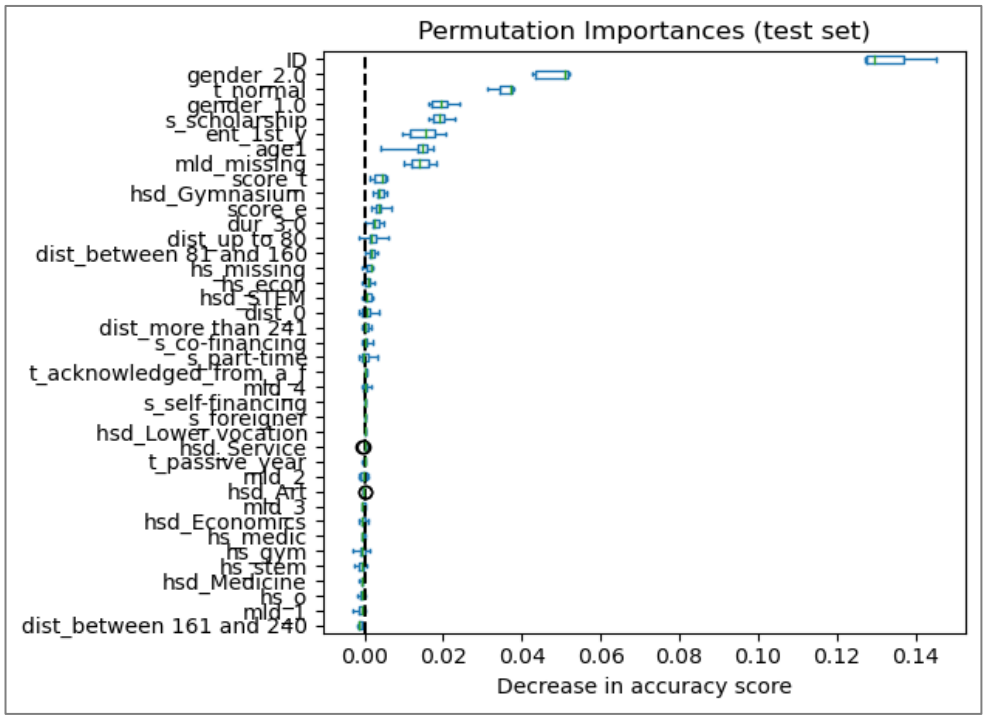


Figure 44 – HGBC, enrollment variables, PI, test set.

Source: Author.

[17] Having in mind that enrollment quotas were unchanged for few years, it is possible to identify the year and faculty of enrollment.

The *score_e*, and *score_t* are input variables only in case of HGBC due to large portion of missing data. The variables which don't have any contribution to the model, according to PI (Figure 44) are the last 15 variables, starting at enrollment status: foreigner, *s_foreigner* on the chart above. According to the SHAP importance, those are *s_foreigner, s_self-financing*, and *hsd_Service*.

Global feature importance at the end of freshmen year:

The predictor variables from the end of the first year were initially used as input for the model, together with all previous. Subsequently, 13 top variables were selected based on PI, and the model was fed again with these variables. The results of model evaluation metrics for the top 13 variables are presented in Table 29, while importance for PI and SHAP, sorted by PI for top 20 variables is on the Figure 43, (right). The top 10 variables are the same for PI and for SHAP but ranked differently. There are 6 variables: enrollment faculty (admission) exam score, *score_e,* secondary education in STEM domain*, hs_stem, hsd_STEM, dist_between_81_and_160* km*,* type of enroll: acknowledged from another faculty, *t_acknowledged_from_a_f* and status at the time of enrollment that student is a part-time student, *s_part-time,* which are listed in top 20 by PI are not in the top 20 by SHAP. Among top 5 feature importance is a gender (from pre-enrollment), *ID* (enrollment), and collected amount of ECTS at the end of the freshmen year, *ects_1*, dropout which occurred during the year, *t_dropout*, and the total number of passed courses at the end of the first study year, *npe_1*.

The bees-warm plot of SHAP values for enrollment data and end of first year prediction (with all variables) were not presented because those are basically the same plots, and the only difference is in the size of features impact to the model output (the x-axis size in pre-enroll SHAP plot is 0 – (±1), enroll 0 – (±4), and end of first year 0 – (±6)).

At the end of year the strongest contribution to the model is generated by the number of collected ECTS credits, *ects_1, ID, t_dropout*, female gender, number of passed courses, *npe_1*, and school year of enrollment, *ent_1st_y* (Figure 45, left). The student is in risk of churn if has a few ECTS credits collected at the end of first year, if his/her ID is higher, if is not a female gender, and his/her year of enrollment is higher in a rank (2007-2013), (Figure 45, right).

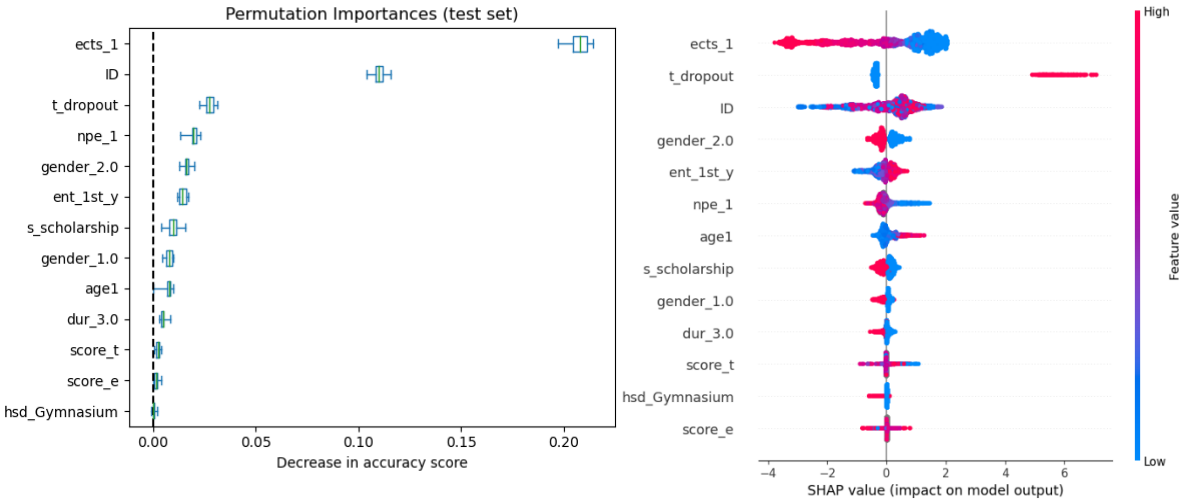


Figure 45 – End of freshman year data set with 13 the most important variables. Left: PI. Right: SHAP global importance.

Source: Author.

Global dependence:

To illustrate the advantages of SHAP through a detailed analysis of individual variables, here we will highlight the most significant variables and their strongest interactions to gain the new knowledge of our data. The most influential variable impacting the HGBC model is the total number of ECTS credits earned at the end of the first year (Figure 46, upper left). The feature with the strongest interaction with *ects_1* is *dur_3.0*, which represents the study duration. Students who have collected fewer than approximately 40 credits are associated with higher SHAP values, indicating a greater likelihood of being classified as dropouts. Conversely, students enrolled in a three-year bachelor study duration are also at risk of attrition. These students are spread both above and below the model's baseline (0.075).

The variable that has the highest interaction with *hsd_Gymnasium* is whether the student has a scholarship or not. Students who have completed Gymnasium and do not have a scholarship are classified as the less risky group for dropping out of studies, and opposite: students who graduated in Gymnasium and are scholarship holders are at risk of churn (Figure 46, upper right).

Students who hold a scholarship and have passed approximately up to 5 exams are considered to be at low risk of interrupting their studies (Figure 46, middle left). Students who have passed more than 5 exams are considered at risk of interrupting their studies, no matter of scholarship statu. This may be because these students are transferring from one study program, faculty, or university to another.

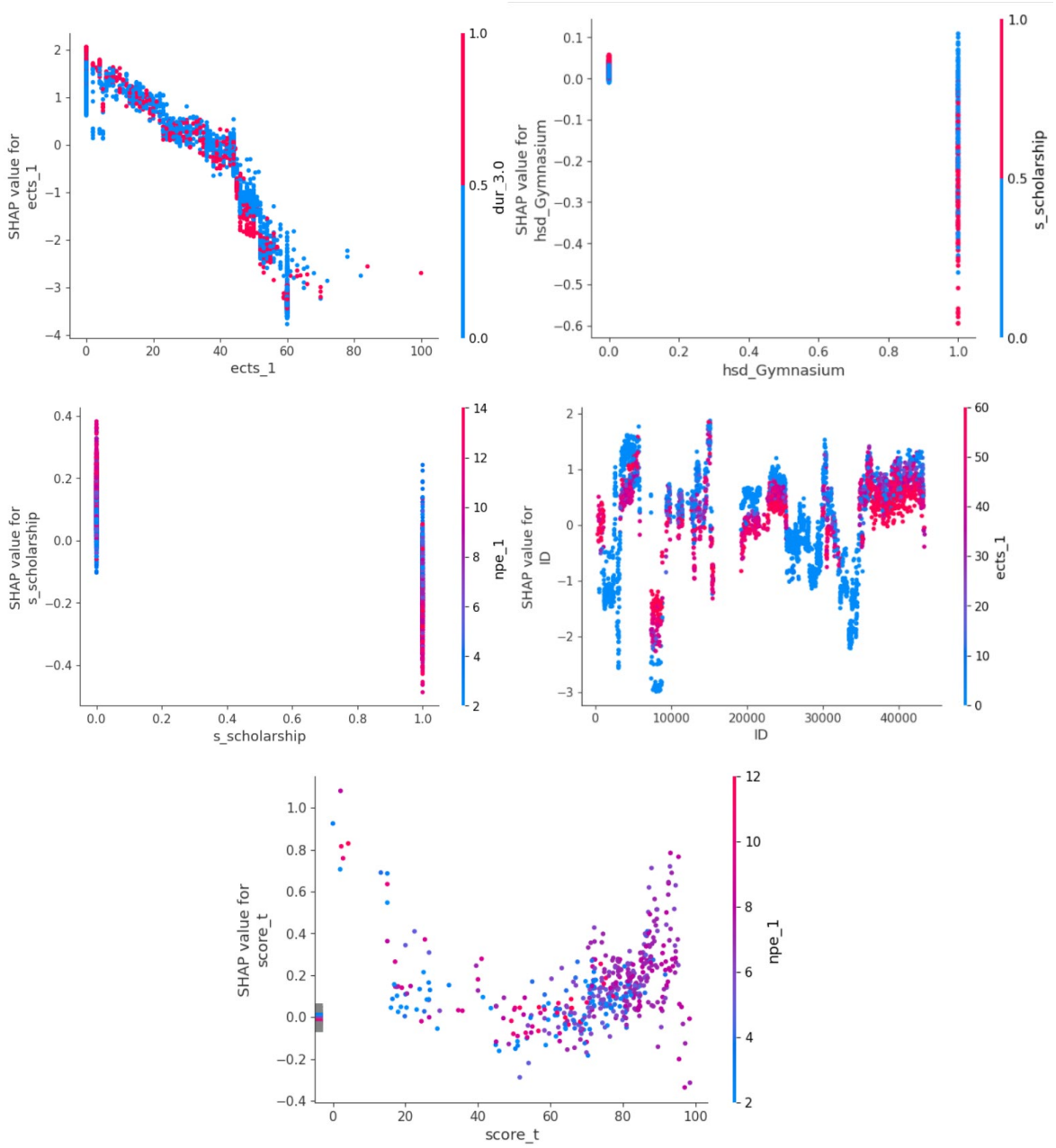


Figure 46 – Dependency plots of some of the top 13 variables at the end of first year and their strongest interactions: ects_1 (upper left), hsd_Gymnasium (upper right), s_scholarship (middle left), ID (middle right), score_e (bottom)

Source: Author

The ID variable was ranked second by PI in the data set at the end of the freshman year, using the top 13 features. In the raw data obtained from UNIBL, students are listed by faculties and study programs. In each such cohort, it was observed that students are usually listed according to the generation (from newest to oldest, i.e., from 2018 back to 2007) and their enrollment rank (which is part of institution's ID and is not unique across

the University). ID is artificial variable added in the first step of data exploration. This why the ID variable has the values from 1 to 45 thousand. According to SHAP analysis the variable which interact the most with ID variable is the *ects_1* (Figure 46, middle right). For SHAP values higher than base value (0.075) students have higher chances to be dropouts, and vice versa. But students with approximately more than 30 collected credits at the end of the year are also those who are in risk of churn. Again model shows that swichers (between faculties and universities) are more in risk of dropping out.

The *score_t* represents the total score at enrollment, which is the combination of secondary education GPA and admission exam score (Figure 46, bottom). It has the strongest correlation with *npe_1*. Students who achieve a total score of 80 or higher are more likely to drop out. Some of these students may choose to drop out despite having a high enrollment rank, possibly because they want to transfer to universities abroad, which may have more lenient acceptance policies for students enrolled at UNIBL.

Individual level of feature importance:

SHAP individual force plots are a visualization method used to illustrate the breakdown of individual predictions. Each part of the plot displays how the positive (red) or negative (blue) contribution of each feature moves the value from the expected model output over the background dataset to the model output for a specific prediction. The models score at each particular case is a sum of base value (0.07496, represent the probability of classifying student as dropout) and positive and negative contributions (Figure 47).

Based on the end-of-year data, when considering the top 13 most important variables, the student was classified as a dropout, even though they had collected 57 ECTS by the end of the year (see Figure 47, a). The model's prediction is 2.45, which is much higher than the base value 0.07496. In this specific case, the red-colored features in the plot move the prediction to the right side, classifying the student as a dropout.

An example of a non-dropout classification occurs when a student has collected 60 credits, is not a scholarship holder, did not attend a Gymnasium, and is male. This is presented in Figure 47, b), where the model output was -4.05, indicating a value below the base value.

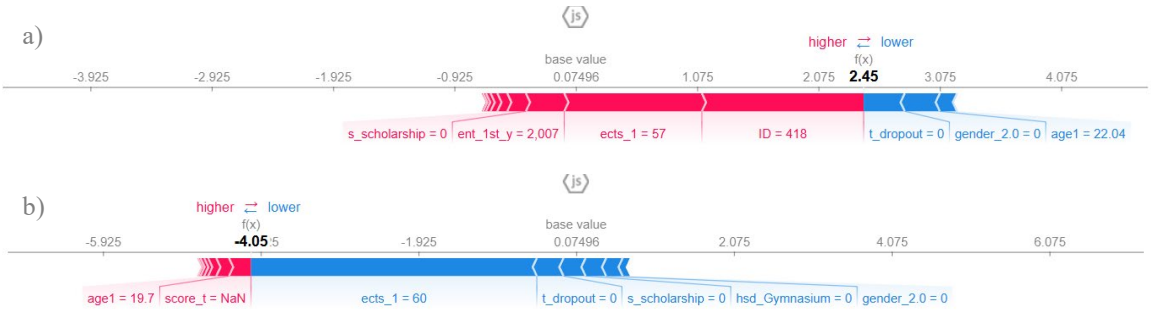


Figure 47 – SHAP local: individual cases of dropout (a) and non-dropout (b) prediction.

Source: Author.

Some researchers found out that a waterfall plot is more understandable to end users. This is why both are presented here for two extreme cases of model outcome values from above (Figure 47). Waterfall plots show the same thing in a different way. The starting point of a waterfall plot is the expected value of the model output, the same as at plots above (rounded at two digits). Plots showing the all 13 features and their impact to the classification task.

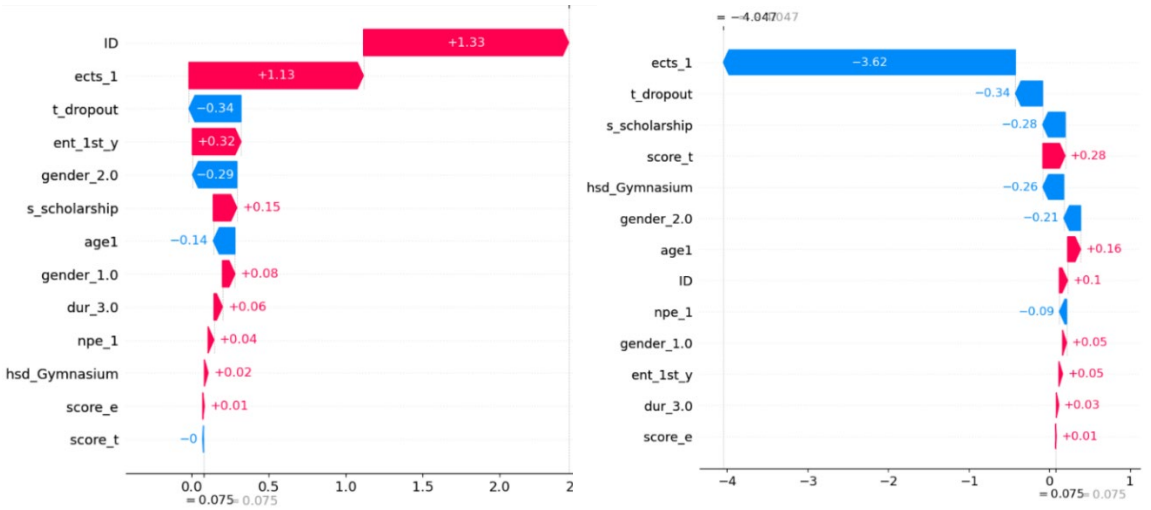


Figure 48 - SHAP local: individual cases of dropout (left) and non-dropout (right) prediction

Source: Author.

Presenting both methods to gain insights into a particular case demonstrates the power of SHAP explanations at an individual level. This can be beneficial for faculty management and student counselors in helping students to stay engaged in their studies.

6.2.1 Imbalance scenario: change in definition of dropout

In the raw dataset obtained from UNIBL, there was a variable indicating the type of enrollment, which included a category for dropout. This category mainly involved cases where students withdrew from the program at their own request. In interviews with administrative officers, it was discovered that systematic changes were made to the data only at the Faculty of Natural Science and Mathematics. This involved removing students from the dataset who had not updated their statuses for several years, and the administration staff withdrew them (as dropouts) from the records. At the start of the modeling process, all the models were fed with a target variable that was only True if the variable type of "enrollment: dropout" was also True. This dataset revealed an imbalanced ratio, with the minor class (dropout = True) representing only 0.28 percent of the dominant class (dropout = False). In interviews with administration chiefs at various educational institutions, it was revealed that the current definition of dropout is not comprehensive enough. This is because the majority of institutions do not change the student's status unless the student initiates the change themselves. As a result, a step was taken to broaden the definition of dropout to better reflect reality according to the one established in the chapter 5.

Still, from the ML point of view, it is important to monitor changes in the definition and assess their impact on the model's output. In this chapter, we examine the behavior of the best performing model, HGBC, in both imbalanced and balanced scenarios for all three time predictions (Table 30). The overall accuracy on the test set is not an appropriate measure when the data is imbalanced. This is because the model tends to learn to identify the dominant class better than minor class. That's why metrics like recall, precision, and PR score are more relevant in this case. They allow us to evaluate the model's performance in a way that accommodates imbalanced data.

Table 28 – Summary of HGBC imbalanced and balanced metrics in three prediction times

HGBC	Imb.	Bal.	Imb.	Bal.	Imb.	Bal.	Imb.	Bal.
	Pre enroll		Enroll data		End of 1st year		top N*	
Accuracy	0.79403	0.66225	0.86389	0.82679	0.88340	0.85377	0.88051	0.84630
True Negative	3173	2234	3150	2839	3153	2882	3162	2839
False Positive	86	1025	109	420	106	377	97	420
False Negative	769	377	456	299	378	230	399	218
True Positive	123	515	436	593	514	662	493	674
True Negative	76.33%	53.74%	75.78%	68.29%	75.85%	69.33%	76.06%	68.29%

HGBC	Imb.	Bal.	Imb.	Bal.	Imb.	Bal.	Imb.	Bal.
	Pre enroll		Enroll data		End of 1st year		top N*	
False Positive	2.07%	24.66%	2.62%	10.10%	2.55%	9.07%	2.33%	10.10%
False Negative	18.50%	9.07%	10.97%	7.19%	9.09%	5.53%	9.60%	5.24%
True Positive	2.96%	12.39%	10.49%	14.27%	12.36%	15.92%	11.86%	16.21%
F1	0.22343	0.42352	0.60682	0.62257	0.67989	0.68566	0.66532	0.67875
Precision	0.58852	0.33442	0.80000	0.58539	0.82903	0.63715	0.83559	0.61609
Recall	0.13789	0.57735	0.48879	0.66480	0.57623	0.74215	0.55269	0.75561
ROCAUC	0.55575	0.63142	0.72767	0.76796	0.77185	0.81324	0.76146	0.81337
Matthews corr.	0.20945	0.22349	0.55380	0.51250	0.62654	0.59402	0.61511	0.58432
PR score	0.40082	0.39910	0.74205	0.73869	0.80212	0.79799	0.79737	0.79562
Specificity	0.97361	0.68549	0.96655	0.87113	0.96747	0.88432	0.97024	0.87113
Train accuracy	0.80100	0.72932	0.88833	0.84779	0.89948	0.87358	0.89303	0.85991

* Top N are sorted PI of the top 10 of a balanced data set at the end of the year, excluding variables representing missing data: *ID, t_dropout, ects_1, npe_1, dur_3.0, ent_1st_y, age1, hsd_STEM, gender_1.0, gender_2.0*.
Source: Author.

The modeling results for pre-enrollment, enrollment, and end-of-year datasets show poor performance in both imbalance and balance scenarios, when compared to the broad definition outlined in Chapter 6.2, with equally representation of both classes, Table 29. In the imbalance scenario, the specificity is high because of the large number of correctly identified True Negative values (non-dropouts), as expected. The RF, SVMs, and NNs showed less performances on imbalanced data using up-sampling comparing to HGBC. For those models, imbalance and balance scenarios are not displayed, as that would greatly exceed the scope of this study.

The PI and SHAP importance for imbalance scenario were done for all three time of prediction and presented in Appendix, Table A27, A28, and A29. The visual interpretation of results at the end of year is on Figure 49. Top 8 variables by SHAP have the same rank for balanced and imbalanced data. The rest of variables are similar. The importance of balanced data prioritizes the *hsd_Economics*, while the imbalanced data prioritizes *hsd_Gymnasium* and STEM.

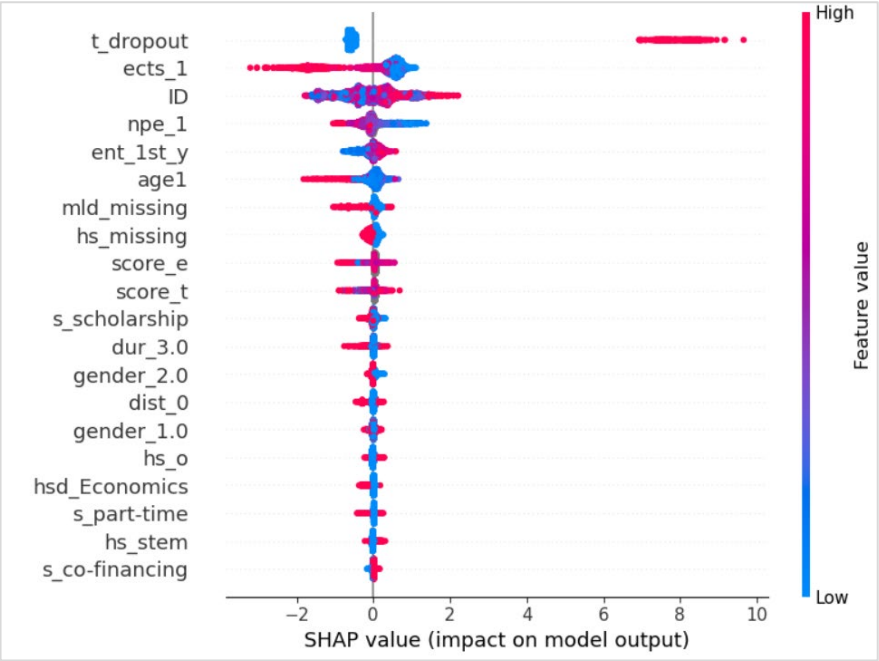


Figure 49 – HGBC, balanced, SHAP importance at the end of first year using all variables.

Source: Author.

Top 10 in SHAP importance (balanced) are similar with SHAP importance at broad definition data set.

6.3 RF performance evaluation

The second best performed model in the time of enrollment week prediction is RF. Adding variables over time improved model metrics. However, removing unimportant variables at the end of the year slightly reversed the improvements of all metrics. Nevertheless, it is the only model which has showed the signs of overfitting, due to significant difference between test set (0.57902) and train set (0.97510) accuracy at pre enroll data set and perfect accuracy on train data at the beginning (1.0) and at the end of school year (1.0), (Table 31). Still, the accuracy of the model at the beginning and at the end of first year reaches the second best place among all models.

Table 29 – Summary of RF

RF	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)*
Accuracy	0.57902	0.72841	0.81236	0.80370
True Negative	1229	1555	1712	1689
False Positive	862	536	379	402
False Negative	888	593	401	414

RF	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)*
True Positive	1178	1473	1665	1652
True Negative	29.56%	37.41%	41.18%	40.63%
False Positive	20.74%	12.89%	9.12%	9.67%
False Negative	21.36%	14.27%	9.65%	9.96%
True Positive	28.34%	35.43%	40.05%	39.74%
F1	0.57379	0.72294	0.81022	0.80194
Precision	0.57745	0.73320	0.81458	0.80428
Recall	0.57018	0.71297	0.80591	0.79961
ROCAUC	0.57897	0.72832	0.81233	0.80368
Matthews corr.	0.15797	0.45688	0.62473	0.60739
Specificity	0.58776	0.74366	0.81875	0.80775
Train accuracy	0.97510	1.00000	1.00000	1.00000

*Top 13 variables according the PI on test set showed decrease in accuracy.
Source: Author.

RF PI on top 4 using pre-enroll features gives the same features as at HGBC PI, and in the same order: female, *gender_2.0*, age before enrollment week, *age1*, year of potential enroll, *ent_1st_y*, and male gender. The fifth feature in a row is *hsd_Gymnasium*. This is the only case when Gymnasium high school vocation reached the top 5. The detail order of variables is presented in Appendix, Table A24. And there is evident that only the top 13 variables contribute to the model based on PI.

With time passing the differences between the HGBC and RF feature importance are growing. Top 5 features on enrollment week are: *ID, t_normal, ent_1st_y, gender_2.0*, and *age1* (Appendix, Table A25).

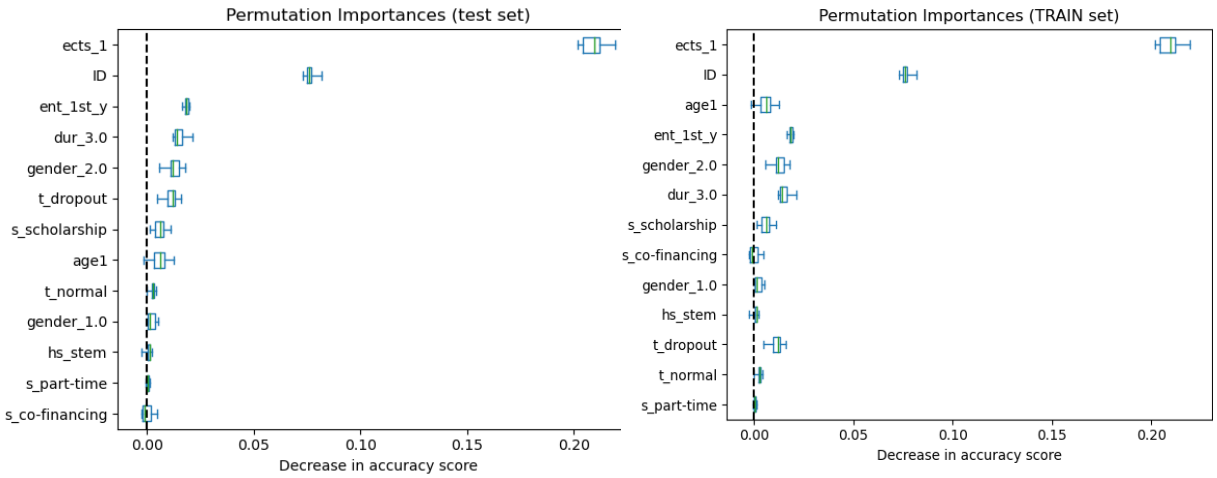


*Figure 50 – RF, end of year prediction, top 13 features, PI on test and train sets.**

*The RF is resource consuming model: it takes 10 minutes to calculate PI importance on test set, and far more for train.
Source: Author.

The model evaluate first 28 variables as important according to PI. The RF had memorized (overfitting) the train set data to much, which is evident at PI test and train charts (Figure 50) at the end of the year, using the 13 most important variables.

SHAP values on top 13 variables have the same direction and dispersion as in HGBC at the end of first year, with exception of *dur_3.0* variable which is straightforward as in HGBC SHAP importance (Figure 51). This variable is *lying down* equally on both sides in RF, unlike in the HGBC case (right for dropout, left for non-dropout classification). Still, the high number of non-dropout occurrences is in the middle. Unlike at HGBC PI, there are new variables among top 13 in RF model SHAP: *t_normal, s_co-financing, hs_stem* and status *s_part-time*.

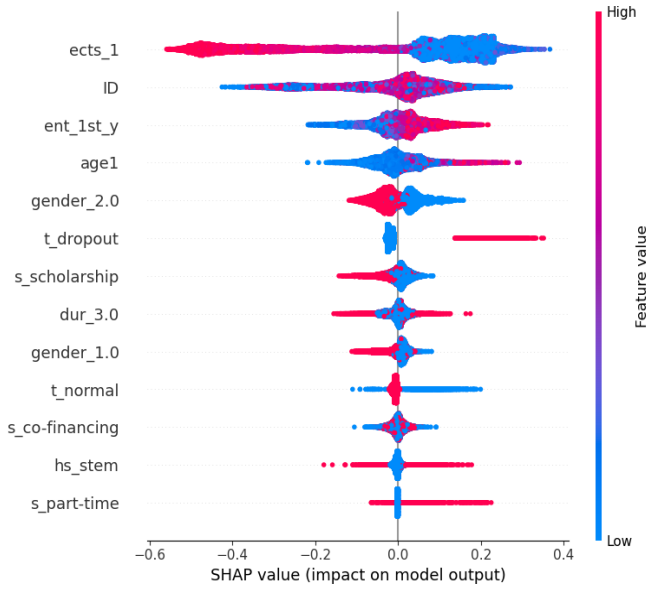


Figure 51 – RF: SHAP global importance at the end of first year, top 13 variables, test set.

Source: Author.

In summary, according to RF, the student has higher chances to be classified as dropout if he/she has lower ECTS score, higher school year of enrollment (i.e. generation, cohort), is older than average student's age, is not a female (or has lack of label), is not a scholarship holder, do not have normal type of enrollment, comes from STEM high school and is a part-time student.

6.4 SVM performance evaluation

The summary statistics of SVM models evaluation shows controversial results: SVM with different kernels has showed one of the lowest performance and one of the best metrics of all trained model. The model that took the last place in general, among all models, was SVM with sigmoid kernel. However, the highest achiever regarding the recall at the end of first year of study (even higher than HGBC) had linear and RBF (gamma = 0.1) kernel (Table 32). Those achieving at the same time took price in low precision and low specificity, which was not the case in HGBC.

However, the SVM by all kernels was big time consumer regarding the training and evaluation phase. The time for fit the model was between 56 seconds and 13 minutes, depend on kernel and data set (the number of inputs). The time for performing the permutation importance on test set was between 1 and 50 minutes, while for calculating the SHAP importance it takes 1 to 5 minutes in average per one sample.

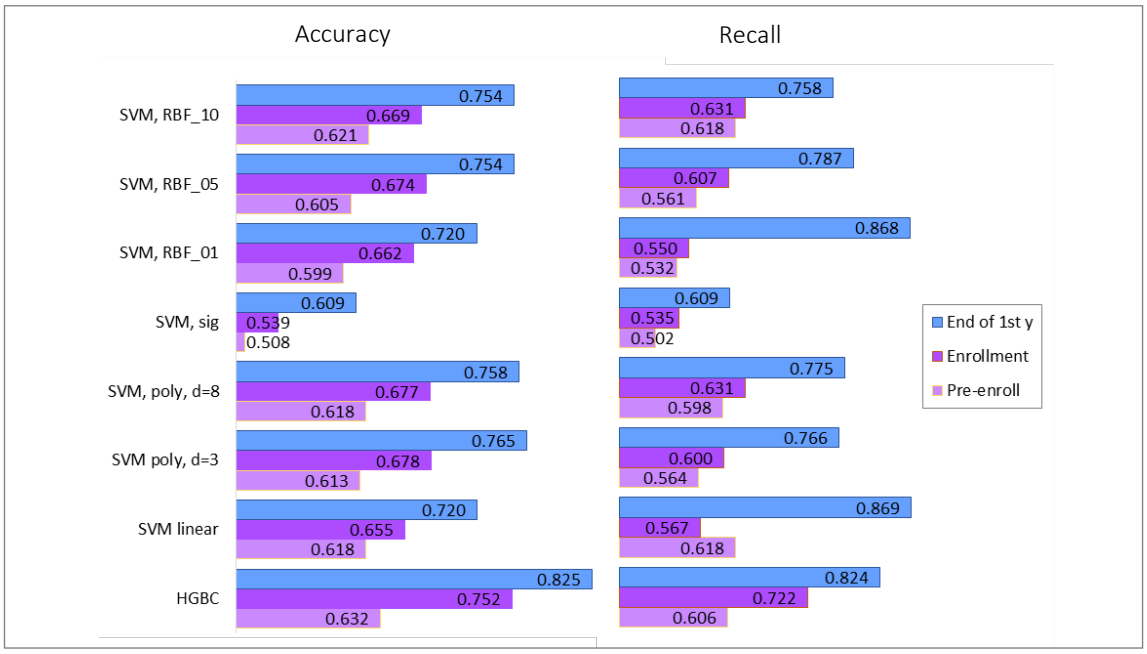


Figure 52 – SVM model accuracy and recall by all kernels and data sets, compared with HGBC

Source: Author.

It is also worth of mention that while training the SVM model with linear kernel on pre enroll data set, model showed higher performance on non-scaled data, in spite of fact that SVM can't handle the non-scaled data, especially the linear kernel. The reason may be in the type of input variables, which all were binary coded, with exception of year of

(potential or expected) enroll into freshmen year, *ent_1st_y*, (i.e. generation, cohort), that takes values between 2007 and 2013. The metrics on scaled set were far worse in some cases: for example the recall was 0.52810 while on non-scaled data it is 0.61762 (Table 32).

Table 30 – Summary for SVM model, kernel: linear.

SVM, kernel: linear	Pre enroll*	Enroll	End of 1st year	End of 1st year (top N)
Accuracy	0.61847	0.65456	0.72023	0.72023
True Negative	1295	1550	1198	1198
False Positive	796	541	893	893
False Negative	790	895	270	270
True Positive	1276	1171	1796	1796
True Negative	31.15%	37.29%	28.82%	28.82%
False Positive	19.15%	13.01%	21.48%	21.48%
False Negative	19.00%	21.53%	6.50%	6.50%
True Positive	30.70%	28.17%	43.20%	43.20%
F1	0.61672	0.61990	0.75542	0.75542
Precision	0.61583	0.68400	0.66791	0.66791
Recall	0.61762	0.56680	0.86931	0.86931
ROCAUC	0.61847	0.65403	0.72112	0.72112
Matthews corr.	0.23694	0.31297	0.46264	0.46264
PR score	0.66111	0.75017	0.78449	0.68804
Specificity	0.61932	0.74127	0.57293	0.57293
Train accuracy	0.61414	0.66129	0.71958	0.71946

*The only exception where data for SVM were not scaled. The scaled data shows accuracy of 0.5980, recall of 0.5281, and F1 of 0.5661.

Considering the polynomial kernel, by adding more degrees (from 3 to 8) it slightly improved the most of model's performances, but both models did not stand out by general performance among others SVM kernels, and the rest of non-SVM models (Appendix, Table A30, A31).

The sigmoid kernel had lowest performance, taking into account all models, where almost all metrics took values between 0.50 and 0.60 for all data sets and time of prediction (Appendix, Table A32). Model is interesting due to its results in permutation importance and SHAP values, when is compared with others.

Analyze of RBF were gamma parameter was tuned in three ways: 0.1 – 0.5 – 1.0 shows increase in overall test and train accuracy, recall, ROC AUC and precision. That success is quite straightforward in the rest of evaluation metrics and with all three data sets (Table A33, A34, and A35 in Appendix). The modest performance has been recorded on pre enrollment data using gamma = 0.1. Performance of that model (gamma = 0.1) were

interesting due to increase in False Negative observations which affect specificity and precision. It is interesting that recall moves up on pre enroll and enroll data sets by tuning the gamma parameter from 0.1 to 0.5, which is not the case at the end of first year data set. On the end of first year (top N) variables data set, the highest recall occurs at gamma = 0.1 (0.86834) and falling down to 0.75750 when gamma reaches the 1.0.

Common to all SVM models is the evaluation of *dist_between_81_and_160*, *dist_0*, and *mld_1* variables in the top 5 in pre-enroll set, which are outside the top 20 in the HGBC model. The next common thing is listing the female gender at the first place by importance, with the exception of sigmoid kernel. Sigmoid kernel aligned only 9 variables as important in pre-enroll set (Appendix, Table A24). In general, the top 5 in pre-enrollment are female gender, *ent_1st_y*, male gender, *mld_missing, dist_up_to_80, dist_0*, and *mld_1*.

In the enrollment dataset, excluding the sigmoid and polynomial 8-degree functions, there are no variables that rank in the top 5 and surpass HGBC in the top 20. The top 5 variables are once again dominated by female gender, *s_scholarship, t_normal, gender_1.0*, and *ent_1st_y*. The *ID* variable only appears in the SVM polynomial, 8-degree. Additionally, most models include *s_co-financing* in their top 5 variables, which ranks 19th in HGBC's list (Appendix, Table A25).

At the end of the year, the PI with all variables revealed a surprising fact: the model with the highest performance among all SVMs (using a linear kernel) in terms of recall showed that only 3 variables were important. These variables are ECTS credits 41-60, *t_dropout*, and ECTS credits larger than 60. The next SVM model with the highest recall (using an RBF kernel with gamma = 1.0) placed the following variables in the top 5: ECTS score 21-40, ECTS score less than 20, *s_scholarship, dur_3.0*, and *ID*. The remaining SVM models also followed a similar pattern, with ECTS score 41-60 and *s_co-financing* being placed in the top 5 (Appendix, Table A26).

6.5 NN performance evaluation

Due to the stochasticity of the NN, the training and testing phases were repeated five times for both, the pre-enrollment and enrollment datasets. The average values are reported in the model summary tables (Appendix, Table A36, A37). For the end of the first-year dataset, a single run was executed, and the top N results were selected to enhance the model's performance through simplification. The model was only run once at the end of the year because the PI and SHAP in all five runs in previous cases show

small difference in variable importance rank. Adding hidden layers did not significantly contribute to the overall accuracy in the test set, and the ROC AUC values were consistent for models with 2-3 and 4 hidden layers (Figure 53). Generally, adding a third hidden layer slightly improved the metrics, while adding a fourth layer resulted in worse performance.

When analyzing the pre-enroll data set, it is evident that additional layers did not significantly improve the metrics, except for a slight improvement in recall (Figure 53, Pre-enroll). The test and train accuracy, ROC AUC, and precision remained almost identical at 0.62 8 (Table 33-35, pre-enroll (avg) accuracy).

However, there was a trade-off between the upper and lower parts of the confusion matrix. Adding layers led to a slight increase in recall and a decrease in specificity, meaning that the model identified more non-dropout students as dropouts, while reducing the number of incorrectly identified dropouts.

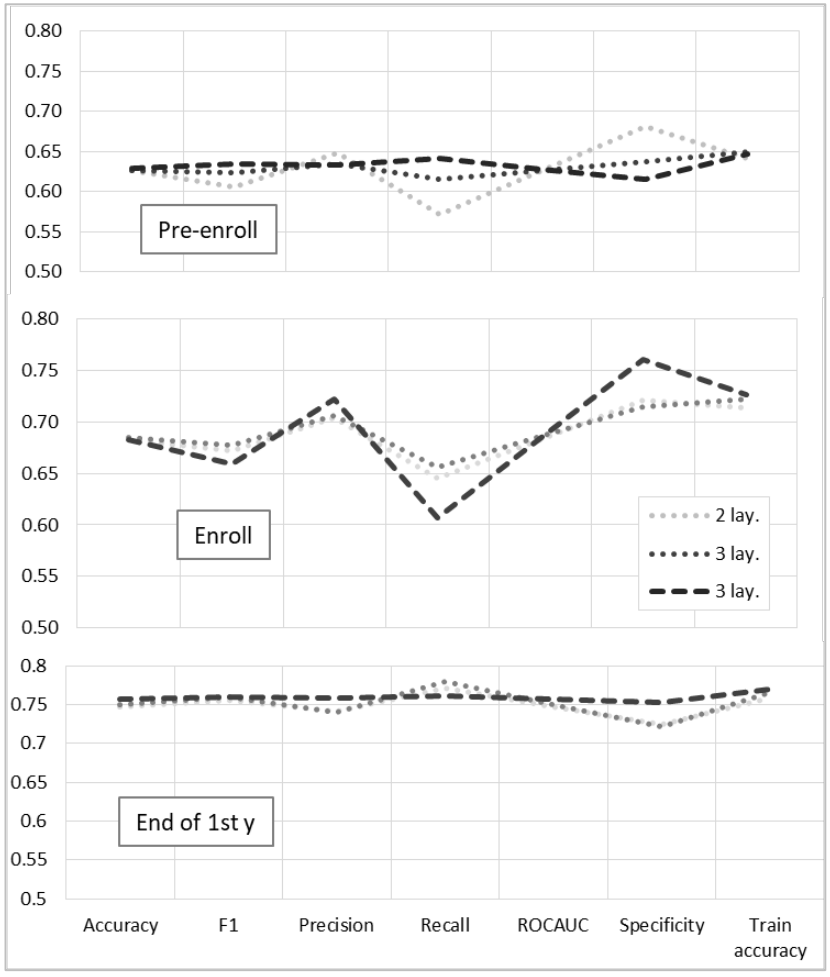


Figure 53 – Summary of NN with 2-3-4 hidden layer in three prediction times.

Source: Author.

In relation to the enroll dataset, the situation mirrors that of the pre-enroll set. The train and test accuracy, as well as the ROC AUC, consistently hold at 0.68 for 2-3 and 4 hidden layers. However, adding the fourth layer led to a decline in recall. The model showed no improvement when transitioning from pre-enroll to enroll variables, even after adding known features from the enroll week (Figure 53, Enroll).

The end of first year data set showed a slight improvement in both test and train accuracy (0.74-0.75) as well as in ROC AUC, with almost identical values for the 2-3 and 4 layer models. However, adding the 4th layer led to a small decrease in recall using top N features at the end of the school year (from 0.77894 with 3 hidden layers to 0.76084, Table 34 and 35), but it significantly improved precision and specificity (Figure 53, End of year).

Table 31 – Summary of NN model, with two hidden layers.

NN 2 layers	Pre enroll (avg)	Enroll (avg)	End of 1st year	End of 1st year (top N)
Accuracy	0.62585	0.68247	0.76962	0.74771
True Negative	1401	1481	1621	1489
False Positive	654	574	434	566
False Negative	900	745	523	482
True Positive	1199	1354	1576	1617
True Negative	33.72%	35.65%	39.02%	0.35845
False Positive	15.75%	13.82%	10.45%	0.13625
False Negative	21.66%	17.93%	12.59%	0.11603
True Positive	28.87%	32.60%	37.94%	0.38926
F1	0.60628	0.67165	0.76710	0.75525
Precision	0.64788	0.70433	0.78408	0.74072
Recall	0.57132	0.64507	0.75083	0.77037
ROCAUC	0.62644	0.68288	0.76982	0.74747
Matthews corr.	0.25500	0.36831	0.53989	0.49556
PR score	0.69336	0.78446	0.86384	0.84094
Specificity	0.68156	0.72068	0.78881	0.72457
Train accuracy	0.64200	0.71348	0.78301	0.75792

Source: Author.

Feeding the NN took 10-12.5 seconds, while calculating SHAP importance took 25 seconds to 1 minute and 10 seconds, depending on input data and number of layers.

Table 32 – Summary of NN model with three hidden layers.

NN 3 layers	Pre enroll (avg)	Enroll (avg)	End of 1st year	End of 1st year (top N)
Accuracy	0.62600	0.68527	0.76938	0.75036
True Negative	1309.8	1470	1599	1482
False Positive	745.2	585	456	573
False Negative	808.4	722.4	502	464
True Positive	1290.6	1376.6	1597	1635
True Negative	31.53%	35.39%	38.49%	35.68%
False Positive	17.94%	14.08%	10.98%	13.79%
False Negative	19.46%	17.39%	12.08%	11.17%
True Positive	31.07%	33.14%	38.44%	39.36%
F1	0.62343	0.67690	0.76927	0.75923
Precision	0.63491	0.70593	0.77789	0.74049
Recall	0.61486	0.65584	0.76084	0.77894
ROCAUC	0.62612	0.68558	0.76947	0.75006
Matthews corr.	0.25322	0.37476	0.53895	0.50108
PR score	0.69516	0.78858	0.86418	0.84784
Specificity	0.63737	0.71533	0.77810	0.72117
Train accuracy	0.64944	0.72240	0.78482	0.76592

Source: Author.

Table 33 – Summary of NN model, with four hidden layers.

NN 4 layers	Pre enroll (avg)	Enroll (avg)	End of 1st year	End of 1st year (top N)
Accuracy	0.62836	0.68310	0.75806	0.75710
True Negative	1265	1563.2	1728	1548
False Positive	790	491.8	327	507
False Negative	753	824.6	678	502
True Positive	1346	1274.4	1421	1597
True Negative	30.44%	37.63%	41.60%	37.27%
False Positive	19.03%	11.84%	7.87%	12.21%
False Negative	18.14%	19.85%	16.32%	12.08%
True Positive	32.39%	30.68%	34.21%	38.44%
F1	0.63420	0.65887	0.73876	0.75993
Precision	0.63343	0.72192	0.81293	0.75903
Recall	0.64107	0.60715	0.67699	0.76084
ROCAUC	0.62822	0.68391	0.75893	0.75706
Matthews corr.	0.25905	0.37248	0.52446	0.51414
PR score	0.69357	0.78418	0.86367	0.85167
Specificity	0.61538	0.76068	0.84088	0.75328
Train accuracy	0.64701	0.72652	0.77904	0.77031

Source: Author.

In the pre-enrollment dataset, the top 5 variables are almost equally important for all models with 2-3-4 layers according to both PI and SHAP values. The most influential

variable is female gender, followed by *ent_1st_y*, male, *mld_missing*, and *distance_up_to_80* km. These findings are consistent across all models. Another notable variable is *hs_missing*, which ranks high among the top 10, followed by *dist_0, mld_1*, and *hs_econ* (Appendix, Table A24).

In the enrollment dataset, all PI are placed in a 2-3-4 layered NN with the same first feature, which is female, and *t_normal* as the second feature. However, all SHAP values list *t_normal* as the first feature and female as the second. The top 5 features also include *s_co-financing*, *ID, hs_missing, dist_up_to_80*, and male gender (Appendix, Table A25).

At the end of the first year, the top N variables included the top 5 dominated female variable and *s_scholarship*, which were also in the top 5 at the beginning of the year. The variables that jumped to the top 5 at the end of the year were *ects_1* and ECTS ranks, along with *dur_3.0* which moved up from the top 10 at the beginning of the year to the top 5 at the end (Appendix, Table A26).

7 DISCUSSION AND ANALYSIS OF RESEARCH RESULTS

Although Bosnia and Herzegovina has the necessary resources to ensure prosperity for its citizens, the reality is quite the opposite. In an effort to contribute to the overall well-being of the country, it was decided to focus on increasing the number of highly educated residents as a piece of the larger puzzle. The specific goal of this thesis is to predict, at an early stage, which students are at risk of dropping out of their studies. To accomplish this, the following research questions were formulated 1) What is the size, structure and reasons for leaving the HE at UNIBL, B&H, between 2007 and 2018? 2) How well does a ML model perform when trained on a dataset that is almost entirely binary? 3) How one can effectively enhance the explainability and interpretability of the black box models, to provide a clearer understanding of its outputs?

The data indicate that UNIBL is not spared from undesired trends, such as student's attrition. The overall dropout rate for the observed 12 years is 47.1 percent, but the permanent dropout rate is lower and amounts to 31.8 percent of all enrolled students. This phenomenon shows a growing tendency and data suggest that around 1/3 of dropouts can be prevented because they are institutional.

Based on evaluated model performances, HGBC is the one that best copes with the challenges of a modest data set and predicting dropouts as early as possible with an accuracy of over 80 percent compared to other models. The model also showed advantages when exposed to reductions in the number of variables and changes in definition, and imbalance data. The most influential variables to the models decision of students' classification are gender, generation of enrollment, particular type and status of enrollment, as well as amount of collected ECTS credits.

7.1 Interpretation of results through the prism of research questions and aim

In this section one would present how the results of research presented in previous chapters are aligned with research questions, and existing literature, as well as contribution to the field.

7.1.1 Research question 1

Our contribution includes the first-ever presentation of data on dropout rates at a HEI in Bosnia and Herzegovina, and where dropouts go after quitting. This pioneering effort provides valuable insights into the dropout patterns specific to this context, serving as a crucial reference point for future studies and policy development in the country. In order to answer the first research question, below are presented three segments: the magnitude, the structure, and reasons for leaving the HE.

In obtained dataset were identified several types of churn, and their share in overall 12 years of observed data dropout rate of 47.1 percent was decomposed to withdrawal initiated by student, and the one "initiated" by law and study rules. These two were further segmented to permanent, and temporary dropout. In that way it was precisely identified each type of churn, and its share in overall dropout, because some authors, like (Xavier and Meneses, 2020) highlighted the problems of broadly defined dropout in research. The final 31.8 percent of quitters for good, 2007 – 2018, at UNIBL were reached by decomposing different types of dropouts. The European countries with similar amount of HE churn are Slovenia – 35, Denmark – 30, and Germany with 28 percent. Still, the report from Slovenia and Denmark are for 2014 – 2015 year only, and overall attrition rate in Denmark include the one who continue education (transfers/switchers). Considering the permanent and temporal overall dropout at UNBIL of 47.1 percent, within 12 years of data, it is comparable only to Denmark methodology, and to the Slovakia (42-51 percent) for 2005-2010, and shows huge magnitude of this undesired phenomenon.

Our contribution to the field involves a refined segmentation of the concept and categories of dropout, which allows for a more accurate and nuanced definition of dropout phenomena. This precise categorization serves as a foundational framework for future researchers, enabling more consistent comparisons of results across studies and improving the reliability of dropout prediction models. By providing this structured approach, we aim to enhance the clarity and comparability of research in this domain, fostering a deeper understanding and more effective interventions.

The magnitude of dropout is analyzed across multiple dimensions: by academic year (cohort, i.e. generation of enrollment), gender, faculty, and geographic origin—considering both the municipality's level of development and its distance from Banja Luka, where UNIBL is located. Additionally, dropout is examined by field of study, program length (3- or 4-year), and various student characteristics, including age at

enrollment, admission score, total admission score, number of accumulated ECTS credits, and number of passed exams.

The data suggest that half of churn occurs during first university year, which is in align with data from France (Rajski, *et.al*, 2018). However, there are cohorts were the share of first year dropout was even higher than half, and comparing with the data from Italy (2011-2013) where churn in first year was 6.7 percent (Modena *et al*., 2020), or Norway (Statistics Norway, 2022b) 6 percent (2015-2020), it is an indicator of a significant waste of resources of the students and the University. Analysis by cohort, utilizing Kaplan-Meier curves, reveals that more recent student cohorts exhibit lower persistence rates compared to older cohorts. This trend is particularly significant for UNIBL, as it highlights inefficiencies in resource utilization and underscores the urgent need to implement an Early Warning System and adopt preventive measures to reduce student attrition.

The gender data portrayed female as more persistent than their male peers, through all fields of study even those predominantly male study programs. The overall dropout by women is 39 percent, and by men is 55 within 12 observed years at UNIBL. Comparing to other European countries those findings support existing literature. In its comprehensive literature review (Behr *et al*., 2020, p. 17) listed the cases of university dropout that is lower for female students in Italy, Germany and Netherland, and stated that "*compared to men, women seem to be more motivated, disciplined and have better time management skills…*". (Guzmán *et al.*, 2021) found that female dropout less if they are coming from rural places in Europe, which opens the window for discussion, since a Bosnia and Hercegovina has more rural than urban population.

Looking at dropout rates per faculty, and comparing with available data for European countries, all faculties at UNIBL that were compared to European have higher dropout rates, with the exception of UNIBL Academy of Arts. Data shows that the highest churn rates at UNIBL have Faculty of Mining – 71, Mechanical Engineering – 68, Electrical Engineering – 64, and Agriculture – 62 percent. To compare the Agriculture data with Latvia's Agriculture university dropout, which is only 27 percent (Paura and Arhipova, 2016), and with Sweden's engineering data are between 17 and 21 percent (Kolm and Svensson, 2017). Natural science and Math at UNIBL reached the 53 percent of churn, while in Poland is 48 percent, (Zając and Komendant-Brodowska, 2019). Economics in Poland is 41-42, while at UNIBL is 49 percent. The share of dropouts at 3-year duration

bachelor study programs at UNIBL ranged between 20 and 80.1 percent, while is 43.8 in Romania, 2015 – 2020, (Herţeliu *et al*., 2022).

Speaking of distance from the place where the UNIBL is located, data suggest that the more distanced the place of students origin is, the lower are dropout rates. The second is that with increase of municipality development, the dropout rates increase too. This is in align with existing research which found that students from urban areas are three times more prone to dropout, in Germany, (Behr *et al*., 2020).

Regarding the admission exam score, total admission score, ECTS collected credits and number of passed exam – all suggesting the same: the dropouts' average score is lower than for non-dropouts students. The variable age shows opposite, the dropouts tend to be a bit older than non-dropouts. The same conclude (Behr *et al*., 2020).

This research contributes by presenting a variety of categories of tertiary education dropout within the context of Bosnia and Herzegovina, observed over a 12-year period. This extensive timeframe classifies the study as longitudinal, offering a comprehensive view of dropout patterns in this region.

For every university it is very important to have at least insight in the structure of reasons for churn, because there are reasons where university could have direct, or partial contribution to affect decision to stay. The analyses of reasons for leaving the HE at UNIBL suggest that around 1/3 of leavers is possible to prevent, since they are directly influenced by University as institution. (Guzmán *et al*., 2021) found out that this share in Europe is 13 percent.

The qualitative part of the research indicate that the main reason for leaving the UNIBL was dispute or conflict with lecturer/professor, 15.6 percent. On the second place with 13.5 percent of dropout is caused by financial reasons, and the same share goes for reason: inability to work and study in the same time. (Kehm *et al*., 2019) placed the financial reason and working while studying in external conditions that may cause the churn, and provided studies which showed that work while study increase the chance of churn.

The financial aspect serves as the intersection between the qualitative survey analysis and the quantitative research using ML modeling. Both approaches emphasize the importance of financial factors in student churn, despite the relatively low tuition fees.

The dispute or conflict with lecturer/professor as reason for attrition was not found in the literature review by other researchers.

The survey respondents reported and other reasons for dropout: personal (almost 2/3 had mental overwhelm, psychological unpreparedness, lack of occupation, or motivation),

and institutional reasons (around half of dropouts is partially caused by insufficient internships, outdated study programs, and boring teaching methods).

Where the students go after dropout and their satisfaction with decision to leave is the last part of the qualitative research in this thesis. Around 40 percent of quitters left HE for good. The rest of quitters continue education at private HEI in the country (1/3), abroad (1/3), and at the UNIBL (1/3). The rise of number and quality of study programs at private HEI in the country pull the share of public university students to the private ones. Also, the large offer of scholarships for study abroad impact the decision to leave the HEI in the country. There is 13 percent of quitters who are not satisfied with their decision to quit the first study they were enrolled in, and 19 percent of them thinks they would have now higher income if obtained the quitted degree.

The contribution of this first-ever research on the reasons for leaving higher education in the country lies in providing valuable insights into the structure and causes of student dropout, as well as a preliminary understanding of their trajectories in the years following their departure.

7.1.2 Research question 2

Looking at the size of the phenomenon of study interruption at UNIBL, and comparing it with other universities in Europe, and the seriousness of the repercussions that the interruption of studies leaves on the individual, the university and society as a whole, there is a need to approach this problem as soon as possible. The way we approached its solution is a model for early prediction of study interruption, even before the student makes the decision to leave. To achieve this, we trained 78 ML models, and ran an additional 11 experiments with the best model, in order to be sure of its quality.

We have run our models using pre-enrollment data with 27 features, and one added 10 more at enrollment week, plus additional 7 at the end of the year. In total each model was fed with 45 variables, and on those 45 variables who were known at the end of first study year - was done feature reduction by PI and SHAP to improve the model performances. The feature reduction to improve performance was not done in previous time points, at the beginning of the year, and prior to enrollment, because the dataset has modest variables, and lack of sociodemographic features, like occupation and education of parents, and student housing, as well as academic data. Other researchers used 77 variables in total, while the best results achieved by redaction to 15 (Márquez-Vera *et al*.,

2013), 19 feature (Ghorbani and Ghousi, 2020), 39 features (Delen, 2010), 34 features (Thammasiri *et al*., 2014), 27 crucial of 40 variables (Kim *et al*., 2023).

Regarding the size of data set, the 37 thousands records, between 2007/08 and 2018/19 academic year, were each record represents one student, were used for churn estimation, and 20 thousands students, since 2007 – 2013, for ML modeling due to request to follow each student at least 6 years in a dataset to be sure he/she is a dropout. In the literature review we found only several dataset of tertiary education sized more than 10 thousand: 39 thousands (Delen *et al*., 2023), 32 thousands of UK student (Waheed *et al*., 2021), 25 thousands (Delen, 2011), 23 thousands (Chai and Gibson, 2015), 21 thousands (Thammasiri *et al.*, 2014), and 16 thousands (Delen, 2010). The most of them are between a few hundreds to few thousands.

The specificity of obtained data set are missing data, categorical variables and quite modest amount of study and socioeconomics related variables. The dataset of predictors contains 41 binary variable, three numerical and additional three numerical variables with missing data (total admission score obtained by admission exam and high school GPA, *score_t,* score at admission exam, *score_e,* number of passed courses at the end of first year, *npe_1*) which served as input only to HGBC model due to "indigestible" data for the rest of models.

The highest achieved accuracy in all three times of prediction is by HGBC (0.6317, 0.7522, and 0.8254). The highest accuracy among other researchers were: 94 percent by DT (Perez *et al*., 2018), 91.2 percent by DT – J48 (Pérez *et al*., 2019), 90.9 with DT-ID 3 (Pal, 2012), but using ensemble models only 63.6 percent for ensemble tree with AdaBoost (Patacsil, 2020). The highest achievers in the literature were models based on DT, RF, NN and SVM, while there is no cases of use HGBC in dropout classification prior to this research.

As we are more interested in the number of dropouts than those who are non-dropouts, the recall measure together with precision and ROC AUC on dataset where both classes are approximately equally represented was our primary concern. The HGBC with top 10 features at the end of first year reached the 82 correctly classified students of 100 who really dropout within at least six years from enrollment. The model with data in enrollment week had recall of 72 of 100 dropouts, and with pre-enrollment data 60 of 100 dropouts. Still, model had slightly lower recall in pre-enrollment and end of year churn prediction comparing with others, but demonstrated higher quality than SVM and NN which showed only higher recall than HGBC. If one insist on recall, the possible approach

is usage of different models in pre-enrollment and end of year prediction, having in mind the high cost respecting the evaluation time, models quality, and small differences in obtained results by HGBC and other models.

Even on imbalanced dataset as consequence of experiment with change in definition to reduce the level of definition's rigidness by meeting the dataset feature label of dropout, the occurred imbalance was 72:28, with top 10 features used, the HGBC showed good-enough recall, 75.6 percent, and precision 61.6 percent at the end of first year. Data acknowledged that adding more variables and using up-sampling strategy improved the model on imbalanced data. This is in align with (Chai and Gibson, 2015) findings. Authors tried to predict dropout in HE at the end of first year using set of 23 thousands of students, had imbalanced data, 83:17, with rich in feature dataset, were in total had 164 selected features, and no missing data. The highest achieved precision was with RF, 71 percent, while the highest recall was 37, with DT, that is far lower than our results.

Our contribution lies in successfully developing high-performance model using a dataset characterized by modest number of variables related to student pre-academic and academic performance, extensive missing data, and predominantly categorical variables. Despite the limitations of the dataset — particularly the scarcity of variables related to student academic performance and socioeconomics features — we achieved strong model performance through effective handling of missing data. This demonstrates the robustness of our methodological approach, even in the face of significant data challenges.

Another contribution of this research is the novel comparison of the Histogram-Based Gradient Boosting Classifier (HGBC) with RF, NNs, and SVM in modeling higher education dropouts. This comparison has been conducted for the first time in this context, offering new insights into the relative performance of these models.

Although in the existing studies of study discontinuation in HE, the highest results were achieved using RF (Rodríguez-Maya *et al*., 2017), NN (Siri, 2015), (Delen, 2011) and SVM (Weng Fu Mei, 2010), on rich in features datasets, and no missing data, our results suggest that HGBC demonstrated superiority when is compared with RF, NN and SVM on challenging dataset, respecting the overall accuracy, ROC AUC, time needed for model evaluation and calculating the PI and SHAP values.

7.1.3 Research question 3

The answer to the third research question was provided by employing two model-agnostic machine learning interpretation techniques—Permutation Importance (PI) and SHAP

(SHapley Additive exPlanations) – to better understand the factors influencing the classification of student dropout after the models were run, i.e., post-hoc on the challenging dataset. Additionally, prior to modeling, we conducted a correlation analysis (pre-hoc) to gain preliminary insights into feature importance and set expectations. The use of multiple interpretability methods enhances the comprehensiveness of our understanding of the model's decision-making process by revealing the underlying impact of different features.

With the passage of time and the introduction of new variables into the model, some variables from pre-enrollment dataset remained in the top 5 until the end of year, while some were replaced by those introduced in enrollment and at the end of the year. The variables which have the most frequent occurrences in general, at the end of the year, are female gender, whether student is a scholarship holder or not, whether is in status of study co-financing or not, artificial variable ID, its belonging to the study duration of 3- or 4-years, and does he/she have certain ECTS score.

The other researchers who also used university dataset for churn prediction and found out the same or similar variables among the most important ones, are listed below.

According to (Baghernejad, 2016) students age, and scholarship (Thammasiri *et al*., 2014), (Delen, 2010) are among the most important predictors for most of students located in USA. Besides those two, also highly important are financial features like family income, loan status, different types of financial aids, and ethnicity, which our dataset does not have.

In our data set student's age was introduced with pre-enrollment features and was in top 5 at HGBC by PI and by SHAP in that time of prediction. It fell to sixth to tenth place of 39 in enrollment and sixth to tenth place of 48 in end of year prediction by HGBC, for both: PI and SHAP. It has high significance (top 5) in pre-enrollment dataset only for few models, while in enrollment phase of modeling reached top 5 only in RF.

The *s_scholarship* status was introduced in the enrollment phase of modeling, at the beginning of the year. In that stage of prediction it was in top 5 by all models except the RF, SVM linear kernel, and NNs. The other SVM models ranked the scholarship in first or second place, even higher than HGBC (4^{th} place by PI and SHAP). It is interesting that almost all NNs does not consider this variable in top 10. SHAP analysis at enrollment and at the end of year indicate that students who are not scholarship holders have higher chance to be classified as in-risk of attrition students (Figure 45, right). This information may be valuable for University and policymakers to manage the quotas in effective way

to support additional number of students or desirable study programs. The UNIBL has good approach to this in a way that every student who passed all courses during first (or any sophomore) study year fells in the category of scholarship holder and in every follow school year may transfer up to 15 ECTS credits in to the next one (it means he/she does not have to pass all courses in the current school year). The way to act in align to this information is to find the way to help more students fell in the category scholarship holder by establishing support programs of graduation by example, students as mentors, and teachers as mentors who will regularly evaluate the students persistence.

There are studies which found out that among the most influential features to the model prediction of churn in Bangladesh (Mustafa *et al*., 2012) and USA (Aguiar, 2015) are gender and age, bat study in Mexico claims that gender and age have the less predictive power (Pérez *et al*., 2019). For (Patacsil, 2020) the gender variable was also less important.

The gender variables were introduced with the pre-enrollment features and female gender remained among top 5 in all three stages of prediction, i.e. with time passing. This means that at the UNIBL the woman students are more persist ones than their men colleagues, which was supported and by results of churn analysis in first research question. Having in mind that female students dropout more in absolute numbers than men, due to gender share of 60 percent at UNIBL, the possible steps to act upon are support programs and groups for pregnant students and those with children.

The place of residence (Weng Fu Mei, 2010), distance from the university (Siri, 2015), and county of residence (Ameri, 2015) were features of highest influent to the model churn decision. In our case the one variable regarding the distance from UNIBL stands among others by its frequency of occurrence among top 5 and top 10: *dist_up_to_80* km. Our data suggest that if student's place of habitation is closer to UNIBL (i.e. by up to 80 km) it is more in risk to dropout from the HE. The variable *dist_up_to_80* was in top 10 in pre-enrollment data by all models, except for RF PI, and SVM, sigmoid, PI. It slightly fell in the range of top 10 in enrollment stage by all models, and even bellow top 10 for HGBC, RF and SVM, polynomial, 3-degree. At the end of the year prediction it was present only in NN 2- and 3-layers, in 13th place among top N variables. This is important because if those students are commuters, it increase their costs in time and financially.

The admission test score and high school GPA were among the most important predictors for (Pal, 2012), (Patacsil, 2020), and (Ameri, 2015), which also highlighted the ECTS score. In our data set admission test score is *score_e*, while *score_t* encompass the *score_e*

plus high school GPA multiplied by 10. Those variable have more than 80 percent of missing data, thus are only input to HGBC. At the end of the year they were in top 13 the most important variables by HGBC, ranged in 11th, and 12th place by PI, and 11th and 13th by SHAP importance. Still, in experiment of running the HGBC without variables that contain the missing data, their deficiency had not significantly decreased the models performance and quality (Table A23). The ECTS variable had huge amount of zero values, thus we introduced the ECTS ranks which were one of the most frequent feature in top 5 at the end of the year by PI and SHAP for all models, and we left *ects_1* as predictor too, even it was the source of derivation of ECTS ranks variables. As expecting the HGBC and RF ranged *ects_1* in 1st place by importance (PI and SHAP) at the end of the year. The same goes for NN 3- and 4-layers, and SVM, polynomial 8-degree. The NN 2-layers and SVM, polynomial, 3-degree, ranged it in 2nd place. The rest of SVM models did not reported importance of *ects_1* in top 10, or at all. Regarding the ECTS ranks, there are three models which did not reported the ECTS ranks as important ones at all in the top N: HGBC, RF and SVM, polynomial, with 8-degree. All other models ranged ECTS scores rank among top 5 the most influential to the model outcome.

In addition to the above variables, our models found some other important variables: year of enrollment, i.e. cohort of students – *ent_1st_y*, duration of study program, *dur_3.0*, artificial variable, *ID*, partial scholarship holder status at enrollment, *s_co-financing*, and as expected – the *t_dropout* status which occurs during the school year and is initiated by student.

Scholarship status, including whether a student is a scholarship holder or co-finances their studies, has been identified as a highly important feature. This finding underscores the financial reasons for student churn, which aligns with insights from the qualitative portion of the research.

We observed that some models, despite their overall poor performance, yield feature importance results that are comparable to those of the highest-ranked model, the HGBC. The NNs, SVMs, and RF demonstrate similar variable importance results to the HGBC at the end of the year. However, despite the strong rationale provided by feature importance analysis, those models not perform as well as the HGBC when evaluated on other metrics.

The results obtained from PI and SHAP are closely align with the correlation analysis conducted prior to modeling, which included all variables. This alignment suggests that

correlation analysis can provide valuable and reliable insights into variable importance, reinforcing its utility as a preliminary step in model interpretation.

This study contributes by demonstrating that, even with a challenging binary dataset, the correlation analysis aligns quite god with the results from PI and SHAP. However, it also reveals that different models do not consistently rank variables in the same way. Still, some models have very similar outcome, like HGBC and RF. By comparing the feature importance outputs across all models, by variety of interpretation techniques, we gain a more comprehensive understanding of the reasons behind the model's decisions, particularly in classifying students as dropouts or non-dropouts. This approach offers a deeper understanding of the factors influencing model predictions at global and individual level, thereby enhancing the confidence in the decision-making process and providing a solid foundation for further model improvement.

7.2 Limitations of the research

The limitations associated with the first research question primarily refer to the scope of the undergraduate, bachelor student sample used in this study. Specifically, the research is based on data from UNIBL, which represents only approximately 20 percent of higher education students in the country. Additionally, the data span from the 2007/08 academic year to the 2018/19 academic year.

A significant issue is the large volume of missing data, exacerbated by the fact that e-services have not been utilized uniformly across all university members since 2007, such as at the Faculty of Medicine. This lack of comprehensive data limits the ability to fully explain, understand, and analyze patterns of study interruptions.

Another concern is the absence of certain variables in the dataset. Although these variables exist in the database, they are not included in this dataset because they were not recorded by the administrative staff. Other researchers have identified variables such as parents' age, education, and occupation; students' housing; ethnicity; GPA in the last year of secondary education, and financial status as important factors in predicting student churn, and thus why our recommendation to add them into modeling is.

Also, there is noticeable a certain amount of errors by human during data entry for date of birth, date of enrollment, and place of students origin, as well those caused by data migration.

The following limitation is regarding the choice of definition of churn: The sample of "blank status in the database" students who were interviewed is modest, as it includes

only 10 people, and indicate the possibility of not capturing the important information which are used to define the dropout.

The dataset variables, University study rules, the long term nature of decision to continue the HE, and choice of our dropout definition contribute to the underestimation of the permanent dropout of students from the oldest generations, 2007-2010, which are classified in category "study" (Figure 23). Sankey's chart shows that some students from the oldest generations are present in the database since 2007/08 school year and still did not graduated or having dropout status. We are aware that discontinuation and continuation of HE may be the long term decision, still the same share of permanently dropout student (1/3) obtained by surveyed dropouts by their request is applied within the whole data span of 12 years which may be misleading.

The small sample of surveyed students who withdrawal at their request, 96 in total, is not large enough to make generalized conclusion, but despite its size limitations, brings valuable insights into share and reasons of churn.

Limitations regarding the second research question, i.e. the modeling and evaluation of classification ML models are related to the data span from 2007 – 2013, with a lot of missing data in several variables (admission scores, ECTS credits, gender label, date of birth, place of students residence while studying, and student's origin etc.) It was not possible to connect the students' dataset with students' grades due to fact that course records started in the 2014/15 school year, and ML modeling set is limited to 2007 – 2013. It was not possible to trace students who were enrolled into more than one study program, i.e. faculty in a dataset due to lack of unique identifier which is recorded in a database but was not provided by the University, due to policy of information safety. However, the answer of the University was that those cases are rare and statistically insignificant.

The usage of ML models is limited to HGBC, RF, SVM's with 7 different kernels and NN with 2-, 3- and 4-layers, which all are black box models, with no opportunity for employment ante-hoc interpretability, like it is possible for DT.

Limitations regarding the last research question are related to interpretation of importance the ID variable which is artificial, still it captured the important information of students' admission success, study program and faculty.

While SHAP importance at the individual level offers valuable insights into the classification of students as at-risk or not, it is limited in its practical application for

students. In its current form, it does not provide actionable guidance on what steps a student should take to reduce their likelihood of dropping out.

Despite its limitations, this research highlights the significant potential of machine learning models in accurately predicting early-stage student dropout in higher education, and novel comparison of HGBC with rest of ML models, even when working with a challenging and modest dataset. Additionally, it provides the first preliminary data on the types and proportions of student dropouts in higher education in Bosnia and Herzegovina.

7.3 Implications and recommendations

In this section one will to present the importance and significance of thesis results, followed by its appliance in practice and for future research potential. At the end one will be provided specific and feasible recommendation for further research and improvements of this work which were addressed in the limitation chapter.

The magnitude of HE interruption at UNIBL suggests the urgency of bringing it into the focus. Since half of the total attrition occurs during the first year of study, this is where most efforts should be concentrated through:

- Establishing the support and graduation programs as some universities in Europe do. The students who participate in those programs have higher persistent rates than those who don't, because those programs increase their motivation to obtain degree.
- Having student as mentor, and professor as mentor for small group of in-risk students from beginning of the school year, who will regularly assist regarding the time management, the way of approaching to the different courses, and direct the student to keep track with study requirement.
- To analyze in which week during semester in first year the attrition occurs and what are the reasons for. In that way one can take preventive measures to prevent the discontinuation.
- The social events where future students may meet each other, professors and sophomore students, before beginning of the class, especially those who comes from undeveloped, distanced towns, and rural areas.
- Having organized supporting groups guided by professionals from Faculty of Psychology.

- Taking into account achievements in the secondary education or acknowledge the credits that student could took before enrollment, by providing summer schools for secondary education pupils.
- The wrong choice of field of study in HE may be reduced by providing more information to the students while they are in secondary education, by online and offline evens which involve inclusion of current students too.
- Harmonization of the secondary education courses' content and university syllabuses, as one of proposals, in order to prevent churn due to lack or modest previous education of domain knowledge.
- Establish the mandatory regularly education of academic stuff in order to build up their presentation and communication skills to make the class more attractive, by making it interactive, interesting and contemporary. The current process of end of semester evaluation of academic stuff by student, make transparent and rewarding for the higher achievers.
- To ensure more opportunities for students' internship in semester and during the semester breaks. Also, due to the reported cases of abuse, mistreatment and corruption, it is necessary to make the existing procedure for reporting and protecting such cases more safe and protective to students when those cases are reported. One possible way is to ensure in some part of the process the anonymous of the student identity.

Acting according to the recommended ways, the University may reduce one third of churn which are institutional, but the University may have partial contribution and impact to the decision of staying in HE and when the reasons for leaving are caused by student's personal characteristics, too. This may be achieved by tacking the preventive measures and actions before, while and after students' enrollment, listed above.

To improve the analysis and prevention of study interruptions, we recommend the following: (a) Create a new field in the database where students can enter the reason for interrupting their studies through the e-Student interface; (b) Update records in the database with reasons for study interruptions based on hard copy records from the Student Service for previous generations and add the date of dropout; (c) Completion of an online standardized questionnaire by a student who is interrupting his studies. In some ways, students need to know that someone cares about their future, and that support can help them endure tough times. (d) Make information of study duration publicly available from

the beginning of their establishment, for each study program and academic year. (e) To reduce the amount of missing data, one can allow to students that they enter all their data using e-Student service in enrollment week, and in the case of any errors, which may be noticed later, the student has to pay some amount of fee in order to be corrected by the administration stuff, otherwise will not be able to receive diploma.

One way to improve the qualitative part of this research is by creating survey and interviews on larger sample of dropouts, with systematically prepared and evaluated questionnaire which should be evaluated not only by the UNIBL staff, while students too. The underestimation of churn definition may be overcome by its change, by aligning it to the most often in the literature – 6 years after enrollment for 4 years study duration, or by aligning it with more strict OECD definition, which we provide on one part of our sample, for the sake of comparison. Another possible solution is to meet the rules from all public HEI to establish the dropout definition.

The way to overcome limitation regarding the second research question: having data scope of students with 3- and 4-years bachelor study duration, without 5- and 6-years may be overcome by zooming in to the level of faculty, or by asking more recent data from the UNIBL.

Respecting the third research question the improvements may be done by creating the interactive feature importance reports in all three times of prediction by employing Tableau, Excel, Python or LaTeX to alleviate the decision making process. The research may be upgraded by exploring the pattern of students churn with employment of unsupervised learning models. Also, the cohort, i.e. generation of enroll students differences among each other's and there is a space for analysis that may provide useful new knowledge, because there are expected changes with time passing. Those changes may discover the different feature importance by cohorts, and help to understand what drives churn in recent years. Adding semester data of students' progress (first semester courses grades, amount of class attendance) and setting prediction at the end of first semester may add valuable contribution to understanding not only dropout, while success too. The further way of exploring the ML models is to evaluate their performance on different parts of dataset by exploring where the higher accuracy come and why, by employing the Responsible AI Python's library. At the individual level, visualization techniques could be further enhanced to not only depict the current risk of attrition but also to offer specific steps for mitigating that risk. This thesis presents the first-ever evaluation of higher education churn in the country, using a sample representing 20

percent of the student population. The findings reveal a high and increasing dropout rate at UNIBL during the observed period. Importantly, the study suggests that at least one-third of these dropouts could be prevented by implementing the recommendations provided. The proposed Early Warning System is capable of accurately predicting dropout for at least 60 out of 100 dropouts, even before their decision to leave, i.e., prior to their enrollment at UNIBL, which provides high potential for in time prediction and successful prevention of churn.

8 CONCLUSION

In this section of the thesis, it will be presented an overview of the summarized results, how they address the research question, and their alignment with the research objective. It will be, also, discussed their contribution to both theoretical and practical implications. Additionally, there is a reflection on acknowledged study limitations, implications, and provided recommendations for future improvements and research.

This study aims to foster the country HE dropout research, and raise the awareness of HE churn implications to the countries development, the effective utilization of HE resources, as well the individual consequences, by providing the first ever HE dropout reasons and estimation on the sample of around 20 percent of tertiary education students population and proposing the ML approach to predict the risk of withdrawal at earliest stage of HE. The results indicate that almost half of students at UNIBL churned, for many reasons. Having in mind the broad consequences of high churn rate and its increase in HE for the country, university, and individual, the size and implications of it highlighted the urgent need for policymakers to address this issue. This urgency is further underscored by the economic indicators of Bosnia and Herzegovina, brain drain and the relatively low proportion of highly educated individuals in the national workforce.

The results suggest that is possible to predict the students churn at earliest stage of education, even on challenging dataset, with missing data, on dataset that is poor in academic and socioeconomics features. By employing the ML models, this research succeed to achieve the high performances and quality, and quite good to overcome the immanence of the black box models by introducing pre-hoc and post-hoc interpretability and explainability techniques.

The contribution of the research in terms of practice, and theory, together in context of Bosnia and Hercegovina are follow:

- The first comprehensive and longitudinal estimation of higher education attrition at the University of Banja Luka, based on a sample of 37 thousands undergraduate students.
- The first survey of reasons for dropout at UNBIL, and students' trajectories in the years following their departure.
- Refined segmentation of churned students into different categories to achieve clarity and comparability of dropout in higher education.

- Training ML models on challenging dataset that is characterized by missing values, binary variables, and poor in socioeconomics, secondary education and academics features of student.
- Novel comparison of Histogram-based Gradient Boosting Classifier (HGBC) with other ML models that are well used by other researchers in the Educational Data Mining field, which offering new insights into models relative performance in three stages of prediction.
- Results confirming existing studies in the field that adding more variables over time improves the model's performance.
- Comparison of the feature importance among variety of trained ML models, by combining the PI and SHAP importance together with pre-hoc technique – correlation, enhancing confidence in the models' decision of classified students.
- Discovered the most important variables regarding the churn at UNIBL in global and individual level.

This thesis provide ML model for early churn detection, referred to as an Early Warning System (EWS) at UNIBL, in three stages: pre-enrollment stage, in enrollment stage (enrollment week), and at the end of first university year. Moving the prediction to the earliest stage in HE ensures timely prediction and prevention, especially considering that half of the overall dropout occurs during the first year of study.

Implementing this EWS, the University of Banja Luka can reduce the waste of resources and growing costs of higher education student churn, which has multifaced consequences on the country's development, economy, society, individuals, as well as higher education institutions.

In terms of contributions to the field and addressing gaps – prior to this research, there was a lack of statistical data, estimations, reports, or any other data related to HE dropout in the country. This study provides a detailed and comprehensive estimation of dropout among approximately 20 percent of undergraduate students over a period of 12 years.

There was also a lack of longitudinal studies of dropout and studies addressing the HE attrition in its earliest stage. This research fulfills this gap by using data from the academic years 2007/08 to 2018/19, providing insights into students' attrition in three stages of time: pre-enrollment, enrollment week and end of first year.

In our literature review, we noticed a lack of large datasets utilized in predicting student churn in higher education. Our dataset for machine learning models consists of 20

thousands students, while the dataset for churn estimation consists of 37 thousands students.

Prior to our study, the Histogram-based Gradient Boosting Classifier had not been used for churn prediction in higher education. We introduced this ensemble black box model, along with commonly used RF, SVM, and NN models, to compare their performance on a challenging dataset.

Our literature review revealed only nine papers that incorporated explainable AI into the field of Educational Data Mining, while seven of them are in higher education. We aim to contribute by providing insights into model outputs through reporting and comparing feature importance using Permutation Importance and SHAP values at three stages of prediction. This allows us to track changes in variable importance over time and identify the most important variables at each stage of prediction.

Our study findings are in alignment with existing literature. In comparing our study results with others in the field, the attrition rate at the University of Banja Luka (UNIBL) is comparable only to the methodology used in Denmark, which reported an overall churn rate in higher education of 30 percent. At UNIBL, the churn rate was 47.1 percent, with recent generations experiencing higher churn rates. When compared to other institutions in Europe, the churn rates at UNIBL by faculties are generally higher. Female students show persistence, and there is a correlation between the location of the university and churn rates, as well as the level of municipality development, which is consistent with findings from other research.

Our data also suggest differences between dropouts and non-dropouts in terms of their age, academic performance, and study field, which is consistent with existing literature. When considering the reasons for churn, some align with existing studies, showing that students drop out due to institutional, personal, and external reasons such as financial constraints and working while studying. This highlights the value of machine learning, as it uncovers hidden aspects that are consistent with both qualitative and quantitative approaches. However, a top reason for churn at UNIBL was disputes or conflicts with lecturers or professors, which has not been reported in the literature before. The identified institutional reasons for churn account for around one third of total dropouts, while this share in Europe is 13 percent.

The HGBC was used for the first time in an EDM context, on a challenging dataset. This means there isn't much information to compare with other research. However, in previous research, the highest accuracy was achieved by DT algorithms, RF, SVM and NN. We

achieved a prediction accuracy of 82.5 percent using HGBC with the top 13 features at the end of the school year, while at the beginning of the year it was 75.2 percent.

In modeling student dropout in three stages of time, even on a challenging dataset, HGBC showed much better results in an experiment with a changed churn definition and imbalanced data, when compared with existing research. We obtained a recall measure of 76 percent, correctly classifying 76 students as dropouts out of 100 who actually dropped out, while the highest achiever in another study with a rich dataset was 37 percent recall. HGBC demonstrated its power compared with RF, SVM, and NNs, even when the definition of churn was changed and when variables with missing data were introduced and later removed in three stages of prediction on a dataset with few variables but a large size.

We found that the strongest predictors of churn in higher education are age, scholarship status, female gender, distance from the university, and ECTS score, which aligns with existing literature. However, the strongest predictors that are different from other researchers are students' cohort (i.e. the academic year of enrollment), duration of the study program, and an artificial ID variable that carries study program, admission score rank and faculty information. Other studies found that financial features, ethnicity, and high school GPA, together with admission test scores, are important predictors of churn. Due to the lack of these variables and more than 80 percent missing data for high school GPA and admission scores, we were unable to check their importance using all models, except for HGBC, which can handle missing values. Among the 13 most important variables, the HGBC reported that admission score and high school GPA were among them, but ranked last, while the number of passed exams was ranked 4th (PI) and 6th (SHAP), higher than those variables.

The importance of some variables fluctuated over time, with some initially decreasing and then increasing again, while others lost their significance as time passed. The results of this study can be applied by practitioners in the following ways: a) Serve as the foundation for creating a dropout prevention policy at the country level or at UNIBL; b) Implement the proposed or modified Early Warning System at UNIBL or other universities; c) Raise awareness about the urgent need for action and information regarding higher education dropout in the entire country by establishing annual reports and statistics on churn; d) Explore the potential for automating annual dropout reports using the Python code developed in this thesis; e) Establish support programs, groups, and mentors, organize socialization events, provide more information about study

programs and requirements to high school students, align secondary and tertiary course syllabuses, offer more internships and summer school opportunities, and pinpoint the specific weeks in semesters when churn occurs; f) Improve the University's database management to better address dropout issues by implementing recommendations related to missing data and variables associated with dropout to accurately identify the weeks when churn occurs.

This research is affected by sampling issues, limited data access, and lack of research experience. The sample size in the qualitative part of the research is not large enough to draw generalized conclusions, which has implications for the chosen dropout definition and may result in an underestimation of dropout rates. The limited data access is evident in the lack of socio-economic, academic, and pre-academic variables, as well as missing data. Other publicly funded universities either did not provide or refused to provide student churn data in time for inclusion in this thesis. The lack of research experience in the field of Educational Data Mining (EDM) is reflected in the time spent understanding the university dataset, higher education law, rules, practices, dataset variables, as well as the order of data preprocessing and evaluation steps in Python.

In terms of recommendations to other researchers for improving this research, it is suggested to introduce a model at the end of the first semester with in-semester data, which may offer insights for preventing attrition in the second semester of the first university year. Since half of the churn occurred in the first year, prevention efforts should be primarily focused there. Zooming in to the faculty level with exam records could also enhance the model. Implementing the Responsible AI library in Python on the current dataset may reveal which parts provide the highest accuracy and why, as well as the importance of features. To make the model's results more actionable for students, it is essential to incorporate additional visualization techniques, such as Partial Dependence Plots, Anchors, Counterfactuals, and Deletion Diagnostics.

Additionally, a systematic approach to a nationwide survey may identify factors that are feasible to address in the short and long term.

One interesting finding that emerged was not investigated further because it did not provide answers to the research questions posed in this study. The finding suggested that the importance of a student's cohort, whether they enrolled in a recent or older academic year within the 2007-2018 span, showed that students in recent cohorts churn more. One recommendation for other researchers in the field is to build upon these results by employing unsupervised learning models, which may detect important patterns in

students' cohorts over time. Understanding what drives attrition can enable more efficient prevention efforts in the future.

Among the challenges facing the higher education institutions of Bosnia and Herzegovina is the waste of resources and growing costs of higher education student churn, which has multi-dimensional consequences on the country's development, economy, society, individuals, as well as higher education institutions. This thesis represents a pioneering effort in providing a comprehensive estimation of higher education categories and share of dropout, covering one-fifth of the student population in Bosnia and Herzegovina over a 12-year period. It offers valuable insights into the reasons for and consequences of withdrawal from the University of Banja Luka. By employing an existing yet novel approach in the field of Educational Data Mining, this study demonstrates that it is possible to effectively challenge widely used machine learning models for early-stage dropout prediction, even with modest datasets and missing data. Furthermore, the integration of multiple pre-hoc and post-hoc explainable AI techniques enhances the reliability and interpretability of used black box models, offering a clearer understanding why and how the models decide.

This study opens the door to numerous inquiries aimed at estimating the size, causes, and structure of study interruptions not only within the country but also across the broader region, where there is a significant lack of data on higher education dropout. The objective of this research extends beyond the application of machine learning models for the early detection of at-risk students; it seeks to inform concrete, systematic actions to mitigate this phenomenon, ultimately contributing to the reduction of brain drain — a challenge in which the country has been among the most affected in the world in recent years.

9 REFERENCES:

Abe, Shigeo (2010): *Support Vector Machines for Pattern Classification*, Springer Science & Business Media.

Agency for Statistics of Bosnia and Herzegovina (2021): Bosnia and Herzegovina in numbers, 2021, Agency for Statistics of Bosnia and Herzegovina.

Agnihotri, L.; Ott, Alexander (2014): Building a Student At-Risk Model: An End-to-End Perspective From User to Data Scientist, in: *PROCEEDINGS OF THE SEVENTH INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING*, London, United Kingdom, S. 209–2012.

Aguiar, Everaldo (2015): Identifying Students at Risk and Beyond: A Machine Learning Approach, (Dissertation) South Bend, Indiana, USA: University Of Notre Dame, doi: 10.7274/1v53jw8435h.

Ahmad, Zahoor; Shahzadi, Erum (2018): Prediction of Students' Academic Performance Using Artificial Neural Network, in: *Bulletin of Education and Research*, Institute of Education and Research, Jg. 40, Nr. 3, S. 157–164.

Alcauter, Iara; Martinez-Villaseñor, Lourdes; Ponce, Hiram (2023): Explaining Factors of Student Attrition at Higher Education, in: *Computación y Sistemas*, Jg. 27, Nr. 4, doi: 10.13053/cys-27-4-4776.

Alkhasawneh, Ruba (2011): Developing a Hybrid Model to Predict Student First Year Retention and Academic Success in STEM Disciplines Using Neural Networks, Virginia Commonwealth University.

Ameri, Sattar (2015): Survival Analysis Approach For Early Prediction Of Student Dropout, (Master Theses) Detroit, Michigan: Wayne State University.

Ameri, Sattar (2016): Survival Analysis Approach For Early Prediction Of Student Dropout,.

Appiah, Elizabeth N. (2017): The Effect of Education Expenditure on Per Capita GDP in Developing Countries, in: *International Journal of Economics and Finance*, Jg. 9, Nr. 10, S. 136, doi: 10.5539/ijef.v9n10p136.

Arce, Maria Elena; Crespo, Barbara; Míguez-Álvarez, Carla (2015): Higher Education Drop-Out in Spain—Particular Case of Universities in Galicia, in: *International Education Studies*, Jg. 8, Nr. 5, S. p247, doi: 10.5539/ies.v8n5p247.

Arnold, Kimberly E.; Pistilli, Matthew D. (2012): Course signals at Purdue: using learning analytics to increase student success, in: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, New York, NY, USA: Association for Computing Machinery (LAK '12), S. 267–270, doi: 10.1145/2330601.2330666.

Aulck, Lovenoor; Velagapudi, Nishant; Blumenstock, Joshua; u. a. (2016): Predicting Student Dropout in Higher Education, in: *arXiv:1606.06364 [cs, stat]*,.

Auria, Laura; Moro, R. A. (2008): Support Vector Machines (SVM) as a Technique for Solvency Analysis, in: *SSRN Electronic Journal*, doi: 10.2139/ssrn.1424949.

Baghernejad, Danielle (2016): ANALYSIS OF MTSU STUDENT RETENTION DATA, Middle Tennessee State University.

Bain, Alexander (1873): *Mind and body. The theories of their relation*, New York : D. Appleton and company.

Baranyi, Máté; Nagy, Marcell; Molontay, Roland (2020): Interpretable Deep Learning for University Dropout Prediction, in: *Proceedings of the 21st Annual Conference on Information Technology Education*, New York, NY, USA: Association for Computing Machinery (SIGITE '20), S. 13–19, doi: 10.1145/3368308.3415382.

Barros, Thiago M.; Souza Neto, Plácido A.; Silva, Ivanovitch; u. a. (2019): Predictive Models for Imbalanced Data: A School Dropout Perspective, in: *Education Sciences*, Multidisciplinary Digital Publishing Institute, Jg. 9, Nr. 4, S. 275, doi: 10.3390/educsci9040275.

Bean, John P.; Metzner, Barbara S. (1985): A Conceptual Model of Nontraditional Undergraduate Student Attrition, in: *Review of Educational Research*, Jg. 55, Nr. 4, S. 485–540, doi: 10.3102/00346543055004485.

Behr, Andreas; Giese, Marco; Teguim Kamdjou, Herve D.; u. a. (2020): Dropping out of university: a literature review, in: *Review of Education*, Jg. 8, Nr. 2, S. 614–652, doi: 10.1002/rev3.3202.

Belle, Vaishak; Papantonis, Ioannis (2021): Principles and Practice of Explainable Machine Learning, in: *Frontiers in Big Data*, Frontiers, Jg. 4, doi: 10.3389/fdata.2021.688969.

Bernardo, Ana; Cervero, Antonio; Esteban, María; u. a. (2017): Freshmen Program Withdrawal: Types and Recommendations, in: *Frontiers in Psychology*, Jg. 8.

Bianca Thaler; Martin Unger (2014): *Dropouts ≠ Dropouts Wege nach dem Abgang von der Universität*, Vienna: Institute for Advanced Studies, Vienna.

Blagojević, Marija; Micić, Živadin (2013): A web-based intelligent report e-learning system using data mining techniques, in: *Computers & Electrical Engineering*, Jg. 39, Nr. 2, S. 465–474, doi: 10.1016/j.compeleceng.2012.09.011.

Boaz Shmueli (2019): Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of, [online] https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a [26.06.2022].

Breiman, Leo; (2001): Random forest, [online] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro [26.04.2019].

Brownlee, Jason (2021): *Imbalanced Classification with Python, Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning*, v1.3.

Cabrera, Alberto F.; Nora, Amaury; Castaneda, Maria B. (1993): College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention, in: *The Journal of Higher Education*, Jg. 64, Nr. 2, S. 123–139, doi: 10.2307/2960026.

Caruana, Rich; Niculescu-Mizil, Alexandru (2006): An empirical comparison of supervised learning algorithms, in: *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, S. 161–168, doi: 10.1145/1143844.1143865.

CBS (o. J.): Figures - Society | Trends in the Netherlands 2019 - CBS, *Trends in the Netherlands 2019*, [online] https://longreads.cbs.nl/trends19-eng/society/figures [02.06.2022].

Chai, Kevin E. K.; Gibson, David (2015): Predicting the Risk of Attrition for Undergraduate Students with Time Based Modelling, in: *International Association for Development of the Information Society*, Dublin, Ireland: International Association for the Development of the Information Society, S. 109–116.

Cortes, Corinna; Vapnik, Vladimir (1995): Support-Vector Networks, in: *Machine Learning*, Jg. 20, Nr. 3, S. 273–297, doi: 10.1023/A:1022627411411.

Criminisi, A; Shotton, J; Konukoglu, E (2011): Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning, in: *Microsoft Research technical report TR-2011-114*, S. 151.

Dass, Sheran; Gary, Kevin; Cunningham, James (2021): Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model, in: *Information*, Multidisciplinary Digital Publishing Institute, Jg. 12, Nr. 11, S. 476, doi: 10.3390/info12110476.

Dech Thammasiri; Dursun Delen; Phayung Meesad; u. a. (2013): A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, in: *Expert Systems with Applications*, S. 3, doi: https://doi.org/10.1016/j.eswa.2013.07.046.

Delen, Dursun (2010): A comparative analysis of machine learning techniques for student retention management, in: *Decision Support Systems*, Jg. 49, Nr. 4, S. 498–506, doi: 10.1016/j.dss.2010.06.003.

Delen, Dursun (2011): Predicting Student Attrition with Data Mining Methods, in: *Journal of College Student Retention: Research, Theory & Practice*, Jg. 13, Nr. 1, S. 17–35, doi: 10.2190/CS.13.1.b.

Delen, Dursun; Davazdahemami, Behrooz; Rasouli Dezfouli, Elham (2023): Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework, in: *Information Systems Frontiers*, Jg. 26, Nr. 2, S. 641–662, doi: 10.1007/s10796-023-10397-3.

Elena Torou; Suzanne Borg; Tatjana Chircop (2022): Dropping Out From Post-Secondary Vocational Education: A Case Study in Malta, in: *MCAST Journal of Applied Research & Practice*, Jg. 6, Nr. 1, S. 208–231, doi: 10.5604/01.3001.0015.8195.

ESLU (2017): A Study focusing on Students dropping out from Post-Secondary Education in Malta Scholastic Year 2015-16, Ministry for Education and Employment in Malta.

European Commission. Directorate General for Education and Culture.; CHEPS.; NIFU. (2015): *Dropout and completion in higher education in Europe: annex 2: short country reports.*, LU: Publications Office.

Febro, January D.; Barbosa, Jocelyn (2017): Mining student at risk in higher education using predictive models, in: *Journal of Advances in Technology and Engineering Research*, doi: https://doi.org/10.20474/jater-3.4.2.

Gandhi, Rohith (2018): Support Vector Machine — Introduction to Machine Learning Algorithms, *Towards Data Science*, [online] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [21.04.2019].

Gašpar, Dražena; Mabić, Mirela; Ćorić, Ivica (2015): *Data Mining and Predicting Student Performance*,.

Ghorbani, Ramin; Ghousi, Rouzbeh (2020): Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques, in: *IEEE Access*, Jg. 8, S. 67899–67911, doi: 10.1109/ACCESS.2020.2986809.

Guryanov, Aleksei (2019): *Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees*, Cham: Springer International Publishing (Lecture Notes in Computer Science. 11832), doi: 10.1007/978-3-030-37334-4_4.

Gutiérrez, Francisco; Seipp, Karsten; Ochoa, Xavier; u. a. (2020): LADA: A learning analytics dashboard for academic advising, in: *Computers in Human Behavior*, Jg. 107, S. 105826, doi: 10.1016/j.chb.2018.12.004.

Guzmán, Alfredo; Barragán, Sandra; Cala Vitery, Favio (2021): Dropout in Rural Higher Education: A Systematic Review, in: *Frontiers in Education*, Jg. 6, doi: doi.org/10.3389/feduc.2021.727833.

Hasan, Md. Rajib; Siraj, Fadzilah; Sainin, Mohd Shamrie (2015): Improving ensemble decision tree performance using Adaboost and Bagging, in: *AIP Conference Proceedings*, Kedah, Malaysia, S. 030008, doi: 10.1063/1.4937027.

HEA (2021): Completion Data Release March 2021, *Higher Education Authority*, [online] https://hea.ie/statistics/data-for-download-and-visualisations/students/completion/completion-data-release-march2021/ [02.06.2022].

Herţeliu, Claudiu; Alexe-Coteţ, Daniela; Hâj, Cezar Mihai; u. a. (2022): Defining and Measuring Dropout Phenomenon in Romanian Public Universities, in: Adrian Curaj, Jamil Salmi, und Cezar Mihai Hâj (Hrsg.), *Higher Education in Romania: Overcoming Challenges and Embracing Opportunities*, Cham: Springer International Publishing, S. 93–118, doi: 10.1007/978-3-030-94496-4_6.

Heublein, Ulrich (2014): Student Drop-out from German Higher Education Institutions, in: *European Journal of Education*, Jg. 49, doi: 10.1111/ejed.12097.

Hiltunen, Heli (2018): Statistics Finland - Discontinuation of education 2018, Statistics Finland, [online] https://www.stat.fi/til/kkesk/2018/kkesk_2018_2020-03-12_tie_001_en.html [02.06.2022].

IBM (2014): IBM Knowledge Center, [online] https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview [07.05.2019].

Isiaka, R. M.; Babatunde, R. S.; Ajao, F. J.; et al. (2019): A Machine Learning Approach to Dropout Early Warning System Modeling, in: *International Journal of Advanced Studies in Computers, Science and Engineering; Gothenburg*, Jg. 8, Nr. 2, S. 1–12.

Išljamović, Sonja (2013): EARLY PREDICTION OF UNIVERSITY STUDENTS' SUCCESS VIA NEURAL NETWORKS - ProQuest, in: *Metalurgia International, Bucharest*, Jg. 18, Nr. 5, S. 120–125.

Jadrić, Mario; Garača, Željko; Čukušić, Maja (2010): Student Dropout Analysis with Application of Data Mining Methods, in: *Management : journal of contemporary management issues*, Jg. 15, Nr. 1, S. 31–46.

James, Gareth; Witten, Daniela; Hastie, Trevor; et al. (2013): *An Introduction to Statistical Learning*, New York, NY: Springer New York (Springer Texts in Statistics), doi: 10.1007/978-1-4614-7138-7.

Jovanovic, Milos; Vukicevic, Milan; Milovanovic, Milos; et al. (2012): Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study, in: *International Journal of Computational Intelligence Systems*, Atlantis Press, Jg. 5, Nr. 3, S. 597–610, doi: 10.1080/18756891.2012.696923.

Kabashi, Qamil; Shabani, Isak; Caka, Nebi (2022): Analysis of the Student Dropout Rate at the Faculty of Electrical and Computer Engineering of the University of Prishtina, Kosovo, From 2001 to 2015, in: *IEEE Access*, Jg. 10, S. 68126–68137, doi: 10.1109/ACCESS.2022.3185620.

Kacapor, Kemal; Lagumdžija, Z (2020): *Rudarenje edukacijskih podataka: korištenje klasteriranja za predikciju studentskog uspjeha*,.

Katsuragi, Miki; Tanaka, Kenji (2022): Dropout Prediction by Interpretable Machine Learning Model Towards Preventing Student Dropout, in: *Transdisciplinarity and the Future of Engineering*, IOS Press, S. 678–683, doi: 10.3233/ATDE220700.

Ke, Guolin; Meng, Qi; Finley, Thomas; et al. (2017): LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: *Advances in Neural Information Processing Systems*, Curran Associates, Inc.

Kehm, Barbara M.; Larsen, Malene Rode; Sommersel, Hanna Bjørnøy (2019): Student dropout from universities in Europe: A review of empirical literature, in: *Hungarian Educational Research Journal*, Akadémiai Kiadó, Jg. 9, Nr. 2, S. 147–164, doi: 10.1556/063.9.2019.1.18.

Khawar Shakeel; Naveed Anwer Butt (2015): Educational Data Mining to Reduce Student Dropout Rate by Using Classification, in: *ResearchGate*, Lexington, Kentucky, USA, S. 8.

Kim, Sangyun; Choi, Euteum; Jun, Yong-Kee; et al. (2023): Student Dropout Prediction for University with High Precision and Recall, in: *Applied Sciences*, Multidisciplinary Digital Publishing Institute, Jg. 13, Nr. 10, S. 6275, doi: 10.3390/app13106275.

Klaus Schwab, World Economic Forum (2019): *The Global Competitiveness Report 2019*, (Insight Report) Switzerland.

Kovač, Romano; Oreški, Dijana (2018): Educational Data Driven Decision Making: Early Identification of Students at Risk by Means of Machine Learning, in: *Central European Conference on Information and Intelligent Systems*, Varazdin, Croatia: Faculty of Organization and Informatics Varazdin, S. 231–237.

Krumm, Andrew E.; Waddington, Richard Joseph; Teasley, Stephanie D.; et al. (2014): A Learning Management System-Based Early Warning System for Academic Advising in Undergraduate Engineering, in: Springer New York, doi: 10.1007/978-1-4614-3305-7_6.

Kulkarni, Ajay; Chong, Deri; Batarseh, Feras A. (2020): 5 - Foundations of data imbalance and solutions for a data democracy, in: Feras A. Batarseh und Ruixin Yang (Hrsg.), *Data Democracy*, Academic Press, S. 83–106, doi: 10.1016/B978-0-12-818366-3.00005-8.

Lakkaraju, Himabindu; Aguiar, Everaldo; Shan, Carl; et al. (2015): A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, Sydney, NSW, Australia: ACM Press, S. 1909–1918, doi: 10.1145/2783258.2788620.

Lee, Sunbok; Chung, Jae Young (2019): The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction, in: *Applied Sciences*, Multidisciplinary Digital Publishing Institute, Jg. 9, Nr. 15, S. 3093, doi: 10.3390/app9153093.

Li, Hui; Sun, Jie (2012): Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples – Evidence from the Chinese hotel industry, in: *Tourism Management*, Elsevier, Jg. 33, Nr. 3, S. 622–634, doi: 10.1016/j.tourman.2011.07.004.

Licskay, Péter (2021): One-third of students never finish higher education in Hungary?, *Daily News Hungary*, [online] https://dailynewshungary.com/one-third-of-students-never-finish-higher-education-in-hungary/ [02.06.2022].

Lin, Jien-Jou (2013): Student Success: Approaches to Modeling Student Matriculation and Retention, Purdue University.

Livingstone, D J; Manallack, D T; Tetko, I V (1997): Data modelling with neural networks: Advantages and limitations, in: *Springer*, S. 8.

López-Zambrano, Javier; Lara Torralbo, Juan Alfonso; Romero, Cristóbal (2021): Early prediction of student learning performance through data mining: A systematic review, in: *Psicothema*, Spain: Colegio Oficial de Psicólogos del Principado de Asturias, Jg. 33, Nr. 3, S. 456–465.

Lundberg, Scott; Lee, Su-In (2017): A Unified Approach to Interpreting Model Predictions, arXiv, doi: 10.48550/arXiv.1705.07874.

Márquez-Vera, Carlos; Cano, Alberto; Romero, Cristóbal; et al. (2013): Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data, in: *Applied Intelligence*, Jg. 38, Nr. 3, S. 315–330, doi: 10.1007/s10489-012-0374-8.

Michal Rajski; Alexandra Davis; Gabriel Goodspeed; Dalton Goree; Theresa Haunold; Nate Johnson; (2018): The Problem with French Universities | EUChicago, *University of Chicago*, [online] https://voices.uchicago.edu/euchicago/the-problem-with-french-universities/ [02.06.2022].

Michelle Roberts (2019): Universities With Highest and Lowest Dropout Rates, *Universities With Highest and Lowest Dropout Rates*, [online] https://www.whatuni.com/advice/news/universities-with-highest-and-lowest-dropout-rates/85809/ [02.06.2022].

Mijović, Vladimir; Isaković, Jasna; Husković, Selma; et al. (2019): Labour Force Survey 2019, Thematic Bulletin, Agency for Statistics of Bosnia and Herzegovina.

Modena, Francesca; Rettore, Enrico; Tanzi, Giulia Martina (2020): The Effect of Grants on University Drop-Out Rates: Evidence on the Italian Case, in: S. 28.

Mustafa, Mohammad Nurul; Chowdhury, Linkon; Kamal, Md.Sarwar (2012): Students dropout prediction for intelligent system from tertiary level in developing country, in: *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, Bangladesh: IEEE, S. 113–118, doi: 10.1109/ICIEV.2012.6317441.

Nagy, Marcell; Molontay, Roland (2023): Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention, in: *International Journal of Artificial Intelligence in Education*, doi: 10.1007/s40593-023-00331-8.

Natthakan Iam-On; Tossapon Boongoen (2017): Generating descriptive model for student dropout: a review of clustering approach, in: *Springer Berlin Heidelberg*, doi: 10.1186/s13673-016-0083-0.

Natthakan Iam-On; Tossapon Boongoen (2015): Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings, in: *Springer-Verlag Berlin Heidelberg*, doi: 10.1007/s13042-015-0341-x.

Natthakan Lam-On; Tossapon Boongoen (2014): Using cluster ensemble to improve classification of student dropout in Thai university - IEEE Conference Publication, in: Kitakyushu, Japan: IEEE, doi: 10.1109/SCIS-ISIS.2014.7044875.

Nielsen, Michael A. (2015): *Neural Networks and Deep Learning*,.

OECD (2022): *Education at a Glance 2022: OECD Indicators*, Paris: Organisation for Economic Co-operation and Development.

OECD (2009): How many students drop out of tertiary education? in Highlights from Education at a Glance 2008, OECD Publishing, Paris.

Olja Jovanović; Ljiljana Plazinić; Jelena Joksimović; Jovan Komlenac; Ana Pešikan; (2017): Developing the Early Warning System for identification of students at risk of dropping out using a collaborative action research process, in: *Psihološka istraživanja*, Филозофски факултет, Универзитет у Београду, Jg. 20, Nr. 1, S. 107–125.

Osmanbegovi, Edin; Agi, Haris; Suljic, Mirza (2013): Data Mining Approach for Making Prediction of Students Success, in: Atlantis Press, S. 705–709, doi: 10.2991/icaicte.2013.145.

Osmanbegovic, Edin; Suljic, Mirza (2012): Data Mining Approach For Predicting Student Performance, in: *Economic Review: Journal of Economics and Business*, University of Tuzla, Faculty of Economics, Jg. 10, Nr. 1, S. 3–12.

Osmanbegović, Edin; Suljić, Mirza; Agić, Hariz (2014): DETERMINING DOMINANT FACTOR FOR STUDENTS PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS, in: *Tranzicija*, Ekonomski institut Tuzla, JCEA Zagreb, DAEB, IEP Beograda, feam Bukurest, Jg. 16, Nr. 34, S. 147–158.

Pal, Saurabh (2012): Mining Educational Data Using Classification to Decrease Dropout Rate of Students, in: *International journal of multidisciplinary sciences and engineering*, Jg. 3, Nr. 5, S. 35–39, doi: 10.48550/arXiv.1206.3078.

Pastor, José M.; Peraita, Carlos; Serrano, Lorenzo; et al. (2018): Higher education institutions, economic growth and GDP per capita in European Union countries, in: *European Planning Studies*, Routledge, Jg. 26, Nr. 8, S. 1616–1637, doi: 10.1080/09654313.2018.1480707.

Patacsil, Frederick F. (2020): Survival Analysis Approach for Early Prediction of Student Dropout Using Enrollment Student Data and Ensemble Models, in: *Universal Journal of Educational Research*, Horizon Research Publishing, Jg. 8, Nr. 9, S. 4036–4047, doi: 10.13189/ujer.2020.080929.

Paura, Liga; Arhipova, Irina (2016): STUDENT DROPOUT RATE IN ENGINEERING EDUCATION STUDY PROGRAM, in: *ENGINEERING FOR RURAL DEVELOPMENT*, S. 6.

Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; et al. (2011): Scikit-learn: Machine Learning in Python, in: *Journal of Machine Learning Research*, Jg. 12, Nr. 85, S. 2825–2830.

Perez, Boris; Castellanos, Camilo; Correal, Dario (2018): Applying Data Mining Techniques to Predict Student Dropout: A Case Study, in: *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, S. 1–6, doi: 10.1109/ColCACI.2018.8484847.

Pérez, Petra Norma Maya; Jorge R. Aguilar C; R, Rosa A. Zamora; et al. (2019): Predictive Model Design applying Data Mining to identify causes of Dropout in University Students, in: *Strategy, Technology & Society*, Jg. 7, Nr. 2.

Plak, Simone; Cornelisz, Ilja; Meeter, Martijn; et al. (2022): Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment, in: *Higher Education Quarterly*, Jg. 76, Nr. 1, S. 131–152, doi: 10.1111/hequ.12298.

Qian, Yue (2017): Gender Asymmetry in Educational and Income Assortative Marriage, in: *Journal of Marriage and Family*, Jg. 79, Nr. 2, S. 318–336, doi: 10.1111/jomf.12372.

Rodríguez, Patricio; Villanueva, Alexis; Dombrovskaia, Lioubov; et al. (2023): A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of Chile, in: *Education and Information Technologies*, Jg. 28, Nr. 8, S. 10103–10149, doi: 10.1007/s10639-022-11515-5.

Rodríguez-Maya, Noel Enrique; Lara-Álvarez, Carlos; May-Tzuc, Oscar; et al. (2017): Modeling Students' Dropout in Mexican Universities, in: *Research in Computing Science*, Jg. 139, Nr. 1, S. 163–175, doi: 10.13053/rcs-139-1-13.

Schnepf, Sylke V. (2014): Do Tertiary Dropout Students Really Not Succeed in European Labour Markets?, in: *SSRN Electronic Journal*, doi: 10.2139/ssrn.2409537.

Shapley, L. S. (1953): 17. A Value for n-Person Games, in: *17. A Value for n-Person Games*, Princeton University Press, S. 307–318, doi: 10.1515/9781400881970-018.

Sharma, Deepak Kumar; Chatterjee, Mayukh; Kaur, Gurmehak; et al. (2022): 3 - Deep learning applications for disease diagnosis, in: Deepak Gupta, Utku Kose, Ashish Khanna, et al. (Hrsg.), *Deep Learning for Medical Applications with Unique Data*, Academic Press, S. 31–51, doi: 10.1016/B978-0-12-824145-5.00005-8.

Simeunovic, Vlado; Milic, Sanja (2018): Application of Data Mining in Predicting the Evaluation Preocess at Postsecondary Educational Establishments - RapidMiner, in: *Novi pristupi metodologiji istraživanja odgoja*, S. 79–109.

Simeunovic, Vlado; Preradović, Ljubiša (2014): Using Data Mining to Predict Success in Studying, in: Jg. 16, S. 491–523.

Siri, Anna (2015): Predicting Students' Dropout at University Using Artificial Neural Networks, in: *Italian Journal of Sociology of Education*, Jg. 7, Nr. 2, S. 225–247.

Sisovic, Sabina; Matetic, Maja; Bakaric, Marija Brkic (2016): Clustering of imbalanced moodle data for early alert of student failure, in: *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, S. 165–170, doi: 10.1109/SAMI.2016.7423001.

Sofia Berlin Kolm; Fredrik Svensson (2017): Report 2017:17 E (summary of Swedish full report 2017:17) Early higher education drop-out rates in SwedenAnalyses of throughput rates of ten large study programmes, Swedish higher education authority.

Sokolova, Marina; Japkowicz, Nathalie; Szpakowicz, Stan (2006): Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, *ResearchGate*, [online] https://www.researchgate.net/publication/225215404_Beyond_Accuracy_F-Score_and_ROC_A_Family_of_Discriminant_Measures_for_Performance_Evaluation [28.04.2019].

Solis, Martin; Moreira-Mora, Tania; Gonzalez, Roberto; et al. (2018): Perspectives to Predict Dropout in University Students with Machine Learning, in: *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, San Carlos, Costa Rica, S. 1–6, doi: 10.1109/IWOBI.2018.8464191.

Statista (2022): Belgium: annual average dropout rate 2021, *Statista*, [online] https://www.statista.com/statistics/536434/average-dropout-rate-per-year-in-belgium/ [02.06.2022].

Statistics Norway (2022): Completion rates of students in higher education, *SSB*, [online] https://www.ssb.no/en/utdanning/hoyere-utdanning/statistikk/gjennomforing-ved-universiteter-og-hogskoler [02.06.2022].

Stepanovic Ilic, Ivana; Tosković, Oliver; Krstic, Ksenija (2020): DROPOUT AT UNIVERSITY LEVEL IN SERBIA: ANALYSIS OF MEASUREMENT, RESEARCH FINDINGS, SERVICES AND PREVENTION MEASURES (ОСИПАЊЕ НА НИВОУ ВИСОКОГ ОБРАЗОВАЊА У СРБИЈИ: АНАЛИЗА МЕРЕЊА ОСИПАЊА, НАЛАЗА ИСТРАЖИВАЊА И МЕРА ПРЕВЕНЦИЈЕ), in: *Zbornik Instituta za pedagoska istrazivanja*, Jg. 52, S. 498–519, doi: 10.2298/ZIPI2002479S.

Stiburek, Šimon; Vlk, Ales; Švec, Václav (2017): Study of the success and dropout in the higher education policy in Europe and V4 countries, in: *Hungarian Educational Research Journal*, Jg. 1, S. 43–56, doi: 10.14413/herj.2017.01.04.

Stirrup, Jen (2017): What's wrong with CRISP-DM, and is there an alternative?, *Jen Stirrup*, [online] https://jenstirrup.com/2017/07/01/whats-wrong-with-crisp-dm-and-is-there-an-alternative/ [07.05.2019].

Tavares, O.; Sin, C.; Dias, D.; et al. (2018): DROP-OUT AND COMPLETION AMONG PORTUGUESE STUDENTS, in: *EDULEARN18 Proceedings*, Palma, Spain: IATED, S. 1886–1892, doi: 10.21125/edulearn.2018.0545.

Teodorescu, Sandra. (2018): On The Positive Correlation between Education and GDP in Romania, in: S. 936–941.

Thammasiri, Dech; Delen, Dursun; Meesad, Phayung; et al. (2014a): A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, in: *Expert Systems with Applications*, Jg. 41, Nr. 2, S. 321–330, doi: 10.1016/j.eswa.2013.07.046.

Thammasiri, Dech; Delen, Dursun; Meesad, Phayung; et al. (2014b): A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, in: *Expert Systems with Applications*, Jg. 41, Nr. 2, S. 321–330, doi: 10.1016/j.eswa.2013.07.046.

Tinto, Vincent (1987): *Leaving College: Rethinking the Causes and Cures of Student Attrition | Request PDF*, University of Chicago Press.

Troelsen, Rie; Laursen, Per F. (2014): Is Drop-out from University Dependent on National Culture and Policy? The Case of Denmark, in: *European Journal of Education*, Wiley, Jg. 49, Nr. 4, S. 484–496.

University of Banja Luka (2023): About the University, *UNIBL*, [online] https://www.unibl.org/en/university/about-the-university [26.04.2023].

Waheed, Hajra; Anas, Muhammad; Hassan, Saeed-Ul; et al. (2021): Balancing sequential data to predict students at-risk using adversarial networks, in: *Computers & Electrical Engineering*, Jg. 93, S. 107274, doi: 10.1016/j.compeleceng.2021.107274.

Wang, Zhuping; Zhu, Chenjing; Ying, Zelin; et al. (2018): Design and Implementation of Early Warning System Based on Educational Big Data, in: *2018 5th International Conference on Systems and Informatics (ICSAI)*, S. 549–553, doi: 10.1109/ICSAI.2018.8599357.

Weng Fu Mei (2010): Modelling IT student retention at Taiwanese higher education institutions - RMIT University, (PdD dissertation) Melburn, Australia: RMIT University.

Werbos, P.J. (1990): Backpropagation through time: what it does and how to do it, in: *Proceedings of the IEEE*, Jg. 78, Nr. 10, S. 1550–1560, doi: 10.1109/5.58337.

Williams, Nigel; Zander, Sebastian; Armitage, Grenville (2006): A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification, in: *ACM SIGCOMM Computer Communication Review*, Jg. 36, Nr. 5, S. 5, doi: 10.1145/1163593.1163596.

Wirth, Rüdiger; Hipp, Jochen (2000): Crisp-dm: towards a standard process modell for data mining, in.:

Wolter, Stefan C.; Diem, Andrea; Messer, Dolores (2014): Drop-outs from Swiss Universities: an empirical analysis of data on all students between 1975 and 2008, in: *European Journal of Education*, Wiley, Jg. 49, Nr. 4, S. 471–483.

World Bank (2023): World Bank national accounts data, and OECD National Accounts data files.,.

Wright, Mary C.; McKay, Timothy; Hershock, Chad; et al. (2014): Better Than Expected: Using Learning Analytics to Promote Student Success in Gateway Science, in: *Change: The Magazine of Higher Learning*, Routledge, Jg. 46, Nr. 1, S. 28–34, doi: 10.1080/00091383.2014.867209.

Xavier, Marlon; Meneses, Julio (2020): A Literature Review on the Definitions of Dropout in Online Higher Education, in: *EDEN Conference Proceedings*, Nr. 1, S. 73–80, doi: 10.38069/edenconf-2020-ac0004.

Yang, Shudong (2021): Who will dropout from university? Academic risk prediction based on interpretable machine learning, in.:

Zając, Tomasz Zbigniew; Komendant-Brodowska, Agata (2019): Premeditated, dismissed and disenchanted: higher education dropouts in Poland, in: *Tertiary Education and Management*, Jg. 25, Nr. 1, S. 1–16, doi: 10.1007/s11233-018-09010-z.

Zhang, Y.; Li, Y.; You, F.; et al. (2010): Withdrawal prediction using the blackboard learning management system through SOM, in: *The 2nd International Conference on Software Engineering and Data Mining*, S. 340–344.

APPENDIX A

Table A 1 - Domestic literature in EDM domain, Bosnia and Herzegovina

Source	Goal	Model	Other analysis used/implementation	Target group	Sample size	Sample features	the most accurate model
(Osmanbegovic und Suljic 2012)	Predict the success in a course	NB, NN, DT		1st year students Faculty of Economic, University of Tuzla	257 students, data collected from the 2010-2011, among first year students + data taken during the enrollment.	12 variables (gender, distance, earnings, GPA, scholarships, entrance exam, materials, family, internet access, time for studding, grade importance)	NB, 71.2-76.65%
(Osmanbegović et al. 2014)	Classification of performances of students	Rule based algorithm, DT, RF; Function; Bayes	WEKA	High school	1210 students. In year 2011/12 and 2012/13	19 variables	
(Osmanbegovi et al. 2013)	Predict student affiliation to the specific class	C4.5, RF, NB, MLP NN		Tuzla, High school			
(Simeunovic und Preradović 2014)	Predict success in a course	LR, DT, CART and NN	If-then	All students at Faculty of Economics, University of Bijeljina	354 students 2–4-year, survey	the importance of mark, attendance at tests, intellectual capabilities, scholarship, attendance at tutorials, and duration of studies	NN (76.4%), then follows LR (74.8%), DT (71.2%).
(Simeunovic und Milic 2018)	Predict success in studding	LR, NN, CART, DT	Use of Rapid Miner Software + WEKA	Faculty of Pedagogy, Bijeljina 2nd, 3rd, 4th year, University	at three different majors: the sample included 175 students.		CART - 57.97 NN - 55.07
(Kacapor und Lagumdžija 2020)	Detect 4 types of learners	kNN,	R, Python, SQL, MySQL, Rapid Miner	Faculty of Economics,		January 2015 and two semesters after. Moodle data. All courses.	

Source	Goal	Model	Other analysis used/implementation	Target group	Sample size	Sample features	the most accurate model
	related to success prediction			University of Sarajevo			
(Gašpar et al. 2015)	Predict success in a course and final grade	NN, NB, regression, and correlation analysis	association rules	Mostar, University			

Source: Literature review, author's contribution.

Table A 2 – Papers of prediction of student attrition by used ML models

Source: students drop out prediction models	Linear Discriminating	Hybrid model superviesed	N.N – superv.	Logistic Regression	Random Forest	Decision Tree	Naïve Bayes or Regression	kNN	Classification	Coxph	SVM	Unsupervised: SOM	Cluster, k-Means	Neural network	Hybrid model[18]	Other	Target group
(Alkhasawneh 2011)		x[19]	X													Focus groups	1st year, STEM, Virginia, USA
(Baghernejad 2016)				X	X												1st year, Middle Tennessee State University
(Khawar Shakeel und Naveed Anwer Butt 2015)					X	x	X										Byes is the most accurate
(Ameri 2016)				X		x			x	x	x					Survival models	All years of study Wayne State University
(Isiaka et al. 2019)	X			X		x	X	x	x		x						
(Zhang et al. 2010)												x					

[18] Hybrid models are the combinaton of the supervised and unsupervised learning alghorithms.
[19] Inputs were focus groups and genetic algorithm.

Source: students drop out prediction models	Linear Discriminating	Hybrid model superviesed	N.N – superv.	Logistic Regression	Random Forest	Decision Tree	Naïve Bayes or Regression	kNN	Classification	Coxph	SVM	Unsupervised: SOM	k-Cluster, Means	Neural network	Hybrid model[18]	Other	Target group
(Lin 2013)	X		X	X												Structural equation modeling[20]	
(Lakkaraju et al. 2015)				X	X	x					x					AdaBoost	
(Thammasiri et al. 2014b)			X	X		x					x					Imbalanced data	
(Jadrić et al. 2010)						x										WEKA	
(Natthakan Iam-On und Tossapon Boongoen 2015)			X			x	X	x	x							Back propagation	1st year
(Natthakan Iam-On und Tossapon Boongoen 2017), (Natthakan Lam-On und Tossapon Boongoen 2014)													x				
(Delen 2010)									x							Sensitivity analysis[21]	
(Delen 2011)			X	X		x										multi-layer perceptron (MLP) with a back-propagation	1st year
(Aulck et al. 2016)				X	X			x									

[20] Statistical method.
[21] Educational and financial variables are the best predictors.

Table A 3 - Survey structure conducted at UNIBL among students who left study by own request

General data	
Gender	Multiple choice. Offered: Male, Female.
Date of birth.	dd/mm/yyyy. Date question.
Place of birth.	Text question.
Country of birth.	Text question.
The year when you dropped out.	Multiple choice. Offered: 12 years (2007-2019) and answer "Do not remember."
The name of the first faculty where you dropped out.	Multiple choice, 17 faculties offered.
Reasons for churn	
The most important reason for dropout.	Multiple choice. Offered seven options: Moving abroad; Moving to another city in B&H; Inability to finance studies and the need for employment; I worked, and because of work, I did not manage to fulfill my obligations to study; Disagreement or conflict with the lecturer; Health-related reason; Pregnancy and marriage; and Other (to write in).
The second most important reason for dropout.	Multiple choice. We offered seven options (the same as the previous question) and Other (to write in).
The institutional and pedagogical reasons for dropout.	Checkbox. Offered four choices: Not applicable to me; My expectations were unmet since there was not enough internship; Professors' classes are boring; I lost motivation and interest to study during the school year; and Other (to write in).
Personal reasons for dropout.	Checkbox. Offered seven choices: Not applicable to me; At that moment, I was not ready for such kind of commitment; As I got familiar with the study program, I felt that this career path was not for me and that I would not do a job well; It was an exhausting study, mentally, for me; Pregnancy and marriage; It was difficult to study because my family was not close to me; I had very good revenues and I was not motivated to study and Other (to write in).
Financial reasons for dropout	Checkbox. Offered five choices: Not applicable; Parents could not afford to pay for my study, and I left university; I stayed without my scholarship; I could not afford to study anymore; I had to find a job to support my family; and Other (to write in).
After leaving your faculty, have you continued your education somewhere else?	Multiple choice. Yes. No.
Section for the ones who continue their study:	
Where do you start your study again?	Multiple choice. We offered: At another faculty, same University; At another public university in the country; At a private university in the country; At a university abroad; and Other (to write in).
Have you graduated?	Multiple choice. Offered: Yes; Still ongoing; Currently, I took a study break (passive); No, I left my studies. Other (to write in).
Section for the ones who did not continue their study:	
Why did you not start your study again, or somewhere else?	Text.
Section for all respondents:	
Are you satisfied with your decision to leave your study?	Multiple choice. We offered: Yes. No. I don't know. Other (to write in).
Are you currently employed?	Multiple choice. We offered: Yes. No. I am a student. I work part-time. Other (to write in).
Section for employed persons:	
What branch are you currently employed in?	Checkbox. Offered: 20 choices and Other (to write in).
Do you think your income would be higher now if you had finished your studies?	Multiple choice. We offered: No. Yes. Other (to write in).
Section for all respondents:	

If you are interested in the results of this study, please write your email.	

Source: Author's contribution.

Table A 4 - Exam records of Faculty of Economics and Faculty of Law, 2007-2018

Database content	Student's exam records
Time frame	2007-08 – 2018-19 school year
Format	.xls
Received	Via university email
Date of receiving	1st October 2019
Size	6651 KB
Database size	128.579 rows, 10 columns, 1.285.790 cells
Features	Total 10

Source: Author

Table A 5 - Variable description: Dataset of exam records of Faculty of Law and Faculty of Economics

Faculty:	Faculty name, text, 2 unique values.
indexs:	Identifier of the student, system generated, numeric, 8 digits. Representation YYYYBNNN (where the first 4 digits are the year of entry, the second digit represents Bachelor (1), Master (2) and Ph.D. (3), part-time student (5), and the last three digits are the number of students in the faculty in the year of entry).
altid:	Alternative ID of the student, string (numeric data with '/').
course_c:	Abbreviated name of the course (course code) with the name of the examination period and year. String. Unique values, 128,579.
course_n:	Full name of the course with the name of the exam period and year, text. Unique 128,579.
period_e:	One of seven predefined exam periods (January, April, June, September, first one in October, the second one in October, validated exam from another faculty, Exam in front of a committee). Text.
grade:	Numeric, scale 6-10. Also: 5-failed, 4-expelled, 3-did not appear for the examination, and 2-confirmed from previous evidence.
professor:	Name of the professor who administered the examination. Text.
score_type:	Type of score as part of the explanation of the score: 5-failed, 4-excluded, 3-did not appear for the exam, 2-confirmed from previous evidence, 6-10 - successfully passed. Text. Unique values, 8.
exam_date:	Date of taking the exam.

Source: Author

Table A 6 - Gender share, 2007-2018

Year of enroll	Male	Female	Female share (%)
2007	893	1497	62.6%
2008	912	1571	63.3%
2009	1351	1957	59.2%
2010	1294	2093	61.8%
2011	1187	1846	60.9%
2012	1330	2029	60.4%
2013	1467	2236	60.4%

Year of enroll	Male	Female	Female share (%)
2014	1603	2036	55.9%
2015	1470	1927	56.7%
2016	1208	1644	57.6%
2017	1005	1604	61.5%
2018	883	1435	61.9%
Total:	14603	21875	60.0%

Source: Author.

Table A 7 - Number of enrolled students at UNIBL by municipalities, 2007-2018. The rest of 46 municipalities contribute with less than 20 students per municipality.

Municipality	**No of students**	**Municipality**	**No of students**	**Municipality**	**No of students**
Banja Luka	11654	Kneževo	341	Bihać	68
Prijedor	1978	Brčko	316	Petrovo	65
Gradiška	1755	Šamac	246	Ljubinje	63
Prnjavor	1293	Brod	244	Cazin	57
Laktaši	1240	Nevesinje	223	Ključ	50
Doboj	1151	Drvar	178	Oštra Luka	50
Teslić	993	Bileća	177	Berkovići	28
Mrkonjić Grad	739	Ribnik	167	Bratunac	26
Kotor Varoš	692	Bijeljina	147	Milići	26
Novi Grad	670	Bosanski Petrovac	128	Travnik	26
Čelinac	666	Sanski Most	126	Sokolac	25
Derventa	592	Kostajnica	102	Višegrad	25
Kozarska Dubica	587	Gacko	98	Rogatica	23
Modriča	548	Foča	81	Glamoč	22
Srbac	536	Zvornik	76	Livno	21
Trebinje	431	Jajce	73	Velika Kladuša	21
Šipovo	366				

Source: Author.

Table A 8 – UNIBL, number of enrolled students by type of enrollment into first year of study, sample 37,667, 2007/08-2018/19

Year of enroll	Dropout	Passive	Transfered from another faculty	Normal	Transferd from anothre study programe	Total
2007	44	0	23	2477	1	2545
2008	97	2	24	2542	0	2665
2009	253	4	36	3248	1	3542
2010	209	3	38	3357	0	3607
2011	220	4	30	2880	0	3134
2012	231	7	31	3199	0	3468

Year of enroll	Dropout	Passive	Transfered from another faculty	Normal	Transferd from anothre study programe	Total
2013	418	21	41	3315	0	3795
2014	399	18	38	3223	0	3678
2015	429	15	30	2959	0	3433
2016	538	9	36	2286	0	2869
2017	347	10	49	2200	6	2612
2018	381	6	25	1828	79	2319
Total	3566	99	401	33514	87	37667
Share:	9.5%	0.3%	1.1%	89.0%	0.2%	

Source: Author

Table A 9 - UNIBL, number of enrolled students by status of enrollment (of finance) into first year of study, sample 37,667, 2007/08-2018/19

Year of enroll	Part-time	Self-finance	Foreigner	Co-finance	Scholarship	Total
2007	9	0	1	1328	1207	2545
2008	8	2	1	1402	1252	2665
2009	10	0	5	1807	1720	3542
2010	32	0	9	1737	1829	3607
2011	232	0	16	1885	1001	3134
2012	209	0	16	1864	1379	3468
2013	161	0	43	2186	1405	3795
2014	112	0	39	2212	1315	3678
2015	66	0	26	2325	1016	3433
2016	63	11	34	1818	943	2869
2017	49	30	37	1492	1004	2612
2018	66	92	42	1086	1033	2319
Total	1017	135	269	21142	15104	37667
Share:	2.7%	0.4%	0.7%	56.1%	40.1%	

Source: Author

Table A 10 – Summary table of numerical variables description, sample size 37,667 used for churn classification before ML modeling

Variable in database	count	mean	Std	min	25%	50%	75%	Max
Faculty	37667	108.57	4.63	101.00	105.00	108.00	113.00	117.00
Index	37667							
Gender	36478	1.60	0.49	1.00	1.00	2.00	2.00	2.00
score_t	10913	70.79	116.81	0.00	59.39	70.54	80.93	8208.00*
score_e	11827	38.86	24.39	2.49	34.38	39.29	44.58	2511.00*
Duration	37667	0.89	2.30	0.00	0.00	0.00	0.00	201.77*
type_2007	2703	1.03	0.23	1.00	1.00	1.00	1.00	3.00
sy_2007	2703	1.00	0.02	1.00	1.00	1.00	1.00	2.00
ects_2007	37667	1.66	8.91	0.00	0.00	0.00	0.00	98.00*

Variable in database	count	mean	Std	min	25%	50%	75%	Max
npe_2007	1602	6.90	4.00	1.00	4.00	7.00	10.00	20.00
type_2008	4761	1.11	0.31	1.00	1.00	1.00	1.00	3.00
sy_2008	4761	1.31	0.47	1.00	1.00	1.00	2.00	3.00
ects_2008	37667	2.45	10.26	0.00	0.00	0.00	0.00	110.00*
npe_2008	2742	6.27	4.31	1.00	2.00	6.00	9.00	20.00
type_2009	7587	1.16	0.40	1.00	1.00	1.00	1.00	3.00
sy_2009	7587	1.57	0.73	1.00	1.00	1.00	2.00	5.00
ects_2009	37667	4.27	13.45	0.00	0.00	0.00	0.00	118.00*
npe_2009	4243	6.99	4.24	1.00	4.00	6.00	9.00	22.00
type_2010	10047	1.23	0.50	1.00	1.00	1.00	1.00	4.00
sy_2010	10047	1.89	1.23	1.00	1.00	2.00	2.00	10.00
ects_2010	37667	6.01	15.79	0.00	0.00	0.00	0.00	117.00*
npe_2010	5948	6.97	4.30	1.00	4.00	7.00	10.00	22.00
type_2011	11678	1.30	0.57	1.00	1.00	1.00	1.00	5.00
sy_2011	11678	2.36	1.94	1.00	1.00	2.00	3.00	10.00
ects_2011	37667	6.98	16.77	0.00	0.00	0.00	0.00	120.00*
npe_2011	7099	6.55	4.04	1.00	3.00	6.00	9.00	23.00
type_2012	13497	1.34	0.66	1.00	1.00	1.00	2.00	5.00
sy_2012	13497	2.74	2.43	1.00	1.00	2.00	3.00	10.00
ects_2012	37667	7.58	16.97	0.00	0.00	0.00	0.00	110.00*
npe_2012	8118	6.22	3.95	1.00	3.00	6.00	9.00	38.00
type_2013	14904	1.44	0.78	1.00	1.00	1.00	2.00	6.00
sy_2013	14904	2.98	2.75	1.00	1.00	2.00	3.00	10.00
ects_2013	37667	8.27	17.42	0.00	0.00	0.00	0.00	165.00*
npe_2013	9149	6.02	3.87	1.00	3.00	6.00	9.00	33.00
type_2014	15509	1.48	0.86	1.00	1.00	1.00	2.00	7.00
sy_2014	15509	3.09	2.82	1.00	1.00	2.00	3.00	10.00
ects_2014	37667	8.63	17.60	0.00	0.00	0.00	5.00	140.00*
npe_2014	9869	5.95	4.03	1.00	3.00	5.00	9.00	56.00
type_2015	15340	1.52	0.90	1.00	1.00	1.00	2.00	8.00
sy_2015	15340	3.21	2.96	1.00	1.00	2.00	3.00	10.00
ects_2015	37667	8.52	17.62	0.00	0.00	0.00	5.00	202.00*
npe_2015	9657	6.00	4.13	1.00	3.00	5.00	9.00	55.00
type_2016	14399	1.57	0.99	1.00	1.00	1.00	2.00	9.00
sy_2016	14399	3.39	3.07	1.00	1.00	2.00	4.00	10.00
ects_2016	37667	7.63	16.71	0.00	0.00	0.00	0.00	193.50*
npe_2016	8879	5.84	4.01	1.00	2.00	5.00	8.00	33.00
type_2017	13108	1.58	1.04	1.00	1.00	1.00	2.00	10.00
sy_2017	13108	3.46	3.12	1.00	1.00	2.00	4.00	10.00
ects_2017	37667	8.76	17.78	0.00	0.00	0.00	5.00	173.00*
npe_2017	9988	5.95	3.90	1.00	3.00	6.00	9.00	30.00
type_2018	11998	1.59	1.10	1.00	1.00	1.00	2.00	10.00
sy_2018	11998	3.53	3.14	1.00	1.00	2.00	4.00	10.00
ects_2018	37667	6.88	15.29	0.00	0.00	0.00	0.00	94.00*
npe_2018	8701	5.37	3.63	1.00	2.00	5.00	8.00	24.00

*Outliers.

Source: Authors contribution.

Table A 11 - Summary of correlation by Pearson, Spearman, Kendall and Phi for all variables and the target variable dropout, sorted by Phi coefficient

Variable	Pearson	Spearman	Kendall	Phi
npe_1	0.5407	0.5551	0.4706	0.7283
ects_1	0.4464	0.4118	0.3636	0.6514
ECTS_between 41 and 60	0.4487	0.4487	0.4487	0.6477
ECTS_less than 20	0.3943	0.3943	0.3943	0.5804
t_dropout	0.2738	0.2738	0.2738	0.4165
t_normal	0.2663	0.2663	0.2663	0.4058
score_e	0.1065	0.2268	0.1854	0.3806
score_t	0.2388	0.2756	0.2251	0.3805
gender_2.0	0.1997	0.1997	0.1997	0.3082
ID	0.0832	0.0952	0.0778	0.3007
sci_1	0.1638	0.1638	0.1638	0.2541
sci_2	0.1559	0.1559	0.1559	0.2421
s_scholarship	0.1328	0.1328	0.1328	0.2066
mld_missing	0.1306	0.1306	0.1306	0.2033
dist_0	0.1293	0.1293	0.1293	0.2013
f_Electrical_Engineering	0.1284	0.1284	0.1284	0.1998
f_Philosophy	0.1255	0.1255	0.1255	0.1953
s_co-financing	0.1152	0.1152	0.1152	0.1795
gender_1.0	0.1147	0.1147	0.1147	0.1787
f_Political	0.1117	0.1117	0.1117	0.1739
f_Agriculture	0.0980	0.0980	0.0980	0.1526
f_ACEG	0.0877	0.0877	0.0877	0.1366
f_Mehanical_Engineering	0.0854	0.0854	0.0854	0.1329
hs_missing	0.0837	0.0837	0.0837	0.1305
ECTS_more than 60	0.0804	0.0804	0.0804	0.1246
hsd_Gymnasium	0.0758	0.0758	0.0758	0.1181
ent_1st_y	0.0943	0.0938	0.0820	0.1140
hs_econ	0.0701	0.0701	0.0701	0.1090
ECTS_ between 21 and 40	0.0697	0.0697	0.0697	0.1085
f_Academy_of_Arts	0.0691	0.0691	0.0691	0.1071
dist_between 81 and 160	0.0659	0.0659	0.0659	0.1026
hs_stem	0.0645	0.0645	0.0645	0.1004
dur_3.0	0.0629	0.0629	0.0629	0.0978
dur_4.0	0.0629	0.0629	0.0629	0.0978
age1	0.0790	0.1207	0.0986	0.0939
f_Natural_Sciences_and_Math	0.0603	0.0603	0.0603	0.0938
hsd_Economics	0.0584	0.0584	0.0584	0.0907
hs_o	0.0573	0.0573	0.0573	0.0890
mld_1	0.0541	0.0541	0.0541	0.0841
hs_gym	0.0538	0.0538	0.0538	0.0835
f_Philology	0.0518	0.0518	0.0518	0.0802
dist_more than 241	0.0509	0.0509	0.0509	0.0787
s_part-time	0.0476	0.0476	0.0476	0.0734

Variable	Pearson	Spearman	Kendall	Phi
f_Technology	0.0464	0.0464	0.0464	0.0716
hsd_STEM	0.0443	0.0443	0.0443	0.0684
mld_2	0.0427	0.0427	0.0427	0.0659
f_Mining	0.0405	0.0405	0.0405	0.0619
mld_3	0.0365	0.0365	0.0365	0.0559
t_passive_year	0.0354	0.0354	0.0354	0.0528
dist_up to 80	0.0309	0.0309	0.0309	0.0471
f_Security	0.0307	0.0307	0.0307	0.0459
dist_between 161 and 240	0.0273	0.0273	0.0273	0.0409
f_Medical	0.0271	0.0271	0.0271	0.0407
sci_3	0.0271	0.0271	0.0271	0.0407
f_PES	0.0259	0.0259	0.0259	0.0386
hs_medic	0.0213	0.0213	0.0213	0.0311
mld_4	0.0206	0.0206	0.0206	0.0300
t_acknowledged_from_a_f	0.0193	0.0193	0.0193	0.0274
mld_Montenegro	0.0198	0.0198	0.0198	0.0250
hsd_Medicine	0.0127	0.0127	0.0127	0.0157
hsd_Art	0.0122	0.0122	0.0122	0.0130
mld_Croatia	0.0130	0.0130	0.0130	0.0130
hsd_Lower vocation	0.0109	0.0109	0.0109	0.0108
ECTS_0	0.0119	0.0119	0.0119	0.0060
f_Economics	0.0063	0.0063	0.0063	0.0000
f_Forestry	0.0059	0.0059	0.0059	0.0000
f_Law	0.0067	0.0067	0.0067	0.0000
hsd_Service	0.0014	0.0014	0.0014	0.0000
mld_Serbia	0.0039	0.0039	0.0039	0.0000
s_foreigner	0.0005	0.0005	0.0005	0.0000
s_self-financing	0.0099	0.0099	0.0099	0.0000
t_enroll_from_a_sp	0.0099	0.0099	0.0099	0.0000

Source: Author.

Table A 12 - Logit model – odds ratio

Variable	Positive score (1)	Variable	Negative score (0)
f_Philosophy	0.7875	ECTS collected at the end of 1st year	-1.8060
f_Philology	0.6220	ID	-1.3160
f_Forestry	0.5598	Enroll: normal	-1.1727
f_Technology	0.3153	f_ACEG	-0.9826
Enter 1st year	0.3102	f_Economics	-0.8098
f_Law	0.2786	Gender: female	-0.7151
Social science domain	0.2041	f_Political	-0.5697
f_Mining	0.1796	Number of passed courses	-0.4897
f_Agriculture	0.1738	Gender: male	-0.4682
f_PES	0.1720	T:acknowledged from another faculty	-0.4290
study duration is 4 years	0.1266	STEM domain	-0.2164

Variable	Positive score (1)	Variable	Negative score (0)
Part time student	0.0989	f_Natural_Sciences_and_Math	-0.1720
STEM High school	0.0957	Scholarship holder	-0.1284
Co-finansing study	0.0922	Status: passive year	-0.1268
High school: other	0.0748	High school degree: STEM	-0.1132
Age at enrollment	0.0730	hsd_Gymnasium	-0.1067
ECTS_0	0.0729	mld_Montenegro	-0.1061
ECTS_between 41 and 60	0.0502	score_t	-0.0968
Medical science	0.0387	hs_missing	-0.0871
ECTS_more than 60	0.0376	f_Academy_of_Arts	-0.0845
mld_3	0.0335	f_Electrical_Engineering	-0.0768
dist_up to 80	0.0334	ECTS_less than 20	-0.0729
mld_Croatia	0.0279	s_self-financing	-0.0694
ECTS_ between 21 and 40	0.0268	f_Security	-0.0600
mld_Serbia	0.0261	hsd_Medicine	-0.0596
hsd_Lower vocation	0.0251	t_enroll_from_a_sp	-0.0527
s_foreigner	0.0241	mld_missing	-0.0384
score_e	0.0202	dist_more than 241	-0.0340
mld_1	0.0112	hsd_Service	-0.0337
mld_4	0.0096	hs_econ	-0.0332
hsd_Economics	0.0087	dist_between 161 and 240	-0.0247
hsd_Art	0.0022	hs_gym	-0.0191
mld_2	0.0019	dist_between 81 and 160	-0.0147
		f_Mehanical_Engineering	-0.0056
		dist_0	-0.0033
		hs_medic	-0.0014

Source: Logit in Python

Table A 13 – Sankey chart data source: What happens with students after enrollment, by cohort

Year of enroll	Enrolled by generation (a)	Graduated (2007-18) (b)	Drop out by low (c)=(f)+(g)	Drop out by request, total (d)	Study (e)=(a-b-c-d)	Dropout by request permanetn (f) = 0.3158*(c)	Dropout by request to continue (g)=0.6842*(c)
2007	2545	710	827	246	762	78	168
2008	2665	762	864	290	749	92	198
2009	3542	976	959	741	866	234	507
2010	3607	1000	1024	727	856	230	497
2011	3134	823	867	734	710	232	502
2012	3468	843	1032	843	750	266	577
2013	3795	764	1088	991	952	313	678
2014	3678	498	959	1,046	1,175	330	716
2015	3433	136	805	998	1,494	315	683
2016	2869	17	534	846	1,472	267	579
2017	2612	0*	343	584	1,685	184	400
2018	2319	0*	0**	381	1,938	120	261

Year of enroll	Enrolled by generation (a)	Graduated (2007-18) (b)	Drop out by low (c)=(f)+(g)	Drop out by request, total (d)	Study (e)=(a-b-c-d)	Dropout by request permanetn (f) = 0.3158*(c)	Dropout by request to continue (g)=0.6842*(c)
Total:	37667	6529	9302	8427	13409	2661	5766
Share in total:		17.3%	24.7%	22.4%	35.6%	7.1%	15.3%

*Generation started in 2017 and 2018 do not have graduated students.
**Generation 2018 does not meet the dropout by law criteria.
Source: Authors contribution.

Table A 14 – Total permanent dropout at UNIBL, by generation and study year

Generation	1st year	2nd year	3rd year	4th year	5th year	6th year
2007	17.7	5.7	2.8	2.6	1.7	1.1
2008	17.1	5.8	3.8	2.5	1.7	1.3
2009	14.7	5.4	3.4	3.0	1.7	2.2
2010	15.0	5.5	4.3	2.6	2.7	1.9
2011	12.6	6.7	5.3	4.3	2.5	1.7
2012	15.5	7.3	6.6	4.0	2.0	1.6
2013	15.6	9.5	5.0	3.9	2.5	0.5
2014	15.8	8.7	5.6	4.1	0.8	
2015	18.9	7.5	5.0	1.2		
2016	18.1	8.1	1.7			
2017	17.3	2.8				
2018	5.2					
Average:	15.3	6.1	3.6	2.3	1.3	0.9

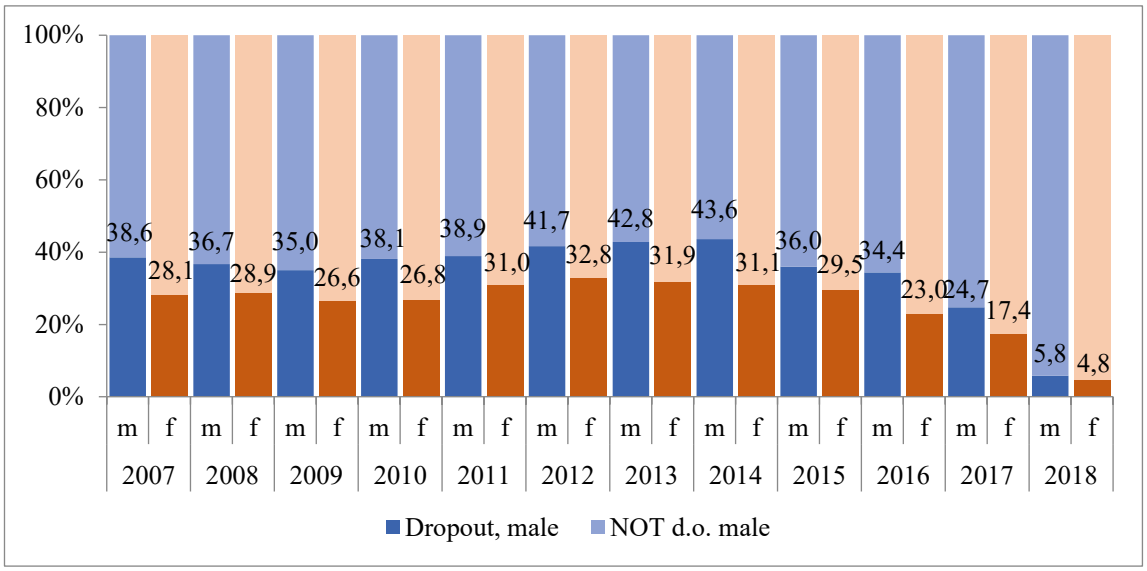
Source: Author.

Table A 15 – Total dropout rates by gender (Male, Female), UNIBL; 2007-2018 (in percentages), sample size Male 14,603; Female 21,800.

	1st year		**2nd year**		**3rd year**		**4th year**		**5th year**		**6th year**	
Generation	M	F	M	F	M	F	M	F	M	F	M	F
2007	15.4	12.1	7.9	5.0	3.5	2.5	3.5	2.2	1.9	1.6	1.6	1.0
2008	13.0	11.0	7.5	5.5	4.9	3.6	2.6	2.6	2.0	1.7	2.0	1.1
2009	12.5	8.2	6.5	5.1	4.6	2.9	3.2	3.1	1.5	2.0	2.2	2.4
2010	14.2	7.9	7.2	5.0	4.5	4.5	2.9	2.6	3.1	2.7	2.5	1.7
2011	14.3	10.9	6.6	6.3	6.6	4.2	4.0	4.4	3.1	1.9	1.9	1.7
2012	17.9	12.0	8.2	6.4	6.3	7.0	4.6	3.6	2.3	1.9	1.9	1.4
2013	19.1	11.8	9.7	9.4	6.2	4.4	4.8	3.4	2.6	2.5	0.5	0.5
2014	18.8	12.7	9.6	8.1	6.1	5.2	8.4	4.1	0.7	0.9		
2015	22.1	15.5	8.0	7.3	4.7	5.3	1.1	1.4				
2016	22.1	15.0	10.5	6.2	1.8	1.7						
2017	21.8	14.6	2.9	2.8								
2018	5.8	4.8										
Avg:	16.4	11.4	7.7	6.1	4.9	4.1	3.9	3.0	2.1	1.9	1.8	1.4

Source: Author's contribution

Figure A 1 - Total dropout rates by gender (Male, Female), UNIBL; 2007-2018, sample size Male 14,603; Female 21,800.



Source: Author

Table A 16 – Grouping faculties by science category

Faculty at UNIBL	Science category classification	Total number of enrolled students in freshmen years from 2007/08-2018/19
Faculty of Economics		
Faculty of Law		
Faculty of Security Science		
Faculty of Political Science	Social science	17,214
Faculty of Physical Education and Sport		
Faculty of Philology		
Faculty of Philosophy		
Faculty of Medicine	Medical science	4,018*
Academy of Arts		
Faculty of Architecture, Civil Engineering and Geodesy		
Faculty of Electrical Engineering		
Faculty of Agriculture	STEAM science	16,410
Faculty of Natural Sciences and Mathematics		
Faculty of Mining		
Faculty of Technology		
Faculty of Forestry		

*Faculty of Medicine in the period of 2007-2018 did not use the database in full capacity.

Table A 17 – Dropout in social science, years after enrollment, UNIBL, 2007-2018 (in percentages), sample size 17,084.

Generation	1st year	2nd year	3rd year	4th year	5th year	6th year
2007	12.9	6.2	3.1	2.9	2.0	0.9
2008	10.8	5.0	4.3	2.4	0.9	1.2
2009	9.1	5.7	2.9	1.9	1.1	2.0
2010	10.6	4.5	3.4	2.7	3.1	1.2
2011	10.9	4.2	5.4	4.1	1.7	1.9
2012	11.7	5.9	7.1	3.5	1.6	1.9
2013	13.6	10.9	4.3	3.2	3.0	0.9
2014	13.9	5.2	4.3	5.2	1.2	
2015	17.7	6.1	5.3	2.0		
2016	15.7	7.1	2.4			
2017	18.1	2.6				
2018	5.6					
Avg:	12.5	5.3	3.5	2.3	1.2	0.8

Source: Author's contribution

Table A 18 - Dropout rates in social science,1-6 years after enrollment, by gender (Male, Female), UNIBL; 2007-2018 (in percentages), sample size Male 5,259; Female 11,856.

	1st year		2nd year		3rd year		4th year		5th year		6th year	
Generation	M	F	M	F	M	F	M	F	M	F	M	F
2007	15.6	11.7	9.1	5.1	4.5	2.4	4.4	2.2	2.5	1.8	1.6	0.7
2008	16.3	8.7	7.2	4.1	5.6	3.9	1.7	2.7	0.9	0.9	2.4	0.7
2009	16.8	6.4	6.8	5.2	4.4	2.5	3.0	1.5	1.2	1.1	2.6	1.8
2010	20.5	6.6	6.2	3.8	3.3	3.4	3.2	2.5	3.6	2.9	1.1	1.2
2011	17.0	8.3	4.2	4.2	7.7	4.4	5.0	3.6	1.9	1.6	2.3	1.7
2012	19.3	8.1	5.9	5.7	6.4	7.3	5.0	3.0	1.6	1.6	2.6	1.7
2013	18.4	10.6	13.9	9.4	5.8	3.7	4.5	2.6	3.6	2.8	0.9	0.7
2014	19.0	10.4	6.2	4.6	4.8	4.1	5.3	5.2	0.8	1.4		
2015	20.5	14.6	7.7	5.4	4.9	5.6	1.7	2.2				
2016	21.7	12.5	12.2	4.6	2.5	2.3						
2017	25.1	14.9	2.3	2.8								
2018	6.2	5.3										
Avg:	18.0	9.8	6.8	4.6	4.2	3.3	2.8	2.1	1.3	1.2	1.1	0.7

Source: Author's contribution

Table A 19 – Dropout rates in STEAM science, UNIBL, 2007-2018 (in percentages), sample size 16,410.

Generation	1st year	2nd year	3rd year	4th year	5th year	6th year
2007	27.3	6.2	2.0	2.2	1.1	0.8
2008	26.6	3.5	3.2	2.1	1.8	1.0
2009	19.7	5.0	3.1	3.1	1.7	1.7
2010	21.8	5.7	4.5	2.4	2.1	2.6
2011	11.6	8.1	5.6	4.5	3.4	1.8
2012	19.3	9.3	5.7	3.9	2.6	1.2
2013	17.8	7.8	5.2	4.0	1.9	0.3
2014	17.9	11.7	6.3	3.4	0.5	
2015	21.1	9.0	5.0	0.5		
2016	21.3	9.1	1.1			

Generation	1st year	2nd year	3rd year	4th year	5th year	6th year
2017	17.5	2.8				
2018	5.8					
Avg:	19.0	6.5	3.5	2.2	1.3	0.8

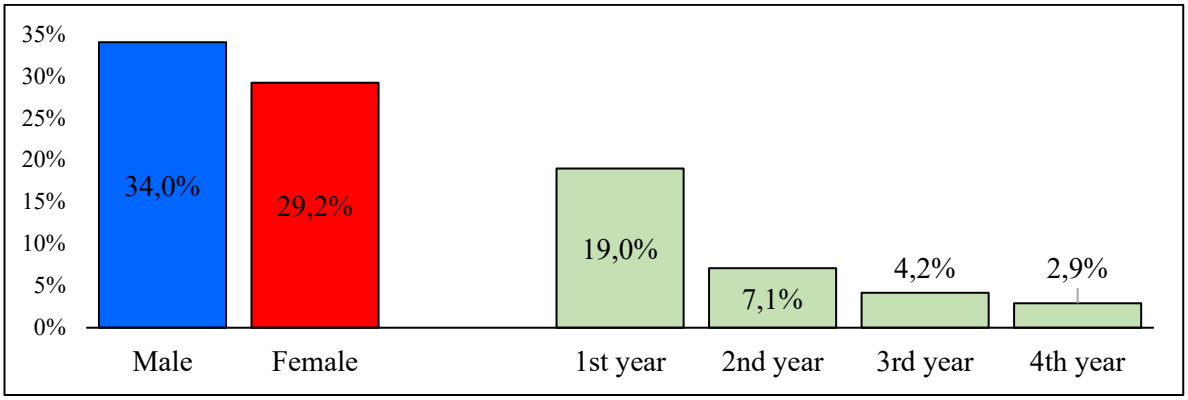
Source: Author's contribution

Table A 20 - Dropout rates in STEAM discipline, by gender (Male, Female), UNIBL; 2007-2018 (in percentages), sample size Male 8,341; Female 7,028.

	1st year		2nd year		3rd year		4th year		5th year		6th year	
Generation	M	F	M	F	M	F	M	F	M	F	M	F
2007	14.4	12.6	7.9	7.1	2.5	1.9	3.3	1.6	1.4	0.9	1.1	0.8
2008	10.5	14.0	4.4	3.7	4.4	3.0	2.3	2.6	2.3	2.0	1.4	1.1
2009	9.1	4.2	6.6	2.7	4.0	1.8	2.9	3.5	1.5	2.6	1.8	2.0
2010	10.3	11.8	6.8	6.0	4.6	5.7	2.9	2.4	2.8	2.1	3.8	2.1
2011	10.8	11.0	7.2	8.0	5.7	4.5	3.4	5.2	3.8	2.3	1.7	2.2
2012	16.9	17.4	10.0	8.1	5.8	6.2	4.0	3.7	2.9	2.5	1.4	1.2
2013	19.3	13.3	7.3	8.7	6.0	4.6	4.6	3.3	2.0	1.8	0.3	0.3
2014	18.9	15.8	11.4	12.3	6.4	6.1	3.8	2.9	0.7	0.3		
2015	23.9	17.5	8.4	9.9	4.8	5.4	0.8	0.2				
2016	22.7	19.4	9.8	8.3	1.4	0.8						
2017	20.1	15.0	3.1	2.4								
2018	6.1	5.5										
Avg:	15.2	13.1	6.9	6.4	3.8	3.3	2.3	2.1	1.4	1.2	1.0	0.8

Source: Author's contribution

Figure A 2 – STEAM dropout by gender and average, by year of study by generations 2007/08-2018/19 school year.



* Due to the lack of gender labels in the database, the average dropout by year of study represents both genders plus students with missing gender labels.
Source: Appendix, Table A7.

Table A 21 - Dropout in medical science, UNIBL, 2007-2018 (in percentages), sample size 4,081.

Generation	1st year	2nd year	3rd year	4th year	5th year	6th year
2007	14.3	2.6	3.5	2.0	1.7	2.6
2008	13.7	14.2	3.7	3.5	4.0	2.5
2009	14.6	4.6	5.7	6.2	2.7	3.0
2010	6.4	10.6	7.6	3.0	3.0	2.4
2011	24.3	11.5	4.2	4.8	2.2	1.0
2012	15.2	4.8	8.7	6.8	0.7	1.3
2013	14.8	10.8	7.2	6.9	2.8	0.1
2014	13.0	8.6	8.0	3.4	0.7	
2015	13.5	7.2	3.5	1.3		
2016	13.7	7.2	1.9			
2017	13.3	4.0				
2018	2.7					
Avg:	13.3	7.2	4.5	3.2	1.5	1.1

Source: Author's contribution

Table A 22 - Dropout rates in the medical discipline, by gender (Male, Female), UNIBL; 2007-2018 (in percentages), sample size Male 1,003; Female 2,991.

	1st year		2nd year		3rd year		4th year		5th year		6th year	
Generation	M	F	M	F	M	F	M	F	M	F	M	F
2007	19.0	12.7	14.3	2.7	3.6	3.5	0.0	2.7	1.2	1.9	3.6	2.3
2008	12.0	14.3	22.0	11.6	5.0	3.3	7.0	2.3	5.0	3.7	3.0	2.3
2009	23.0	11.5	4.0	4.8	10.0	4.1	6.0	6.3	2.0	3.0	4.0	2.6
2010	10.0	5.4	17.1	8.5	11.4	6.6	1.4	3.5	2.9	3.1	0.0	3.1
2011	34.9	21.5	15.9	10.1	9.5	2.8	3.2	5.3	3.2	2.0	1.6	0.8
2012	19.5	13.9	5.2	4.8	10.4	8.2	7.8	6.5	0.4	0.9	2.6	0.9
2013	22.0	12.9	8.5	1.1	9.8	6.6	8.9	6.4	2.4	3.0	0.4	0.1
2014	16.5	11.8	8.8	8.8	9.9	7.5	2.2	3.5	2.1	0.1		
2015	14.4	13.3	6.7	7.5	2.2	3.9	3.9	0.4				
2016	17.7	12.2	7.6	6.6	4.8	0.8						
2017	17.6	12.0	12.1	1.3								
2018	7.3	1.0										
Avg:	17.8	11.9	10.2	5.6	6.4	3.9	3.4	3.1	1.6	1.5	1.3	1.0

Source: Author's contribution

Table A 23 – Summary of HGBC models performance without variables that contain large amount of missing data, and with inclusion of faculty variables

HGBC	Enroll*	End of 1st year**	Enroll data***
Accuracy	0.74669	0.82223	0.755112
True Negative	1615	1726	1634
False Positive	476	365	457
False Negative	577	374	561
True Positive	1489	1692	1505
True Negative	38.85%	41.52%	0.393072
False Positive	11.45%	8.78%	0.109935
False Negative	13.88%	9.00%	0.134953
True Positive	35.82%	40.70%	0.36204

HGBC	Enroll*	End of 1st year**	Enroll data***
F1	0.73877	0.82076	0.747269
Precision	0.75776	0.82256	0.767074
Recall	0.72072	0.81897	0.728461
ROCAUC	0.74654	0.82221	0.754953
Matthews corr.	0.49380	0.64444	0.510699
PR score	0.85293	0.92026	
Specificity	0.77236	0.82544	0.781444
Train acc.	0.80463	0.86111	0.807519

*Without variables *score_e, score_t*, that are part of enrollment dataset.
**Without variables *npe_1*, (part of end of first year dataset) *score_t, score_e*.
***With variables of belonging to the faculties, that are part of enrollment dataset.

Table A 24 - Pre-enrollment feature importance by SHAP and PI, sorted by HGBC PI.

HGBC PI Pre enroll variable	HGBC PI	HGBC Sh.	RF PI	RF Sh.	SVM linear PI	SVM linear Sh.	SVM poly d=3 PI	SVM poly d=3 Sh.	SVM poly d=8. PI	SVM poly d=8 Sh.	SVM sigmoid PI	SVM sigmoid Sh.	SVM RBF g=0.1 PI	SVM RBF g=0.1 Sh.	SVM RBF g=0.5 PI	SVM RBF g=0.5 Sh.	SVM RBF g=1.0 PI	SVM RBF g=1.0 Sh.	NN 2L PI	NN 2L Sh.	NN 3L PI	NN 3L Sh.	NN 4L PI	NN 4L Sh.
gender_2.0	1	1	1	1	1	1	1	1	1	1	3	4	1	1	1	1	1	1	1	1	1	1	1	1
age1	2	4	2	3	3	4	3	3	3	3	6	1	3	4	7	3	3	3	3	4	3	4	3	4
ent_1st_y	3	3	3	2	4	3	10	4	14	27	4	5	5	7	27	4	5	27	7	3	5	12	19	3
gender_1.0	4	2	4	4	5	7	7	27	22	7	1	26	6	27	5	27	22	4	4	7	7	6	26	7
dist_up to 80	5	7	9	7	7	27	4	5	5	6	2	27	26	3	14	26	7	26	5	8	4	3	21	5
hs_missing	6	6	16	6	22	9	22	16	16	5	25	7	18	26	10	7	9	7	29	6	10	26	5	8
mld_missing	7	27	7	5	9	26	5	26	27	26	9	6	14	19	22	16	14	14	6	5	27	7	10	6
hs_stem	8	8	10	26	21	5	16	7	26	22	15	3	12	5	26	14	10	5	2	12	2	5	6	26
hsd_Gymnasium	9	10	17	27	27	6	26	22	7	16	16	22	27	9	19	9	16	16	28	26	12	27	8	14
hsd_STEM	10	5	24	8	2	22	21	8	10	8	19	9	21	10	15	22	8	21	22	10	6	14	12	27
dist_more than 241	11	26	21	14	8	21	27	14	8	19	24	19	9	6	21	6	27	22	10	21	26	2	27	18
hs_econ	12	9	23	10	11	12	14	21	4	14	23	8	7	16	11	21	26	10	8	14	16	18	3	10
dist_between 161 and 240	13	22	25	16	14	14	8	10	9	18	20	10	11	14	13	5	15	9	16	2	14	21	16	21
hs_o	14	16	20	9	6	2	9	18	12	4	17	16	2	12	2	10	4	19	12	18	21	10	2	2
mld_4	15	19	8	22	18	16	2	6	19	24	13	18	17	8	20	18	6	8	14	16	9	24	19	12
hsd_Economics	16	11	15	18	13	8	15	9	25	21	11	14	22	11	25	12	19	12	21	27	11	8	9	16
hsd_Service	17	12	11	21	16	11	13	13	18	9	10	12	13	21	23	11	13	18	18	24	18	25	22	9
hs_gym	18	13	13	12	10	18	24	19	11	12	8	21	19	22	24	15	12	13	9	19	8	16	15	22
mld_3	19	14	19	19	12	19	25	12	15	25	12	11	15	13	8	13	24	25	19	25	17	9	17	11
hsd_Lower vocation	20	21	12	11	20	10	12	15	2	2	14	25	25	15	12	25	20	24	20	22	13	22	13	17

The rest of variables sorted by PI importance for HGBC model are: 21) *mld_2*, 22) *dist_between_81_and_160*, 23) *hsd_Art*, 24), *hsd_Medicine*, 25) *hs_medic*, 26) *mld_1*, 27) *dist_0*.

Table A 25 - Enrollment feature importance by SHAP and PI, sorted by HGBC PI

HGBC PI enroll test importance	HGBC PI	HGBC Sh.	RF PI	RF Sh.	SVM linear PI	SVM linear Sh.	SVM poly d=3 PI	SVM poly d=3 Sh.	SVM poly d=8 PI	SVM poly d=8 Sh.	SVM sigmo1 PI	SVM sigmo1 Sh.	SVM RBF g=0.1 PI	SVM RBF g=0.1 Sh.	SVM RBF g=0.5 PI	SVM RBF g=0.5 Sh.	SVM RBF g=1.0 PI	SVM RBF g=1.0 Sh.	NN 2L PI	NN 2L Sh.	NN 3L PI	NN 3L Sh.	NN 4L PI	NN 4L Sh.
ID	1	1	1	1	2	2	2	4	4	4	3	19	2	2	3	4	3	4	2	3	2	3	2	3
gender_2.0	2	3	3	2	3	5	3	2	2	2	2	4	3	4	4	2	4	19	19	2	3	2	3	2
t_normal	3	2	7	3	5	7	4	5	3	1	20	2	4	5	13	19	19	2	3	19	1	35	1	19
s_scholarship	4	4	2	7	19	19	5	7	1	39	24	8	5	19	7	7	13	23	16	12	8	19	12	5
gender_1.0	5	5	6	6	12	3	7	8	7	23	15	23	7	8	2	8	2	8	1	5	16	16	19	1
age1	6	8	13	4	4	12	1	39	13	7	7	39	12	23	19	23	7	39	8	35	19	1	8	16
ent_1st_y	7	7	4	8	7	4	15	26	16	37	35	12	19	7	1	39	1	5	12	23	12	5	23	35
mld_missing	8	6	8	23	29	39	16	1	23	12	18	7	1	3	35	1	35	7	5	16	7	7	5	12
hsd_Gymnasium	9	23	16	13	13	8	13	23	19	11	9	5	35	39	5	3	12	12	23	1	4	12	7	7
score_t	10	39	35	19	6	1	19	37	39	3	37	13	13	11	15	5	16	1	13	39	13	37	16	37
dist_between_81 a nd_160	11	9	5	5	17	16	8	3	12	20	6	16	17	12	16	37	15	16	35	7	5	23	13	4
dist_up_to_80	12	10	23	16	35	15	17	34	5	34	22	3	18	1	8	16	5	37	15	4	35	39	4	13
dur_3.0	13	22	15	15	22	13	35	16	17	35	32	32	24	13	21	34	11	34	37	37	37	13	15	23
score_e	14	34	12	39	9	18	34	15	18	8	21	11	34	16	37	15	37	11	39	8	39	8	35	39
hsd_STEM	15	13	19	12	20	23	12	11	37	18	28	1	16	15	12	12	8	3	29	18	18	4	39	9
hs_missing	16	12	37	35	24	17	29	20	15	17	38	37	37	32	11	9	21	15	18	11	15	22	18	8
hs_econ	17	14	20	9	23	6	24	18	11	32	26	35	32	9	24	11	29	18	4	9	23	17	11	17
hs_gym	18	18	29	34	27	35	6	12	34	22	30	15	9	35	23	20	22	32	7	13	11	9	20	22
s_co-financing	19	29	32	37	32	32	11	35	20	6	33	9	38	37	32	13	9	20	34	22	34	18	29	29
mld_2	20	30	39	18	33	9	23	13	8	19	27	34	21	34	22	17	39	24	17	17	9	11	34	11

The rest of variables sorted by PI importance for HGBC model are: 21) *mld_4*, 22) *t_acknowledged_from_a_f*, 23) *dist_0*, 24) *dist_more_than241*, 25) *s_foreigner*, 26) *s_self-financing*, 27) *hsd_Lower_vocation*, 28) *hsd_Service*, 29) *s_part-time*, 30) t_*passive_year*, 31) *hs_medic*, 32) *mld_3*, 33) *hsd_Art*, 34) *hsd_Economics*, 35) *hs_stem*, 36) *hsd_Medicine*, 37) *hs_0*, 38) *dist_between_161_and_240*, 39) *mld_1*.

Table A 26 – End of year feature importance by SHAP and PI, sorted by HGBC PI

HGBC End of 1st year	HGBC PI	HGBC Sh.	RF PI	RF Sh.	SVM linear PI	SVM linear Sh	SVM poly d=3 PI	SVM poly d=3 Sh	SVM poly d=8 PI	SVM poly d=8 Sh	SVM sigmoid PI	SVM sigmoid	SVM RBF σ=0.1 PI	SVM RBF σ=0.1 Sh	SVM RBF σ=0.5 PI	SVM RBF σ=0.5 Sh	SVM RBF σ=1.0 PI	SVM RBF σ=1.0 Sh	NN 2L PI	NN 3L PI	NN 4L PI	NN 2L Sh.	NN 3L Sh.	NN 4L Sh.
ects_1	1	1	1	1	44	44	30	30	1	30	44	8	44	30	24	30	24	30	1	1	1	3	1	1
ID	2	3	2	2	3	30	44	44	30	2	30	21	8	44	30	44	30	44	**30**	30	30	1	3	3
t_dropout	3	2	6	30	39	21	13	8	44	1	4	4	24	8	8	8	13	8	13	4	13	30	44	30
gender_2.0	4	4	13	44		8	8	7	13	44	3	9	13	4	13	21	8	21	7	13	2	4	4	13
npe_1	5	6	3	6		24	2	6	2	6	1	44	21	21	2	6	2	4	2	2	4	44	30	4
ent_1st_y	6	5	4	10		3	1	21	8	11	34	24	2	6	44	4	11	6	4	44	3	34	7	6
mld_missing	7	10	30	3		34	4	13	4	16	7	30	11	24	6	1	6	41	3	7	11	13	13	9
s_scholarship	8	7	10	4		19	3	1	11	8	26	7	4	13	4	41	24	1	8	8	6	21	11	11
gender_1.0	9	8	34	8		22	11	9	34	24	24	16	28	9	34	24	21	26	**24**	11	44	8	6	7
age1	10	9	8	13		31	24	4	6	18	29	26	39	1	11	9	34	7	6	3	8	6	21	8
hs_missing	11	13	44	7		20	6	16	28	21	43	41	6	7	28	13	4	16	11	6	21	9	34	21
score_e	12	48	7	34		43	42	41	17	7	36	6	14	11	3	2	3	9	**28**	24	7	16	15	34
dur_3.0	13	11	47	11		28	17	3	41	13	27	11	42	41	9	7	9	2	44	14	24	7	9	2
hs_o	14	14	17	21			28	17	3	9	40	13	15	34	17	26	28	11	26	9	26	41	16	15
hs_stem	15	26	20	9			7	24	42	26	25	1	43	3	7	14	7	14	17	16	9	26	2	14
dist_up_to_80	16	12	14	26			34	42	9	4		34	3	16	21	11	26	17	42	28	17	14	14	41
hsd_STEM	17	42	37	41			21	28	24	14		3	20	26	14	16	42	43	34	17	28	11	8	24
dist_between 81 and 160	18	41	11	30			39	11	45	28		28	17	2	42	15	14	46	14	15	14	28	41	44

HGBC End of 1st year	HGBC PI	HGBC Sh.	RF PI	RF Sh.	SVM linear PI	SVM linear Sh.	SVM poly d=3 PI	SVM poly d=3 Sh.	SVM poly d=8 PI	SVM poly d=8 Sh.	SVM sigmoid PI	SVM sigmoid	SVM RBF σ=0.1 PI	SVM RBF σ=0.1 Sh.	SVM RBF σ=0.5 PI	SVM RBF σ=0.5 Sh.	SVM RBF σ=1.0 PI	SVM RBF σ=1.0 Sh.	NN 2L PI	NN 3L PI	NN 4L PI	NN 2L Sh.	NN 3L Sh.	NN 4L Sh.
t_acknowledged from_a_f	19	29	15	16			41	26	10	3		18	22	28	39	34	43	18	15	26	41	2	28	29
s_part-time	20	20		28			14	14	15	17		17	34	14	1	43	17	24	9	41	16	15	24	26

The rest of variables sorted by PI importance for HGBC model are: 21) s_co-financing, 22) t_passive_year, 23) dist_more_than_241, 24) ECTS_between21_and_40, 25) hsd_Medicine, 26) dist_0, 27) dist_between_161_and_240, 28) hsd_Economics, 29) hs_gym, 30) ECTS_less_than_20, 31) s_foreigner, 32) s_self-financing, 33) t_enroll_from_a_sp, 34) t_normal, 35) hsd_Art, 36) hsd_Lower_vocation, 37) hsd_Service, 38) ECTS_0, 39) ECTS_more_than_60, 40) hs_medic, 41) mld_1, 42) hsd_Gymnasium, 43) mld_2, 44) ECTS_between41, 45) hs_econ, 46) mld_4, 47) mld_3, 48) score_t.

Table A 27 – HGBC feature importance for pre-enrollment data set at imbalanced and balanced data

IMBALANCED						BALANCED					
SHAP feat. no.	**Pre-enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Pre-enrollment**	**PI values**	**SHAP feat. no.**	**Pre-enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Pre-enrollment**	**PI values**
0	ent_1st_y	0.31515	25	hsd_STEM	0.02115	**0**	ent_1st_y	0.33693	**9**	gender_2.0	0.07155
9	gender_2.0	0.20155	9	gender_2.0	0.01484	**1**	age1	0.21394	**8**	gender_1.0	0.06268
1	age1	0.17321	4	hs_stem	0.01238	**9**	gender_2.0	0.21327	**6**	hs_missing	0.04914
6	hs_missing	0.11458	8	gender_1.0	0.01166	**6**	hs_missing	0.13451	**1**	age1	0.03228
4	hs_stem	0.08884	0	ent_1st_y	0.00887	**21**	hsd_Economics	0.12646	**22**	hsd_Gymnasium	0.03050
8	gender_1.0	0.07624	1	age1	0.00703	**8**	gender_1.0	0.09113	**0**	ent_1st_y	0.02428
21	hsd_Economics	0.07489	6	hs_missing	0.00655	**19**	dist_up to 80	0.07950	**25**	hsd_STEM	0.02048
22	hsd_Gymnasium	0.06858	22	hsd_Gymnasium	0.00385	**22**	hsd_Gymnasium	0.07931	**4**	hs_stem	0.01455

IMBALANCED						BALANCED					
SHAP feat. no.	**Pre-enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Pre-enrollment**	**PI values**	**SHAP feat. no.**	**Pre-enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Pre-enrollment**	**PI values**
19	dist_up to 80	0.06111	21	hsd_Economics	0.00284	**4**	hs_stem	0.07369	**21**	hsd_Economics	0.01007
14	mld_missing	0.05934	26	hsd_Service	0.00116	**14**	mld_missing	0.06726	**15**	dist_0	0.00665
2	hs_gym	0.05536	7	hs_o	0.00063	**25**	hsd_STEM	0.03957	**7**	hs_o	0.00318
25	hsd_STEM	0.04019	15	dist_0	0.00063	**15**	dist_0	0.03739	**2**	hs_gym	0.00279
15	dist_0	0.03675	19	dist_up to 80	0.00043	**2**	hs_gym	0.03469	**24**	hsd_Medicine	0.00251
10	mld_1	0.01872	14	mld_missing	0.00043	**3**	hs_econ	0.03132	**26**	hsd_Service	0.00212
26	hsd_Service	0.01589	24	hsd_Medicine	0.00043	**10**	mld_1	0.03023	**19**	dist_up to 80	0.00183
3	hs_econ	0.01419	13	mld_4	0.00039	**26**	hsd_Service	0.01855	**14**	mld_missing	0.00169
24	hsd_Medicine	0.00836	2	hs_gym	0.00034	**24**	hsd_Medicine	0.01692	**18**	dist_more than 241	0.00125
13	mld_4	0.00807	3	hs_econ	0.00029	**13**	mld_4	0.01451	**11**	mld_2	0.00019
7	hs_o	0.00766	5	hs_medic	0.00024	**12**	mld_3	0.01429	**5**	hs_medic	0.00000
18	dist_more than 241	0.00638	10	mld_1	0.00014	**7**	hs_o	0.01314	**23**	hsd_Lower vocation	-0.00005
12	mld_3	0.00477	18	dist_more than 241	0.00014	**11**	mld_2	0.01171	**20**	hsd_Art	-0.00010
11	mld_2	0.00457	12	mld_3	0.00005	**16**	dist_between 161 and 240	0.01150	**3**	hs_econ	-0.00014
17	dist_between 81 and 160	0.00354	20	hsd_Art	0	**17**	dist_between 81 and 160	0.01085	**13**	mld_4	-0.00106
16	dist_between 161 and 240	0.00209	11	mld_2	0	**18**	dist_more than 241	0.01048	**17**	dist_between 81 and 160	-0.00260
5	hs_medic	0.00117	23	hsd_Lower vocation	0	**5**	hs_medic	0.00380	**12**	mld_3	-0.00275
20	hsd_Art	0.00000	17	dist_between 81 and 160	-0.00005	**23**	hsd_Lower vocation	0.00013	**16**	dist_between 161 and 240	-0.00308
23	hsd_Lower vocation	0.00000	16	dist_between 161 and 240	-0.00005	**20**	hsd_Art	0.00007	**10**	mld_1	-0.00487

Table A 28 - HGBC feature importance for Enrollment data set at imbalanced and balanced data

IMBALANCED						BALANCED					
SHAP feat. no	**Enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Enrollment**	**PI values**	**SHAP feat. no.**	**Enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Enrollment**	**PI values**
19	t_normal	0.8669	**19**	t_normal	0.09762	**19**	t_normal	0.82212	**1**	ID	0.08779
1	ID	0.6253	**1**	ID	0.03262	**1**	ID	0.64610	**19**	t_normal	0.08003
0	ent_1st_y	0.1760	**18**	t_acknowledged from_a_f	0.00756	**0**	ent_1st_y	0.20126	**27**	dist_0	0.00689
4	age1	0.1448	**0**	ent_1st_y	0.00511	**4**	age1	0.14353	**18**	t_acknowledged from_a_f	0.00602
16	s_scholarship	0.0992	**27**	dist_0	0.00222	**16**	s_scholarship	0.10679	**34**	hsd_Gymnasium	0.00477
26	mld_missing	0.0923	**26**	mld_missing	0.00116	**26**	mld_missing	0.07313	**0**	ent_1st_y	0.00463
12	gender_2.0	0.0910	**34**	hsd_Gymnasium	0.00082	**27**	dist_0	0.06119	**33**	hsd_Economics	0.00405
27	dist_0	0.0626	**21**	dur_3.0	0.00072	**3**	score_e	0.06104	**4**	age1	0.00294
9	hs_missing	0.0564	**7**	hs_stem	0.00063	**11**	gender_1.0	0.05922	**11**	gender_1.0	0.00275
34	hsd_Gymnasium	0.0533	**37**	hsd_STEM	0.00043	**2**	score_t	0.05548	**21**	dur_3.0	0.00246
2	score_t	0.0439	**20**	t_passive_year	0.00029	**13**	s_co-financing	0.05500	**7**	hs_stem	0.00246
3	score_e	0.0431	**22**	mld_1	0.00029	**34**	hsd_Gymnasium	0.05328	**37**	hsd_STEM	0.00226
7	hs_stem	0.0368	**29**	dist_between_81 and 160	0.00014	**9**	hs_missing	0.04395	**29**	dist_between 81 and 160	0.00130
18	t_acknowledged from_a_f	0.0356	**25**	mld_4	0.00010	**12**	gender_2.0	0.04053	**10**	hs_o	0.00130
13	s_co-financing	0.0317	**36**	hsd_Medicine	0.00010	**7**	hs_stem	0.03850	**12**	gender_2.0	0.00116
11	gender_1.0	0.0313	**24**	mld_3	0.00010	**18**	t_acknowledged from_a_f	0.03560	**16**	s_scholarship	0.00092
31	dist_up to 80	0.0298	**33**	hsd_Economics	0.00010	**10**	hs_o	0.03432	**31**	dist_up to 80	0.00077
21	dur_3.0	0.0219	**14**	s_foreigner	0.00005	**33**	hsd_Economics	0.03262	**30**	dist_more than 241	0.00034
33	hsd_Economics	0.0205	**15**	s_part-time	0.00005	**37**	hsd_STEM	0.02698	**8**	hs_medic	0.00029
23	mld_2	0.0186	**8**	hs_medic	0.00000	**21**	dur_3.0	0.02433	**5**	hs_gym	0.00024
37	hsd_STEM	0.0182	**17**	s_self-financing	0.00000	**22**	mld_1	0.02151	**22**	mld_1	0.00019
10	hs_o	0.0170	**35**	hsd_Lower vocation	0.00000	**31**	dist_up to 80	0.01515	**38**	hsd_Service	0.00014

IMBALANCED						BALANCED					
SHAP feat. no	**Enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Enrollment**	**PI values**	**SHAP feat. no.**	**Enrollment**	**SHAP feat. imp. values**	**PI feat. No.**	**Enrollment**	**PI values**
22	mld_1	0.0115	**32**	hsd_Art	0.00000	**23**	mld_2	0.01180	**32**	hsd_Art	0.00005
5	hs_gym	0.0090	**10**	hs_o	0.00000	**36**	hsd_Medicine	0.00975	**20**	t_passive_year	0.00000
6	hs_econ	0.0081	**28**	dist_between 161 and 240	-0.00005	**30**	dist_more than 241	0.00913	**17**	s_self-financing	0.00000
36	hsd_Medicine	0.0065	**6**	hs_econ	-0.00005	**8**	hs_medic	0.00866	**35**	hsd_Lower vocation	0.00000
15	s_part-time	0.0054	**30**	dist_more than 241	-0.00019	**15**	s_part-time	0.00842	**14**	s_foreigner	0.00000
30	dist_more than 241	0.0040	**5**	hs_gym	-0.00019	**6**	hs_econ	0.00779	**9**	hs_missing	-0.00024
20	t_passive_year	0.0040	**38**	hsd_Service	-0.00029	**5**	hs_gym	0.00690	**24**	mld_3	-0.00029
29	dist_between 81 and 160	0.0039	**23**	mld_2	-0.00034	**29**	dist_between 81 and 160	0.00565	**15**	s_part-time	-0.00034
24	mld_3	0.0037	**2**	score_t	-0.00048	**20**	t_passive_year	0.00391	**25**	mld_4	-0.00039
38	hsd_Service	0.0033	**12**	gender_2.0	-0.00053	**38**	hsd_Service	0.00383	**36**	hsd_Medicine	-0.00058
28	dist_between 161 and 240	0.0020	**31**	dist_up to 80	-0.00082	**28**	dist_between 161 and 240	0.00380	**26**	mld_missing	-0.00063
25	mld_4	0.0012	**13**	s_co-financing	-0.00082	**24**	mld_3	0.00219	**28**	dist_between 161 and 240	-0.00087
8	hs_medic	0.0007	**3**	score_e	-0.00145	**25**	mld_4	0.00197	**6**	hs_econ	-0.00087
14	s_foreigner	0.0004	**11**	gender_1.0	-0.00159	**32**	hsd_Art	0.00014	**23**	mld_2	-0.00106
32	hsd_Art	0.0000	**16**	s_scholarship	-0.00164	**17**	s_self-financing	0.00000	**3**	score_e	-0.00226
35	hsd_Lower vocation	0.0000	**9**	hs_missing	-0.00164	**14**	s_foreigner	0.00000	**13**	s_co-financing	-0.00352
17	s_self-financing	0.0000	**4**	age1	-0.00202	**35**	hsd_Lower vocation	0.00000	**2**	score_t	-0.00487

Source: Author

Table A 29 - HGBC feature importance for end of year data set at imbalanced and balanced data

IMBALANCED						BALANCED					
SHAP feat. no	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**	**SHAP feat. no**	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**
21	t_dropout	1.09779	**21**	t_dropout	0.08576	**21**	t_dropout	1.08742	**1**	ID	0.10826
4	ects_1	0.68692	**1**	ID	0.05623	**4**	ects_1	0.73484	**21**	t_dropout	0.07025
1	ID	0.55720	**4**	ects_1	0.01845	**1**	ID	0.55863	**4**	ects_1	0.05406
5	npe_1	0.25361	**5**	npe_1	0.01845	**5**	npe_1	0.22333	**5**	npe_1	0.01821
0	ent_1st_y	0.13622	**0**	ent_1st_y	0.00443	**0**	ent_1st_y	0.15461	**30**	**mld_missing**	0.01392
6	age1	0.10244	**30**	**mld_missing**	0.00385	**6**	age1	0.14413	**25**	dur_3.0	0.00573
30	mld_missing	0.09655	**6**	age1	0.00140	**30**	mld_missing	0.12291	**0**	ent_1st_y	0.00554
11	hs_missing	0.07378	**41**	hsd_STEM	0.00130	**11**	hs_missing	0.09113	**6**	age1	0.00275
18	s_scholarship	0.04107	**3**	score_e	0.00130	**3**	score_e	0.05499	**31**	**dist_0**	0.00275
31	dist_0	0.03981	**31**	**dist_0**	0.00106	**2**	score_t	0.04695	**41**	hsd_STEM	0.00236
14	gender_2.0	0.03907	**18**	s_scholarship	0.00087	**18**	s_scholarship	0.04286	**13**	gender_1.0	0.00212
2	score_t	0.03068	**25**	dur_3.0	0.00082	**25**	dur_3.0	0.04123	**11**	**hs_missing**	0.00154
3	score_e	0.02694	**43**	ECTS_ between 21 and 40	0.00067	**14**	gender_2.0	0.04042	**14**	gender_2.0	0.00120
25	dur_3.0	0.02242	**2**	score_t	0.00067	**31**	dist_0	0.03679	**12**	hs_o	0.00072
26	mld_1	0.02107	**12**	hs_o	0.00058	**13**	gender_1.0	0.02751	**40**	hsd_Medicine	0.00058
9	hs_stem	0.02068	**9**	hs_stem	0.00053	**12**	hs_o	0.02434	**36**	hsd_Art	0.00048
12	hs_o	0.01704	**11**	hs_missing	0.00048	**37**	hsd_Economics	0.02316	**23**	t_normal	0.00043
38	hsd_Gymnasium	0.01702	**13**	gender_1.0	0.00043	**17**	s_part-time	0.02079	**28**	mld_3	0.00043
13	gender_1.0	0.01532	**27**	mld_2	0.00039	**9**	hs_stem	0.01934	**17**	s_part-time	0.00039
41	hsd_STEM	0.01415	**15**	s_co-financing	0.00034	**15**	s_co-financing	0.01870	**9**	hs_stem	0.00039
37	hsd_Economics	0.01335	**14**	gender_2.0	0.00034	**41**	hsd_STEM	0.01576	**37**	hsd_Economics	0.00024
17	s_part-time	0.01227	**26**	mld_1	0.00024	**38**	hsd_Gymnasium	0.01568	**8**	hs_econ	0.00010
7	hs_gym	0.01185	**28**	mld_3	0.00024	**26**	mld_1	0.01365	**34**	dist_more than 241	0.00010
27	mld_2	0.00840	**17**	s_part-time	0.00019	**35**	dist_up to 80	0.01339	**42**	hsd_Service	0.00010
15	s_co-financing	0.00558	**7**	hs_gym	0.00019	**27**	mld_2	0.01175	**39**	hsd_Lower vocation	0.00000
33	dist_between 81 and 160	0.00407	**42**	hsd_Service	0.00014	**40**	hsd_Medicine	0.01008	**45**	ECTS_between 41 and 60	0.00000

IMBALANCED						BALANCED					
SHAP feat. no	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**	**SHAP feat. no**	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**
40	hsd_Medicine	0.00395	**8**	hs_econ	0.00010	**34**	dist_more than 241	0.00881	**46**	ECTS_less than 20	0.00000
35	dist_up to 80	0.00353	**39**	hsd_Lower vocation	0.00000	**28**	mld_3	0.00837	**44**	ECTS_0	0.00000
8	hs_econ	0.00285	**45**	ECTS_between 41 and 60	0.00000	**7**	hs_gym	0.00710	**24**	t_passive_year	0.00000
28	mld_3	0.00264	**44**	ECTS_0	0.00000	**8**	hs_econ	0.00699	**47**	ECTS_more than 60	0.00000
43	ECTS_ between 21 and 40	0.00226	**35**	dist_up to 80	0.00000	**45**	ECTS_between 41 and 60	0.00697	**22**	t_enroll_from_a_sp	0.00000
42	hsd_Service	0.00178	**46**	ECTS_less than 20	0.00000	**23**	t_normal	0.00691	**10**	hs_medic	0.00000
32	dist_between 161 and 240	0.00069	**36**	hsd_Art	0.00000	**32**	dist_between 161 and 240	0.00610	**19**	s_self-financing	0.00000
20	t_acknowledged _from_a_f	0.00060	**24**	t_passive_year	0.00000	**33**	dist_between 81 and 160	0.00486	**20**	t_acknowledged_from_ a_f	-0.00005
29	mld_4	0.00043	**23**	t_normal	0.00000	**42**	hsd_Service	0.00405	**43**	ECTS_ between 21 and 40	-0.00019
34	dist_more than 241	0.00041	**22**	t_enroll_from_a_sp	0.00000	**29**	mld_4	0.00280	**16**	s_foreigner	-0.00024
16	s_foreigner	0.00040	**20**	t_acknowledged_fro m_a_f	0.00000	**43**	ECTS_ between 21 and 40	0.00167	**33**	dist_between 81 and 160	-0.00024
24	t_passive_year	0.00007	**19**	s_self-financing	0.00000	**46**	ECTS_less than 20	0.00143	**32**	dist_between 161 and 240	-0.00039
36	hsd_Art	0.00000	**16**	s_foreigner	0.00000	**20**	t_acknowledged_fro m_a_f	0.00082	**15**	s_co-financing	-0.00063
39	hsd_Lower vocation	0.00000	**10**	hs_medic	0.00000	**36**	hsd_Art	0.00051	**29**	mld_4	-0.00082
23	t_normal	0.00000	**47**	ECTS_more than 60	0.00000	**10**	hs_medic	0.00024	**38**	hsd_Gymnasium	-0.00087
22	t_enroll_from_a _sp	0.00000	**38**	hsd_Gymnasium	-0.00005	**16**	s_foreigner	0.00014	**26**	mld_1	-0.00092
19	s_self-financing	0.00000	**32**	dist_between 161 and 240	-0.00005	**24**	t_passive_year	0.00007	**27**	mld_2	-0.00096
10	hs_medic	0.00000	**29**	mld_4	-0.00005	**39**	hsd_Lower vocation	0.00000	**7**	hs_gym	-0.00096

IMBALANCED						BALANCED					
SHAP feat. no	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**	**SHAP feat. no**	**End of year**	**SHAP feat. imp. values**	**PI feat. no**	**End of year**	**PI feat. imp. values**
44	ECTS_0	0.00000	**34**	dist_more than 241	-0.00010	**44**	ECTS_0	0.00000	**35**	dist_up to 80	-0.00116
45	ECTS_between 41 and 60	0.00000	**37**	hsd_Economics	-0.00014	**22**	t_enroll_from_a_sp	0.00000	**2**	score_t	-0.00178
46	ECTS_less than 20	0.00000	**40**	hsd_Medicine	-0.00014	**19**	s_self-financing	0.00000	**18**	s_scholarship	-0.00198
47	ECTS_more than 60	0.00000	**33**	dist_between 81 and 160	-0.00019	**47**	ECTS_more than 60	0.00000	**3**	score_e	-0.00400

Source: Author

Table A 30 – Summary of SVM model, kernel: polynomial, degree = 3.

SVM, kernel: polynomial degree = 3	**Pre enroll**	**Enroll**	**End of 1st year**	**End of 1st year (top N)**
Accuracy	0.61342	0.67837	0.76570	0.76546
True Negative	1385	1580	1606	1599
False Positive	706	511	485	492
False Negative	901	826	489	483
True Positive	1165	1240	1577	1583
True Negative	33.32%	38.01%	38.63%	38.47%
False Positive	16.98%	12.29%	11.67%	11.84%
False Negative	21.67%	19.87%	11.76%	11.62%
True Positive	28.03%	29.83%	37.94%	38.08%
F1	0.59182	0.64972	0.76405	0.76455
Precision	0.62266	0.70817	0.76479	0.76289
Recall	0.56389	0.60019	0.76331	0.76621
ROCAUC	0.61313	0.67791	0.76568	0.76546
Matthews corr.	0.22739	0.36031	0.53137	0.53091
PR score	0.64681	0.76122	0.84067	0.84436
Specificity	0.66236	0.75562	0.76805	0.76471
Train accuracy	0.63579	0.69871	0.77907	0.76740

Source: Author.

Table A 31 – Summary of SVM model, kernel: polynomial, degree =8.

SVM, kernel: polynomial degree = 8	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)
Accuracy	0.61823	0.67741	0.75487	0.75848
True Negative	1334	1512	1613	1552
False Positive	757	579	478	539
False Negative	830	762	541	465
True Positive	1236	1304	1525	1601
True Negative	32.09%	36.37%	38.80%	37.33%
False Positive	18.21%	13.93%	11.50%	12.97%
False Negative	19.97%	18.33%	13.01%	11.19%
True Positive	29.73%	31.37%	36.69%	38.51%
F1	0.60902	0.66042	0.74957	0.76129
Precision	0.62017	0.69251	0.76136	0.74813
Recall	0.59826	0.63117	0.73814	0.77493
ROCAUC	0.61811	0.67714	0.75477	0.75858
Matthews corr.	0.23643	0.35584	0.50987	0.51737
PR score	0.63427	0.71741	0.81590	0.83832
Specificity	0.63797	0.72310	0.77140	0.74223
Train accuracy	0.65762	0.74791	0.82364	0.78454

Source: Author.

Table A 32 – Summary of SVM model, kernel: sigmoid.

SVM kernel: sigmoid	**Pre enroll**	**Enroll**	**End of 1st year**	**End of 1st year (top N)**
Accuracy	0.50806	0.53885	0.63219	0.60933
True Negative	1074	1134	1350	1274
False Positive	1017	957	741	817
False Negative	1028	960	788	807
True Positive	1038	1106	1278	1259
True Negative	25.84%	27.28%	32.48%	30.65%
False Positive	24.46%	23.02%	17.83%	19.65%
False Negative	24.73%	23.09%	18.96%	19.41%
True Positive	24.97%	26.61%	30.74%	30.29%
F1	0.50376	0.53572	0.62570	0.60792
Precision	0.50511	0.53611	0.63299	0.60645
Recall	0.50242	0.53533	0.61859	0.60939
ROCAUC	0.50802	0.53883	0.63211	0.60933
Matthews corr.	0.01605	0.07766	0.26431	0.21866
PR score	0.53692	0.58862	0.65743	0.57627
Specificity	0.51363	0.54232	0.64562	0.60928
Train accuracy	0.53221	0.55176	0.63260	0.61131

Source: Author.

Table A 33 – Summary of SVM model, kernel: RBF, gamma = 0.1.

SVM kernel: RBF gamma=0.1	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)
Accuracy	0.59851	0.66250	0.73899	0.71975
True Negative	1388	1618	1397	1198
False Positive	703	473	694	893
False Negative	966	930	391	272
True Positive	1100	1136	1675	1794
True Negative	33.39%	38.92%	33.61%	28.82%
False Positive	16.91%	11.38%	16.69%	21.48%
False Negative	23.24%	22.37%	9.41%	6.54%
True Positive	26.46%	27.33%	40.29%	43.16%
F1	0.56862	0.61823	0.75536	0.75489
Precision	0.61009	0.70603	0.70705	0.66766
Recall	0.53243	0.54985	0.81075	0.86834
ROCAUC	0.59811	0.66182	0.73942	0.72064
Matthews corr.	0.19797	0.33223	0.48358	0.46149
PR score	0.64767	0.74966	0.83381	0.76387
Specificity	0.66380	0.77379	0.66810	0.57293
Train accuracy	0.60138	0.66869	0.74538	0.71922

Source: Author.

Table A 34 – Summary of SVM model, kernel: RBF, gamma = 0.5.

SVM Kernel: RBF gamma=0.5	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)
Accuracy	0.60476	0.67380	0.75559	0.75391
True Negative	1356	1546	1513	1509
False Positive	735	545	578	582
False Negative	908	811	438	441
True Positive	1158	1255	1628	1625
True Negative	32.62%	37.19%	36.40%	36.30%
False Positive	17.68%	13.11%	13.90%	14.00%
False Negative	21.84%	19.51%	10.54%	10.61%
True Positive	27.86%	30.19%	39.16%	39.09%
F1	0.58500	0.64925	0.76217	0.76059
Precision	0.61173	0.69722	0.73799	0.73629
Recall	0.56050	0.60745	0.78800	0.78654
ROCAUC	0.60450	0.67341	0.75579	0.75410
Matthews corr.	0.20983	0.34996	0.51253	0.50917
PR score	0.66011	0.75350	0.83075	0.82751
Specificity	0.64849	0.73936	0.72358	0.72166
Train accuracy	0.62045	0.68577	0.76481	0.75771

Source: Author.

Table A 35 – Summary of SVM model, kernel: RBF, gamma = 1.0.

SVM Kernel: RBF gamma=1.0	Pre enroll	Enroll	End of 1st year	End of 1st year (top N)
Accuracy	0.62136	0.66947	0.75439	0.75391
True Negative	1307	1479	1537	1569
False Positive	784	612	554	522
False Negative	790	762	467	501
True Positive	1276	1304	1599	1565
True Negative	31.44%	35.58%	36.97%	37.74%
False Positive	18.86%	14.72%	13.33%	12.56%
False Negative	19.00%	18.33%	11.23%	12.05%
True Positive	30.70%	31.37%	38.47%	37.65%
F1	0.61852	0.65495	0.75800	0.75367
Precision	0.61942	0.68058	0.74268	0.74988
Recall	0.61762	0.63117	0.77396	0.75750
ROCAUC	0.62134	0.66924	0.75451	0.75393
Matthews corr.	0.24268	0.33952	0.50933	0.50786
PR score	0.65618	0.74144	0.80240	0.80037
Specificity	0.62506	0.70732	0.73505	0.75036
Train accuracy	0.63471	0.69720	0.78045	0.75952

Source: Author.

Table A 36 – Summary of five iterations and their average of NN 2 layers, pre-enroll data

NN 2 LAYERS	NN (1st)	NN (2nd)	NN (3rd)	NN (4th)	NN (5th)	NN (avg)
Accuracy	0.62181	0.63024	0.62494	0.62181	0.63048	0.62585
True Negative	1485	1417	1282	1355	1464	1401
False Positive	570	638	773	700	591	654
False Negative	1001	898	785	871	944	900
True Positive	1098	1201	1314	1228	1155	1199
True Negative	35.75%	34.11%	30.86%	32.62%	35.24%	33.72%
False Positive	13.72%	15.36%	18.61%	16.85%	14.23%	15.75%
False Negative	24.10%	21.62%	18.90%	20.97%	22.73%	21.66%
True Positive	26.43%	28.91%	31.63%	29.56%	27.80%	28.87%
F1	0.58296	0.60995	0.62781	0.60988	0.60078	0.60628
Precision	0.65827	0.65307	0.62961	0.63693	0.66151	0.64788
Recall	0.52311	0.57218	0.62601	0.58504	0.55026	0.57132
ROCAUC	0.62287	0.63086	0.62493	0.62220	0.63134	0.62644
Matthews corr.	0.25063	0.26344	0.24985	0.24503	0.26606	0.25500
PR score	0.69268	0.69620	0.69121	0.69121	0.69548	0.69336
Specificity	0.72263	0.68954	0.62384	0.65937	0.71241	0.68156
Train accuracy	0.63031	0.64097	0.64446	0.64500	0.64927	0.64200

Source: Author.

Table A 37 - Summary of five iterations and their average of NN 2 layers, enroll data

2 LAYERS	NN (1st)	NN (2nd)	NN (3rd)	NN (4th)	NN (5th)	NN (avg)
Accuracy	0.67935	0.67718	0.68175	0.68344	0.69066	0.68247
True Negative	1637	1449	1460	1345	1514	1481
False Positive	418	606	595	710	541	574
False Negative	914	735	727	605	744	745
True Positive	1185	1364	1372	1494	1355	1354
True Negative	39.41%	34.88%	35.15%	32.38%	36.45%	35.65%
False Positive	10.06%	14.59%	14.32%	17.09%	13.02%	13.82%
False Negative	22.00%	17.69%	17.50%	14.56%	17.91%	17.93%
True Positive	28.53%	32.84%	33.03%	35.97%	32.62%	32.60%
F1	0.64019	0.67043	0.67486	0.69440	0.67835	0.67165
Precision	0.73924	0.69239	0.69751	0.67786	0.71466	0.70433
Recall	0.56455	0.64983	0.65364	0.71177	0.64555	0.64507
ROCAUC	0.68057	0.67747	0.68205	0.68313	0.69114	0.68288
Matthews corr.	0.37092	0.35539	0.36460	0.36693	0.38372	0.36831
PR score	0.77719	0.78489	0.78525	0.78624	0.78875	0.78446
Specificity	0.79659	0.70511	0.71046	0.65450	0.73674	0.72068
Train accuracy	0.69550	0.70513	0.71951	0.71885	0.72842	0.71348

Source: Author.