From Black Box to Transparent Academic Support

Daniela Valente

Dissertation written under the supervision of Professor Ana Guedes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, January 2024

Abstract

The thesis explores the evolution of education in Portugal, from high illiteracy rates to alignment with European averages in access to higher education. Despite progress, persistent school dropout remains a challenge. Using machine learning models emphasising explainability, the aim is to predict academic performance and identify dropout risks, highlighting the importance of transparency to build confidence in making practical decisions, such as preventative support strategies. The study uses public data from a Portuguese university, exploring demographic, socioeconomic, macroeconomic and academic factors. Two models were developed, Model A (after one academic year) and Model B (at enrolment), using the CatBoost algorithm. The results indicated substantially better performance for Model A, but both face challenges in the confusion matrix, with more false positives than false negatives. Predicting a false positive is more costly than predicting a false negative, according to the aim of the analysis. To solve this problem, an individualised analysis adapted to each model is suggested. The interpretability technique results highlight that after one year, first year grades have a significant impact on student performance, while at the time of enrolment, age, holding a scholarship and gender also emerged as influential factors. The significance of this analysis aims to formulate proactive strategies and personalised support systems to mitigate dropout risks and increase success in Portuguese higher education.

Keywords: Machine Learning, Explainability, Education, CatBoost, Academic Support.

Title: From Black Box to Transparent Academic Support.

Author: Daniela Valente.

Resumo

A tese aborda a evolução educativa em Portugal, desde elevadas taxas de analfabetismo até à convergência com as médias europeias no acesso ao ensino superior. Apesar dos progressos, o persistente abandono escolar ainda é um desafio. Utilizando modelos de aprendizagem automática com foco em explicabilidade, o objetivo é prever o desempenho académico e identificar riscos de abandono escolar, destacando a importância da transparência para inspirar confiança na tomada de decisões práticas, como estratégias de apoio preventivas. O estudo utiliza dados públicos de uma instituição universitária portuguesa, explorando fatores demográficos, socioeconómicos, macroeconómicos e académicos. Dois modelos, Modelo A (após um ano curricular) e Modelo B (momento da matrícula), foram desenvolvidos recorrendo ao algoritmo CatBoost. Os resultados indicam um desempenho substancialmente melhor para o Modelo A, mas ambos mostram desafios na matriz de confusão, com mais falsos positivos do que falsos negativos. Prever um falso positivo apresenta um maior custo que prever um falso negativo, de acordo com o objetivo da análise. Para mitigar isso, sugere-se uma análise individual ajustada a cada modelo. Os resultados da técnica de interpretabilidade destacam que após um ano as notas do primeiro ano impactam significativamente o desempenho dos alunos, enquanto, no momento da matrícula, a idade, a posse de uma bolsa e o género também surgiram como fatores influentes. A importância desta análise visa desenvolver estratégias proativas e sistemas de apoio personalizados para reduzir riscos de abandono e melhorar o sucesso no ensino superior português.

Palavras-Chave: Aprendizagem Automática, Explicabilidade, Educação, CatBoost, Apoio Académico.

Título: From Black Box to Transparent Academic Support.

Autor: Daniela Valente.

Contents

List of Figures

List of Tables

Introduction

Portugal was characterised by massive illiteracy rates until the 1970s, with only a minority of the population achieving education and qualifications (DGEEC and Ministério da Educação 2021). The illiteracy rate has been higher among females since 1970 until 2011, and there has been a decrease in the illiteracy rate in Portugal from 25.7% in 1970 to 5.2% in 2011, and a higher decrease in 2021 to 3.1%, according to the last INE census (INE 2021; DGEEC and Ministério da Educação 2021). However, the development of the education system, in recent decades, has led to a very fast evolution in education and qualification rates. This is reflected in levels of higher education graduates, which are already close to the European averages among the younger population: in 2019, Portugal registered 25% of the population with higher education, while the European Union (EU) registered 30.7%; and taking into account secondary education, Portugal registered 24.8% and the EU 47.1%. Several policies have contributed to this development, such as a gradual expansion of mandatory education (12 years); and expansion and diversification of education and training offerings in secondary and higher education, for example.

Despite the improvement in education in Portugal, school dropout and educational failure are problems that directly affect and impact higher education students, the higher education system and, consequently, society, since in the perspective of an economy that is increasingly supported by knowledge and innovation, the importance of higher education qualifications is emphasised (Lopes, Pereira, and Vaz 2023). Higher education institutions have the challenge of not only attracting more students but also dealing effectively with their very different academic performances (Delen 2011; Hoffait and Schyns 2017).

Within this challenging scenario, institutions must analyse and understand the reasons that lead students to drop out of their programmes, as this is essential for the further implementation of strategies to prevent dropout, which can be everyone's responsibility, but with a greater focus on the institutions that receive them and the nature of educational policies. It is important to determine the factors associated with dropout rates, characterise the profile of students and their motivations for dropping out and, at the same time, define guidelines so that institutions can develop preventive strategies (Lopes, Pereira, and Vaz 2023; Esteves et al. 2021). To this end, higher education institutions are recognising the potential of studying educational data to improve the quality of their management decisions. According to Delen (2011), more than 50% of student attrition can be attributed to the first year of college, also known as the freshman year. For this reason, it is essential to identify vulnerable students who are likely to drop out of

school in their first year. Identifying at-risk students can allow institutions to move better and faster to achieve their retention management goals. The use of new generation techniques, such as Artificial Intelligence (AI) and in particular machine learning (ML) and data mining (DM), is clearly on the agenda for higher education institutions in order to support the development of educational strategies. Data mining is the process of extracting valuable insights (i.e. non-trivial, logical, previously unknown, and potentially useful patterns) from a large amount of data (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Although AI systems are a relatively recent concept, they have been integrated into all aspects of our lives and have been successfully applied to complex problems in areas such as medicine, healthcare, homeland security, transport, finance, marketing, and entertainment (Mehrabi et al. 2019).

In this project, I apply data mining and machine learning techniques to the higher education system, specifically to the problem of student dropout. In machine learning and especially in the context of education, performance on its own is not enough and a single metric such as classification provides an incomplete description of the phenomenon in question (Doshi-Velez and Kim 2017). Interpretability and explicability play a vital role in understanding model predictions, promoting transparency, upholding fairness, ensuring accountability, and aligning AI systems with human values. This is critical for addressing ethical issues, improving decision-making processes, and supporting the practical applicability of AI in diverse real-world contexts. Improved interpretability in a machine learning model makes it more accessible for individuals to discern the logic behind their decisions or predictions. In essence, a model is considered more interpretable than another when its decision-making process is more understandable to humans.

The objective of this thesis is to predict student academic performance and the risk of students dropping out, employing machine learning models designed with interpretability and explicability in consideration. The ability to acquire knowledge about general trends in student performance is crucial for creating predictive models of student outcomes; however, it also allows me to explore certain situations, such as knowing why a particular student decides to drop out of a course; or more specific trends will also be important, such as: What factors have the greatest impact on dropping out of school?; Are students who enrol with lower admission grades more likely to drop out?; Are older students more likely to drop out?; Are students who are displaced at greater risk of dropping out?; Are students who have a lower academic performance in the first semester more likely to drop out?

In order to proactively address the difficulties that contribute to poor school performance, personalised support plans can be developed for each student as a result of this better understanding.

In the initial phases, followed by data mining, a supervised machine learning approach will be used to predict student performance based on the information about the students that has been acquired. An explainable machine learning technique will be used in the next phase to create a classification decision that is easier to comprehend. Although education is an improved topic nowadays, there is still plenty of opportunity for improvement, which drives many studies. This thesis is concerned with the following research question:

To what extent can interpretability and explainability techniques be integrated into machine learning models to enhance the accuracy and reliability of predictions for students' academic outcomes, such as dropout risk in higher education institutions?

Literature Review

Evolution of Education and Current Situation in Portugal

In terms of higher education, Portugal has seen significant development, aligning closely with its European counterparts. This achievement is particularly remarkable given the substantial decline in the country's birth rate, resulting in fewer young people. In 2000, only 7.5% of the working-age population in Portugal had a higher education, while in 2019, this figure had risen to 23.8% (36.2% in the 30-34 age group, rapidly approaching the European average). With equal opportunity as a constitutional organizing principle, the education system has made remarkable progress in recent decades in ensuring access and success for all students. The early school dropout rate, which stood at over 40% at the beginning of the century, has reduced to 10%, now aligning with the European average (DGEEC and Ministério da Educação 2021). Notably, the dropout rate is higher in regions such as the Algarve, Alentejo, and the North, as opposed to Lisbon and the Centre of Portugal. Regarding higher education, the dropout rate for undergraduate students is 29%, according to a study by the DGEEC published in 2019 (U-World Blog 2021). Official data also shows that less than half of students complete their studies within the three-year duration of a degree program. Particularly, only 8% of students who enter with a perfect 20-point grade average leave their higher education program, while among those entering with a 10-point average, this figure rises to 54%. Similarly, students who are admitted to their first or second choice in the national admission competition have much higher completion rates (53% in the 1st choice and 50% in the 2nd choice) than those who enter as the 5th or 6th option, with completion rates of 42% and 38%, respectively (U-World Blog 2021).

The nature of the process of dropping out of higher education is largely unknown, as Tinto (1975; 2010) states, despite the abundance of literature on dropping out of higher education. The author notes that most of the studies carried out so far have limited themselves to presenting descriptions of how various institutional and individual characteristics are related to dropping out. This is why he says that dropout occurs when students fail to integrate at the social and academic levels (problems with academic performance and career development). Dropping out is caused by a series of relationships between the individual characteristics of the student, the formal and informal characteristics of the educational institution and the community in which they live (Tinto 1975; 2010).

Lopes, Pereira, and Vaz (2023) identify several key factors that influence student performance and the risk of dropping out of higher education, such as vocational issues, academic failure, employability difficulties, economic constraints, professional aspirations, and challenges in the

integration process. These factors often interact, creating a complex web of challenges for students. To address this issue, the authors suggest that higher education institutions can combat dropout rates by implementing volunteer programmes, which offer support, skills development, and a sense of belonging to increase student success and well-being. In addition, implementing preventative measures, such as academic support, career guidance and financial aid, can help mitigate the risk of dropping out and improve students' overall performance.

Kersanszki, Holik, and Sanda (2022) highlight several key factors that influence student performance and the risk of dropping out of higher education, covering economic, institutional, sociological, pedagogical, and psychological dimensions. According to the authors, the initial year, particularly the first semester, is considered a critical phase and success during this period substantially mitigates the risk of dropping out. As students begin their higher education journey, they are faced with greater independence and responsibility, physical separation from their families and the imperative need for effective time management. This transition involves adapting to new learning environments, creating social relationships, and striving for personal and social well-being. However, institutional responsibilities are also emphasised, since the quality of teaching, pedagogical methods, technical resources, and the personality of the trainer significantly shape students' academic commitment, and shortcomings in these aspects can contribute to attrition.

Related Work

In the field of applied machine learning to higher education, research is by no means a new task. Numerous researchers and academics have invested significant effort in exploring various approaches, models, and data. Much of this research revolves around anticipating and understanding student performance, with the aim of discerning whether a student will prosper in a university environment or choose to drop out. This knowledge has immense value when it comes to implementing support strategies and preventative measures to help students facing challenges along their academic journey. Several approaches have been used to address these challenges.

Martins et al. (2021) utilised data from the Institute of Portalegre, excluding pre-college academic information. The primary goal was to predict student outcomes classified into success, relative success, or failure based on the time taken to obtain a degree. The distribution of the classes was imbalanced, with two minority classes failure (28%) and relative success (16%) and a majority class (56%). Due to the class imbalance, the researchers explored data sampling techniques, where data augmentation, particularly through the synthetic minority

oversampling technique (SMOTE) and adaptive synthetic sampling (ADASYN), produced the best results. From the numerous ML algorithms tested, boosting based methods, such as Extreme Gradient Boosting and Random Forest had the best performance, reaching 73% and 72% accuracy and 65% and 62% F1 scores, respectively. To enhance results, they suggested incorporating academic performance data from students' initial semesters into the dataset.

Alhazmi, Sheneamer, and Sheneamer (2017) use the same dataset and explore a predictive model to predict student performance using student characteristics. The characteristics that most influence students' academic success are the admission grade, first-level courses, the academic achievement test (AAT) and the general aptitude test (GAT). The authors mention that they used several classifiers, and the results show that, when only admission grades are considered for training and testing, the best performance is achieved with the Random Forest algorithm with an accuracy of 54%. In addition, the combination of admission grades with different sets of features is explored, with the most optimal results being obtained when admission grades are combined with all first-level computer science grades. The paper also uses t-SNE dimensionality reduction to visualize high-dimensional data and evaluates various combinations of features in predicting student performance, demonstrating the effectiveness of specific combinations in improving prediction accuracy.

Beaulac and Rosenthal (2019) analysed an extensive dataset from a university in Canada and predicted academic success using the Random Forest algorithm. The study examined students' grades in the first few subjects, intending to predict whether a student would complete the course and, if so, what their specialization would be. The results showed an accuracy of 79% in predicting course completion and 47% in predicting the chosen specialization.

Hoffait and Schyns (2017) studied and analysed a dataset of 6,845 students from the University of Liège to identify students most likely to face difficulties when completing their first year of university. Different algorithms, including Random Forest, Logistic Regression and Artificial Neural Networks, were explored, with an overall accuracy of 70% and the researchers mention that it is possible to increase the accuracy of the predictions by adding “uncertain”, this means that individuals that cannot confidently be associated to the ‘failure’ group or the ‘success’ group, were set apart in a new ‘uncertain’ group. The researchers opted to use Random Forest to further improve the model's accuracy and were able to identify with a high rate of confidence, 90%, a subset of 12.2% of students facing a very high risk of failure. Finally, the authors have performed a “what if” sensitivity analysis to identify more precisely the profile of students facing difficulties and to determine some characteristics on which remediation actions could be built.

Miguéis et al. (2018) used data from a European engineering school to predict overall academic performance based on the information available at the end of the first year of the student's academic path. A variety of algorithms were applied, however, the results reveal that Random Forest is the one presenting the best results, performance of the models was assessed by measuring overall accuracy, sensitivity and precision per class. Concluding that the enrolment average grade and the enrolment exam average grade are the most relevant attributes for predicting overall academic performance. One of the main difficulties faced was class imbalance, as one of the categories was much smaller than the other. However, this resulted in high accuracy, above 95%.

In another study, Thammasiri et al. (2014), used different class balancing strategies and classification methods to predict academic dropout. With a comprehensive dataset comprising 21,654 entries, of which 78.7% were positive (retained) and 21.3% were negative (dropped out), the study reveals that the Support Vector Machine combined with the SMOTE achieved the highest rate of accuracy (90.2%) and specificity (95.8%). After a sensitivity analysis on trained prediction models, they concluded that the most important factors are GPA, Tuition Waiver Scholarship, Student Loan, Ethnicity and Major Declared.

Interpretability and Explainability in Education

Machine learning comprises a set of techniques that empower computers to make predictions or enhance behaviour based on data (Molnar 2023). In essence, a machine learning algorithm learns a model by estimating parameters (such as weights) or understanding structures (such as trees). Subsequently, this fully trained machine learning model can be applied to make predictions for new instances. It provides advantages in terms of speed, scalability, and consistent results. However, a notable drawback is the increasing complexity of models, making them challenging to interpret and understand the decision-making process. Despite this, the substantial performance benefits of machine learning make it an indispensable tool for addressing a diverse array of problems (Molnar 2023).

The interpretability and explicability of the model become increasingly important when dealing with large datasets and complex models. The derived analytics and the models will be acceptable and trusted if they are interpretable.

Interpretability is the degree to which a human can understand the cause of a decision (Miller 2017). Explainability in machine learning means that you can explain what happens in your model from input to output, it makes models transparent and solves the black box (Onose 2023).

For the two authors, "explainability" and "interpretability" are often used interchangeably. Although they have the same objective: to understand the model. There are two types of models. Black Box Models, in the realm of machine learning, refer to systems that conceal their internal mechanisms, making it challenging to grasp their operations merely by examining their parameters. These models, while often complex and less interpretable, excel in predictive accuracy. In contrast, White Box Models represent the opposite end of the spectrum. They are characterized by simplicity, and high interpretability, and are favoured in scenarios where comprehending and elucidating the model's decisions holds paramount significance. The choice between these two model types hinges on the trade-off between predictive power and the need for transparency and interpretability in a given application (Molnar 2023). We can approach interpretability and explainability with two types of methods: local, which can explain an individual prediction, or global, which can explain the behaviour of a model. A few methods can be used to capture the interpretability of the models, Partial Dependence Plot (PDP), Individual Conditional Expectations (ICE), Accumulated Local Effects (ALE), Feature Interaction, Feature Importance, Global Surrogate, Local Interpretable Model Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and more.

SHAP and LIME stand out as two of the most widely used methods in explaining machine learning models, nevertheless, they exhibit differences (Molnar 2023). LIME produces local explanations by employing a Kernel-based approach, creating an interpretable model to explain a complex model's decision for a specific observation. Notably, LIME prioritizes interpretability over accuracy, simplifying the model to enhance understanding, even at the cost of reduced accuracy. In contrast, the differentiation lies in the types of explanations they offer, with SHAP providing both local and global perspectives. Employing a game-theoretic approach, SHAP elucidates the contribution of each variable to the final prediction. It assigns a numerical score to each variable, aiming to deliver precise explanations without compromising the accuracy of the model. Despite SHAP's utilization of a more sophisticated methodology to explain variable contributions, it may be more challenging for those unfamiliar with the complexities of the model to comprehend.

Interpretability and explainability have been explored to explain predictions, recommendations, or classifications. In this context, several solutions were developed to be coupled with ML models to provide explanations. However, limited research has been addressed to incorporate those explainable models in dropout prediction or success classification in higher education

institutions. Wang and Zhan (2021) identify interpretability as the main limitation related to artificial intelligence technologies in higher education.

Different data analysis techniques are used to obtain knowledge from educational data, but explainable AI techniques are not yet widely used.

From the few studies, Chitti, Chitti, and Jayabalan (2020) investigate the central role of Artificial Intelligence (AI) and Machine Learning (ML) in education, with a specific focus on predicting student performance. The study highlights how crucial it is to create interpretable models that gain advantage over stakeholders in the education sector and enable data-driven, strategy-based decision-making. Additionally, it promotes the usage of standardised libraries and computationally effective XAI algorithms to enhance the adoption of these approaches in the field of predicting student performance.

Using pedagogical surveys, Leal et al. (2022) investigate interpretable success prediction in Portuguese higher education institutions. The research makes use of a dataset that covers eight years and several courses. With a 98% accuracy and F-measure, the Random Forest classifier produced the most promising results. Three datasets are used in the analysis: one that includes all courses, one that includes large courses, and one that includes small courses. The authors use LIME explanations to pinpoint important factors that impact classification, like the unit mean and the percentage of failed students. The proposed method favours Random Forest as the best classifier, leveraging decision trees for interpretability, and employs LIME to automatically generate explanatory sentences based on relevant branches and survey categories.

Another paper, by Nadaf, Eliëns, and Miao (2021), to investigate and measure the importance of 80 learning factors contributing to student cognitive achievement, using 2018 data from the Programme of International Student Assessment (PISA) and ML approaches. First, they used improved implementations of gradient boosted tree algorithms and second, the Shapley Additive Explanation Framework. They find that a large weekly learning time of more than 35 hours is associated with less positive or even negative effects on the predicted outcome.

Methodology

Dataset Description

The selected dataset is public and originates from a Portuguese higher education institution, *Instituto Politécnico de Portalegre*, including 4,424 records of students who enrolled during the academic years 2008/09 to 2018/2019 (Martins et al. 2021). This dataset with 36 features offers a comprehensive overview of students enrolled in diverse degree programs offered by the institution. It contains demographic data (age at enrolment, gender, marital status, nationality), socio-economic data (parents' qualifications, parents' occupations, fees), macroeconomic data, and information on the time of enrolment and academic performance of the first two semesters. These features can be utilized to analyse potential predictors of student outcomes, such as student performance and dropout rates. Each student record is labelled into three target categories: dropout, enrolled and graduate at the end of the normal duration of the course (3 years). Graduated means that the student obtained the degree in due time; Enrolled means that the student remains enrolled and does not obtain the degree within the expected three years; Dropout means that the student doesn't obtain the degree at all.

Proposed Methodology

Firstly, after becoming familiar with the data set, I performed a careful cleaning of the data to ensure the reliability and quality of the model. This was followed by an exhaustive exploratory analysis of the data. Subsequently, the dataset was divided into two distinct datasets: dataset 1 with all features, which includes all data and academic performance in the first two semesters; and dataset 2, which includes demographic, socioeconomic, and macroeconomic information and academic information at enrolment. This segmentation was intended to facilitate the creation of two models - one applicable immediately after student enrolment or the start of the course, and the other relevant after one year of the course. In the subsequent phase, for both datasets, I prepared the data and carried out an 80/20 split (80% for training and 20% for testing) to facilitate the development of the models and an explainable method. The explainable method aimed to obtain global and local explanations of the model, clarify the importance of the characteristics, and understand their contributions to the final predictions and individual predictions.

Data Preparation

The original dataset contains structured data with 4,424 observations and 37 distinct features, one of which is the target variable, illustrated in Table S1 in the appendix. No missing values, duplicates or inconsistent outliers were detected in the data set.

A significant challenge posed by this dataset was the prevalence of categorical variables with significant cardinality. The presence of high cardinality in a dataset can trigger several problems, including computational and statistical problems, decreased model performance and the curse of dimensionality, in other words, the set of all possible categories may be huge and not known a priori, as the number of different values in the column can indefinitely increase with the number of samples, or the categories may be related with some morphological or semantic links (Cerda and Varoquaux 2019). The classic approach to encode categorical variables is one-hot encoding, however, for high cardinality categories, this leads to feature vectors of high dimensionality. Data engineering practices usually address these problems with data cleaning techniques, such as the aggregation of categories, which require human intervention (Cerda and Varoquaux 2019). To address this problem, variables with more than 4 levels, such as father's qualification, mother's qualification, previous qualification, nationality, marital status, course, order of application and mode of application, were aggregated. This process considered the general aim of the thesis and the frequency of each category, intending to reduce the cardinality of the data set.

Variables such as mother's occupation and father's occupation were excluded due to their high cardinality and some implicit values. The variable international was also eliminated because it transmitted redundant information already captured by nationality. The academic variables, namely the curricular units of the 1st and 2nd semesters, were simplified in order to retain only the essential information for a broader understanding of the 1st year, omitting detailed components such as credited units, evaluations and those without evaluations.

Upon analysing the original target variable (Figure 1), containing the three categories it became evident that it did not align with my primary objective of predicting whether a student successfully complete their course within the expected three-year timeframe or whether they face academic challenges, including both delayed completion and dropout scenarios as cases of failure. To address this, I engineered a new target variable (Figure 2) that clearly focuses on predicting student success, where a positive outcome denotes the timely completion of the

course. This refinement aims to enhance the precision of student identification, with a specific focus on providing timely assistance to those facing academic challenges. So, regarding the new target variable: Success indicates the timely completion of the course, while failure indicates a longer than expected duration or failure to complete the course.

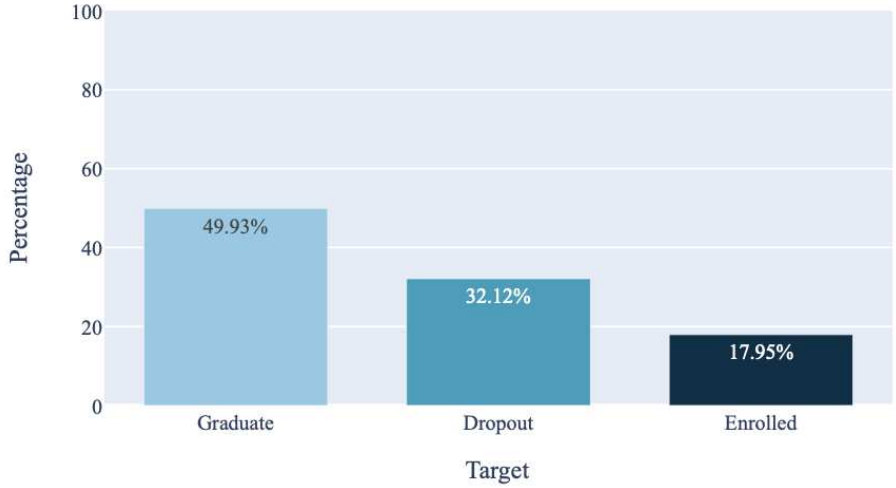

Figure 1: Distribution of the original target variable. The percentage of graduates, dropout and enrolled students varies, with almost 50% of students graduating after the first year.

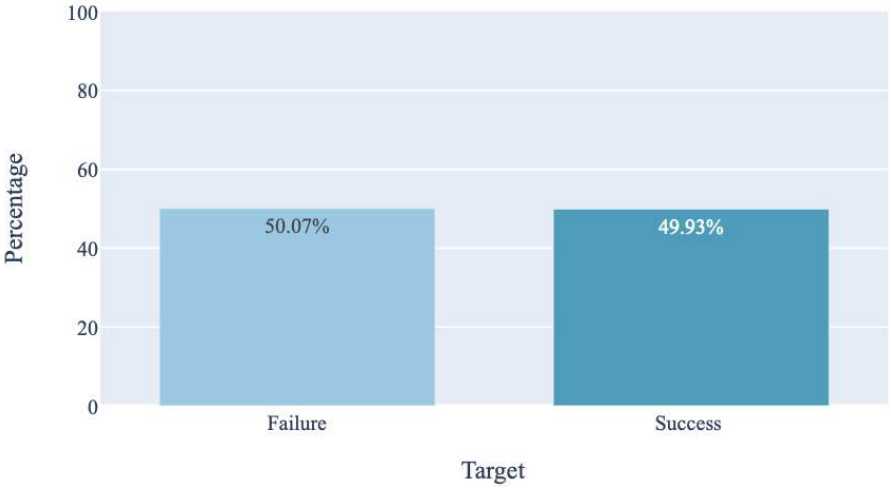

Figure 2: Distribution of the new engineered target variable. Graduate was recoded to success and dropout, or enrolled categories were encoded to failure. The dataset is balanced regarding this new outcome binary variable.

The resulting "clean data" includes 4,424 observations and 28 columns (Table 1), comprising aggregated variables, unchanged numerical variables, and the newly introduced target variable.

Table 1: Data Dictionary of the cleaned dataset. Identification and description of the variables used, categorized by sector and data type.

Variable		Description
Demographic	Marital Status	The marital status of the student. [Categorical]
	Nationality	The nationality of the student. [Categorical]
	Age	The age of the student at the time of enrolment. [Numerical]
	Gender	The gender of the student. [Categorical]
	Displaced	Whether the student is a displaced person. [Categorical]
Social Economic	Mother's Qualification	The qualification of the student's mother. [Categorical]
	Father's Qualification	The qualification of the student's father. [Categorical]
	Scholarship Holder	Whether the student is a scholarship holder. [Categorical]
	Educational Special Needs	Whether the student has any special educational needs. [Categorical]
Macro-economic	Unemployment Rate	From the region. [Numerical]
	Inflation Rate	From the region. [Numerical]
	GDP	From the region. [Numerical]
Academic at enrolment	Application Mode	The method of application used by the student. [Categorical]
	Application Order	The order in which the student applied. [Categorical]
	Course	The course taken by the student. [Categorical]
	Previous Qualification	The qualification obtained by the student before enrolling in higher education. [Categorical]
	Previous Qualification (grade)	The grade obtained by the student before enrolling in higher education. [Numerical]

	Admission Grade	Represents the cumulative academic qualification or score used by higher education institutions to assess an applicant's overall academic suitability. [Numerical]
	Attendance Preference	Whether the student attends classes during the day or in the evening. [Categorical]
Academic at the end of first year	Tuition fees up to date	Whether the student's tuition fees are up to date. [Categorical]
	Debtor	Whether the student is debtor. [Categorical]
	Curricular Units 1st sem (enrolled)	The number of curricular units enrolled by the student in the first semester. [Numerical]
	Curricular Units 1st sem (approved)	The number of curricular units approved by the student in the first semester. [Numerical]
	Curricular Units 1st sem (grade)	The average grade of curricular units in the first semester. [Numerical]
	Curricular Units 2nd sem (enrolled)	The number of curricular units enrolled by the student in the second semester. [Numerical]
	Curricular Units 2nd sem (approved)	The number of curricular units approved by the student in the second semester. [Numerical]
	Curricular Units 2nd sem (grade)	The average grade of curricular units in the second semester. [Numerical]
	Target	Success, Failure. [Categorical]

Modelling

In the modelling process, a Random Forest Classifier and a CatBoost Classifier were implemented using the scikit-learn package and the CatBoost package, respectively. The objective was to appropriately model education performance prediction. It is crucial to note that all steps were carried out on two datasets: one applicable immediately after student enrolment or the start of the course, and the other relevant after one year of the course.

The Random Forest algorithm was my choice of benchmark model, since almost all the studies carried out highlighted its outstanding performance (as described in the related work section of the literature review) (Alhazmi, Sheneamer, and Sheneamer 2017; Beaulac and Rosenthal 2019; Hoffait and Schyns 2017; Miguéis et al. 2018). A widely adopted machine learning algorithm that employs an ensemble method. This approach involves combining the outputs of multiple decision trees to derive a final result through majority voting.

The CatBoost is a machine learning algorithm rooted in gradient boosting with high-performance decision trees. What sets this algorithm apart is its robust support for categorical features. This implies that enhancing training outcomes with CatBoost enables the use of non-numerical factors without the need for extensive pre-processing or the conversion of such factors into numerical representations (Yandex 2023). This capability represents a notable advantage, particularly in datasets predominantly characterized by categorical variables - a relevance that is especially noteworthy in the dataset under consideration for this study (as detailed in Data Preparation).

Model Pipeline

Data Processing

Before starting the modelling process, a series of data pre-processing and feature engineering steps were undertaken.

In the Random Forest model, to convert categorical variables into numerical ones, I utilized label encoding to transform binary categorical variables into numerical ones using Label Encoder. Additionally, I applied one-hot encoding to convert non-binary categorical variables into numerical ones, and redundant columns were subsequently dropped.

In the CatBoost model, owing to its significant advantage, categorical variables are automatically handled, eliminating the need for any pre-processing or feature engineering steps.

To create two models - Model A: after one year of the course, and the other, Model B: applicable immediately after the student's enrolment or the start of the course, two separate datasets were compiled, one for each model. Remarkably, both datasets share the same instances, with one model using all the variables, while the other excludes academic at the end of first year variables. This distinction in model characteristics is crucial for the analysis, allowing it to explore the impact of certain variables on predicting timely course completion.

Model Training

The same modelling pipeline was applied for the 2 defined datasets, with an initial train/test split of 80/20 (Figure 3). As illustrated in Figure 4, both subsets show a balanced distribution of the target variable.

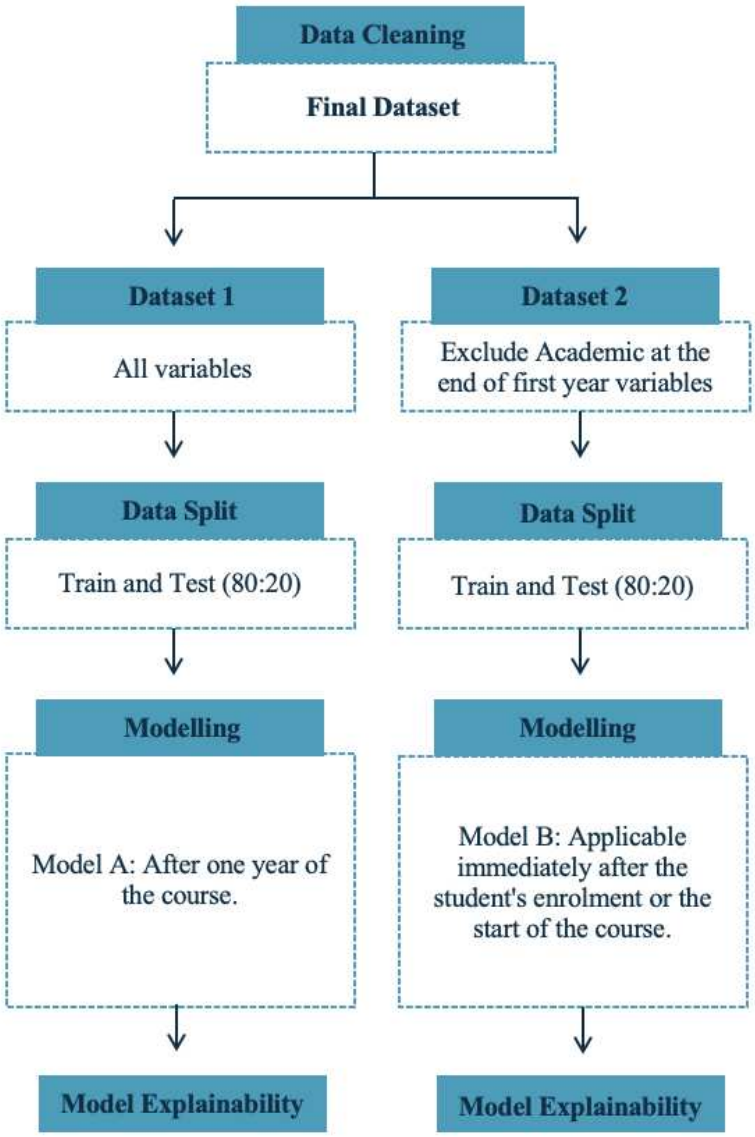

Figure 3: Workflow Process. [1] The data was subject to data cleaning and was divided into 2 distinct final datasets. [2] Each final dataset was split into 2 subsets of data on an 80:20 ratio. [3] Modelling process to predict student performance using the split data to train and evaluates its performance, for each model. [4] Use SHAP model to explain and understand the predictions of the two models.

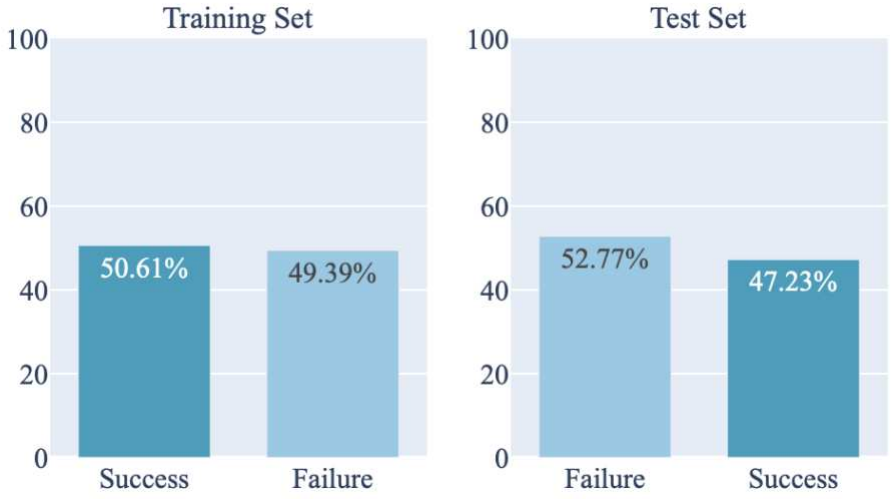

Figure 4: Distribution of the target variable in the two subsets created. The plots demonstrate a consistent distribution of the binary variable in both subsets.

For the 2 pipelines, benchmark model and CatBoost model performance were assessed. Parameter tuning was performed in both models using grid search. The conclusion drawn was that the CatBoost algorithm demonstrated superior performance.

Focusing on the selected algorithm for both datasets, the CatBoost classifier was utilized, employing hyperparameter tuning through grid search with cross-validation and optimizing precision as the scoring metric. The best model was determined based on the optimal combination of hyperparameters, including depth, iterations, learning rate, L2_leaf_reg, and border_count, as illustrated in Table S2 in the appendix. Each of these hyperparameters plays a distinct role in the model (The CatBoost team and Gulin). Depth, which defines the maximum depth of each decision tree, allows for the capture of intricate relationships. However, deeper trees also carry the risk of overfitting. Iterations, representing the number of trees, were adjusted to facilitate the model's learning of complex patterns. However, this adjustment involves a trade-off with computational cost. The learning rate, governing the step size in each iteration towards minimizing the loss function, plays a crucial role. A lower learning rate results in slower convergence but can lead to a more accurate model. L2_leaf_reg, which controls the regularization term for leaf-wise splits, prevents overfitting by penalizing large coefficients in leaf features. Additionally, border_count specifies the number of splits considered for numerical features.

Model Evaluation

The learning curve was visualized, representing the accuracy scores of training and cross-validation on different training set sizes. The metrics chosen to assess the model's performance included accuracy, recall, precision and F1 score through the confusion matrix, the Receiver Operating Characteristic curve (ROC curve) and the Area Under the ROC curve (AUC) were also analysed.

The confusion matrix is a valuable tool for calculating various metrics, including accuracy, precision, recall and F1 score. As illustrated in Figure 5, accuracy tells us about the proportion of positive cases correctly predicted, providing information about the reliability of our model. Meanwhile, recall indicates how effectively our model identifies actual positive cases. The F1 score, which harmonises precision and recall, reaches its maximum when precision equals recall. However, it is crucial to note that the specific metric being maximised is not explicitly mentioned.

In the confusion matrix (Figure 5), two types of errors, Type I and Type II, can occur. Type I error occurs when the model incorrectly predicts that a student was successful in the course when they were not, resulting in a false positive. On the other hand, Type II error occurs when the model decides not to predict a successful student, treating them as a failure, resulting in a false negative. These errors are inversely related, but there is a way to measure the tradeoff between Type I and Type II errors. The selection of a specific threshold or specific weight for a class can be defined to maximize the minimally impactful type of error in the overall problem domain.

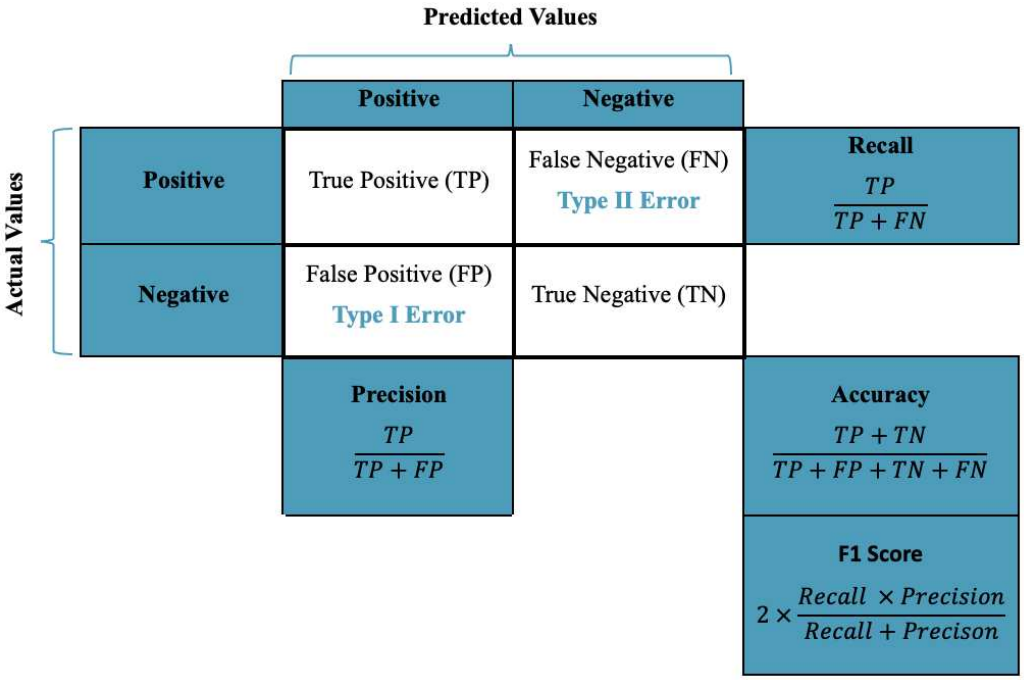

Figure 5: Representation of a Confusion Matrix.

The ROC curve visually represents the performance of a classification model through two key parameters: the true positive rate (recall) and the false positive rate, each assessed at various classification thresholds. The proximity of the graph to the upper and left borders indicates the better accuracy of the model. Furthermore, the AUC serves as a comprehensive measure of performance across all conceivable classification thresholds. A higher AUC value signifies the superior performance of the classifier.

Model Explainability

Explainability can be approached globally, providing an overview of the model and how data features collectively influence the outcome, or locally, offering insights into each instance and individual data feature, showcasing how each characteristic independently impacts the result. The most commonly used explainability methods, as mentioned earlier, are LIME and SHAP - both agnostic methods that explain model interpretation independently of the model type.

In this context, I have chosen to employ SHAP (Shapley Additive Explanations), which is an agnostic model capable of generating interpretable global and local explanations for any machine learning model type. The Shapley value, rooted in game theory, enables us to quantify the contribution of each characteristic to the model's outcomes. It represents the average marginal contribution of a characteristic value in all possible coalitions (Molnar 2023). Shapley values leverage the predicted value and the average predicted value, ensuring that the difference between the prediction and the average prediction is equitably distributed among the instance's feature values - a key characteristic known as the Efficiency property of Shapley values. This property sets Shapley values apart from other methods, such as LIME, which lacks a guarantee of fair distribution of predictions among features. Shapley values uniquely enable contrastive explanations, allowing comparisons not just to the average prediction of the entire dataset but also to subsets or even individual data points. This contrasting capability is absent in local models like LIME. Moreover, Shapley's values stand out as the only explanation method grounded in a solid theoretical framework. The axioms - efficiency, symmetry, dummy, additivity - provide a rational foundation for the explanation, enhancing its credibility (Molnar 2023).

Through the application of SHAP, we can assess the relative contribution of each student characteristic, gaining insights into the key factors influencing the model's predictions - specifically, school success or failure. Consequently, I utilized Shapley values to analyse the predictions of a CatBoost model forecasting student success or failure in the two distinct models

I created: one applicable immediately after student enrolment or the start of the course and the other relevant after one year of the course.

Results

After explaining the modelling process and conducting a comprehensive sensitivity analysis, we can now explore the results from the model for both datasets.

Exploratory Data Analysis

I began by identifying various types of information about the 4,424 students, such as demographic data, socioeconomic data, macroeconomic data about the region they belong to and, finally, academic data at enrolment and of the first year about each student.
First, I analysed the new target variable, where we can conclude that the percentage of success and failure is very similar, 49.93% and 50.07%, respectively (Figure 2).

In terms of demographics, it was noted that of the total students, 2,868 (64.83%) are female, and 1,556 (35.17%) are male (Figure 6). Their ages range from 17 to 70 years, and notably, males exhibit a higher percentage of failures compared to females, with rates of 64.78% and 42.09%, respectively. Regarding the relationship between students' ages and the target variable (Figure 7), it is evident that students who fail tend to have a higher average age (25 vs. 22), compared to students who graduate. However, both groups show similar minimum and maximum age values, ranging from 17 to 70 years for non-graduates and 17 to 62 years for graduates.

Figure 6: Distribution of target variable by Gender. The plot reveals a predominance of females, yet the dropout rate is higher among males.

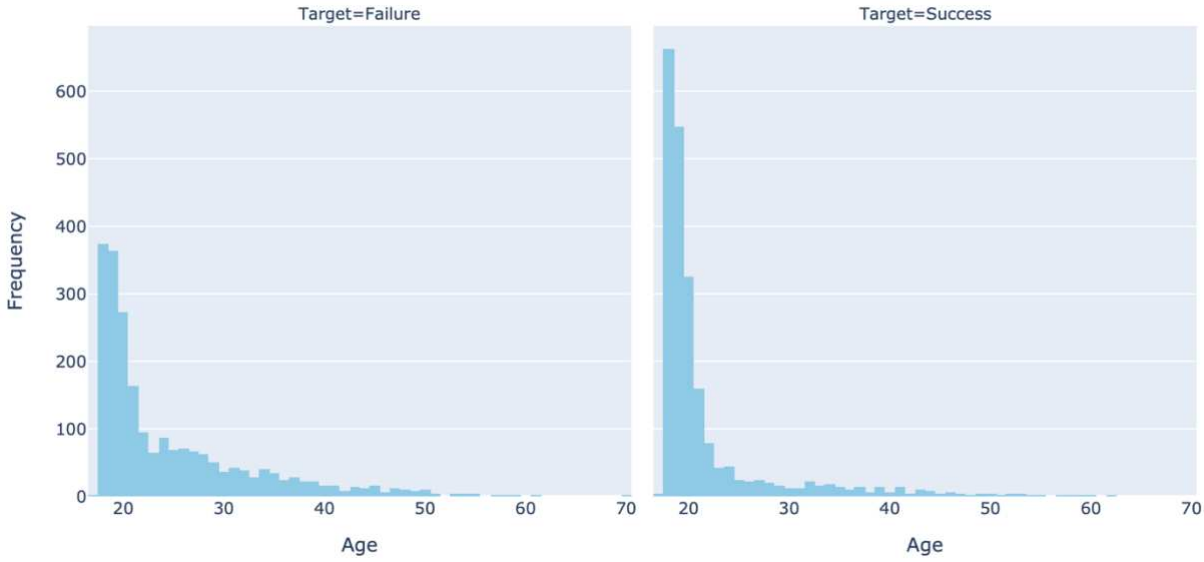

Figure 7: Age distribution by target variable. The plot shows a positively skewed tendency for both classes.

The majority of students are Portuguese (97.51%), and their nationality does not seem to influence success in the course (Figure 8).

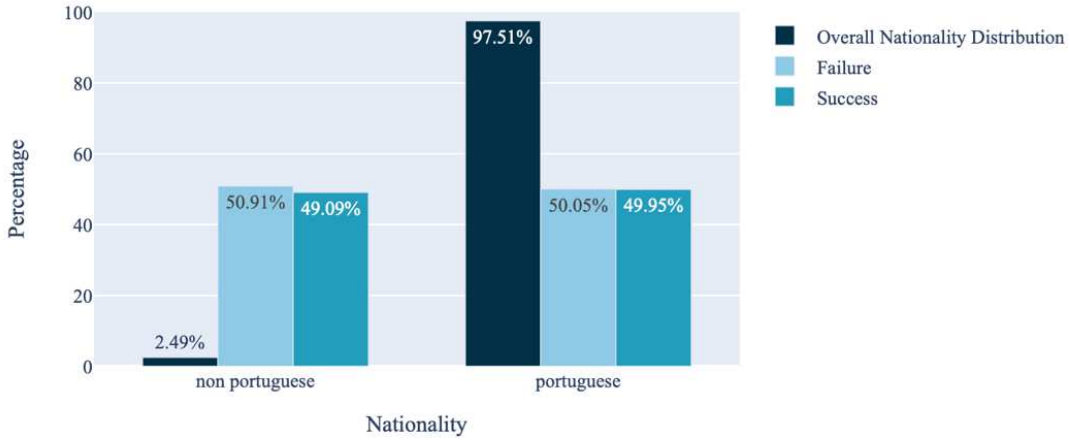

Figure 8: Distribution of target variable by Nationality. The plot reveals a predominance of Portuguese students, yet the dropout rate is higher among non portuguese.

Moreover, most students are unmarried (88.58%), and in this case, only single students exhibit a higher success rate, in contrast to married students and other marital status categories (Figure 9).

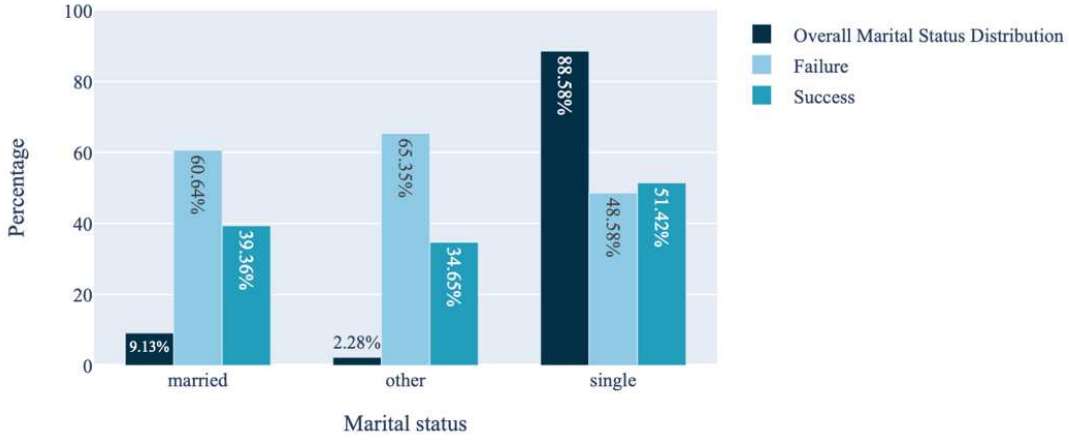

Figure 9: Distribution of target variable by Marital status. The plot reveals a predominance of single students, yet the dropout rate is higher among married and other students.

Regarding the displaced variable, I found that the percentage of students in a different district than their own or in the same district is quite similar, at 54.84% and 45.16%, respectively. However, there is a higher failure rate if students remain in their district without needing to relocate (Figure 10).

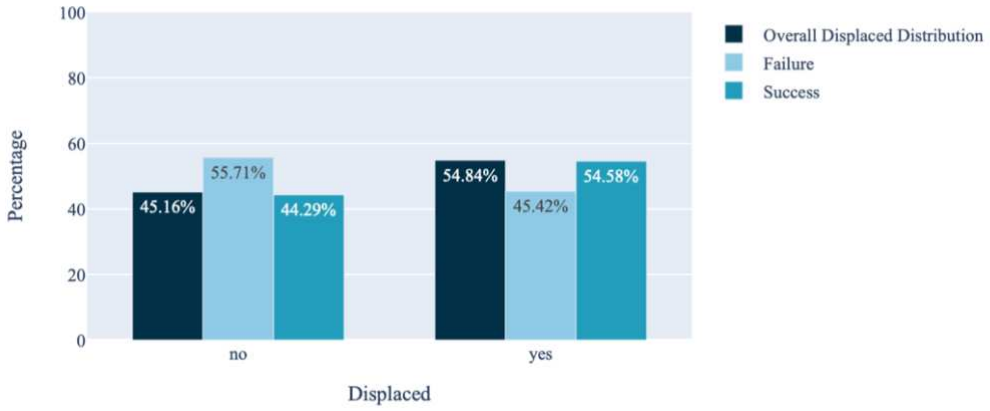

Figure 10: Distribution of target variable by Displaced. The plot reveals a predominance of displaced students, yet the dropout rate is higher among non displaced students.

In terms of socioeconomic factors, it was observed that the majority of students have their tuition fees up to date (88.07%), do not receive a scholarship (75.16%), have no special needs (98.85%), and are not considered debtors (88.63%) (Figure 11, 12, 13 and 14). However, the failure rate among students is higher when they do not have their tuition fees paid, do not receive a scholarship, and are considered debtors. The variable of special needs does not seem to have any discernible influence on the success or failure of students, as it is also represented by a minority and thus lacks significant results.

Figure 11: Distribution of target variable by Tuition fees up to date. The plot reveals a predominance of students that have their tuition fees up to date, yet the dropout rate is higher among students when they do not have their tuition fees paid.

Figure 12: Distribution of target variable by Scholarship holder. The plot reveals a predominance of students that do not have scholarship, and the dropout rate is higher among these students.

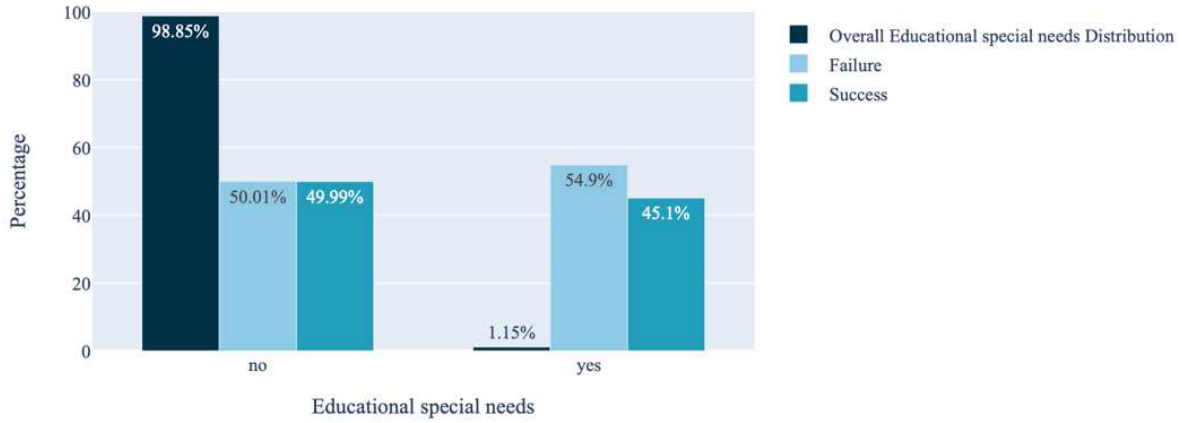

Figure 13: Distribution of target variable by Educational special needs. The plot reveals a predominance of students without educational special needs, yet the dropout rate is higher among the students with educational special needs.

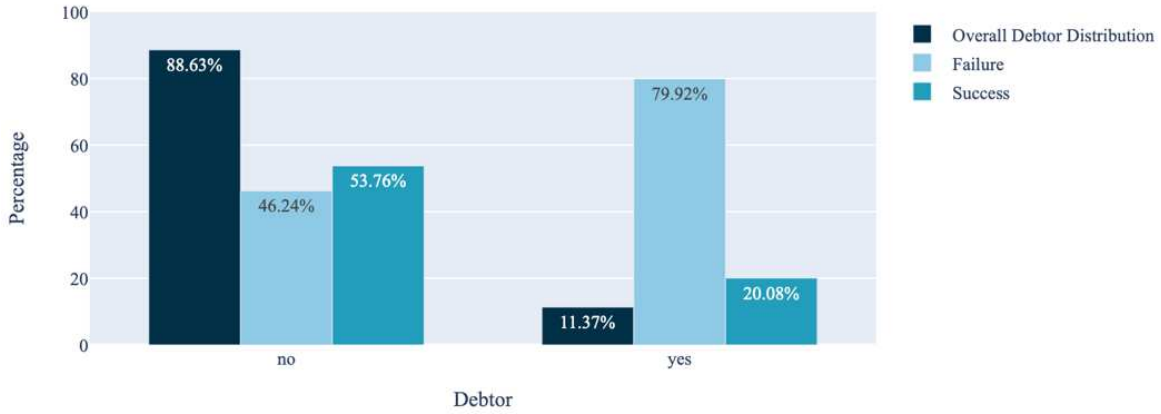

Figure 14: Distribution of target variable by Debtor. The plot reveals a predominance of students without debts, yet the dropout rate is higher among the students with debts.

Concerning the educational background of the student's parents, the majority have only completed basic education, and there doesn't appear to be a discernible influence on the academic success of the students (Figure 15).

Figure 15: Distribution of Parents' qualifications by target variable. The plot shows a similar distribution between parents for both students' outcomes. However, when the students succeed, the father's qualifications in technical courses are higher than the mother's, unlike those who fail.

In terms of macroeconomic information (Figure 16), we can see that both the unemployment rate and GDP have a positive but low correlation, this suggests that, on average, an increase in the unemployment rate or in GDP is weakly associated with higher student success. On the other hand, the inflation rate has a low but negative correlation, this suggests a weak tendency for higher inflation to be associated with a lower likelihood of student success. In addition, it should be noted that these variables do not seem to influence student success or failure, due to their low correlations with the target variable.

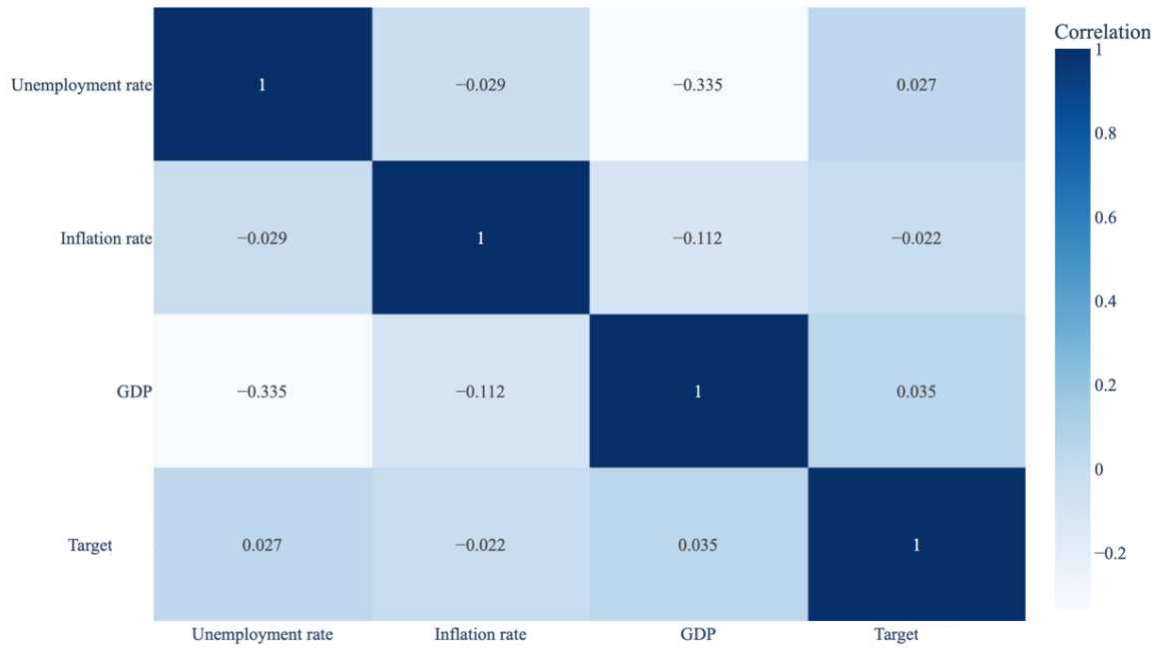

Figure 16: Correlation Matrix of macroeconomic variables with target variable. The matrix reveals that all variables have a low correlation with the target.

Finally, through academic information at enrolment, the data shows that students entering this university with an average grade of 132.6 (with no significant difference based on the course's purpose) predominantly come directly from secondary education (84.02%) (Figure 17). Notably, this category demonstrates a higher success rate in completing the course, while other categories such as technical courses, basic education, and higher education seem to be more prone to dropout, showing a higher failure rate than success (Figure 18).

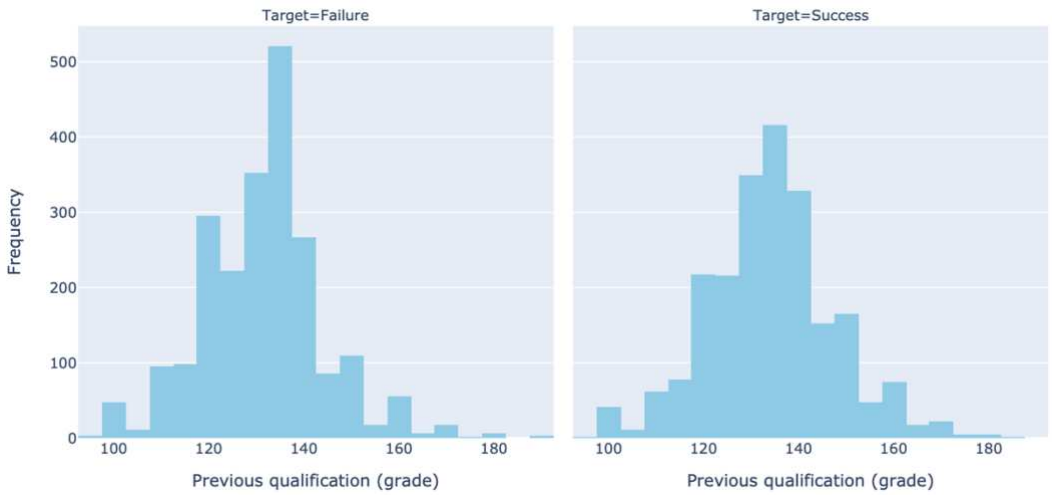

Figure 17: Previous qualification grade distribution by target variable. The plot shows an approximately symmetrical tendency for both classes.

Figure 18: Distribution of target variable by Previous qualification. The plot reveals a predominance of with higher education, yet the dropout rate is higher among students with only basic education.

Concerning the mode of application to the university, the failure rate is consistently higher than the success rate when students already hold other diplomas, are over 23 years old, are transfers, or enrol through special contingents (Figure 19). In the case of the application order, there is a higher failure rate than success for students who enter their first choices (68.42%, where the failure rate is 53.45%) (Figure 20). The fields of Health, Social Services, and Business and Management attract the highest number of students, with Tourism, Science, Business and Management, and Technology presenting higher failure rates than success (Figure 21).

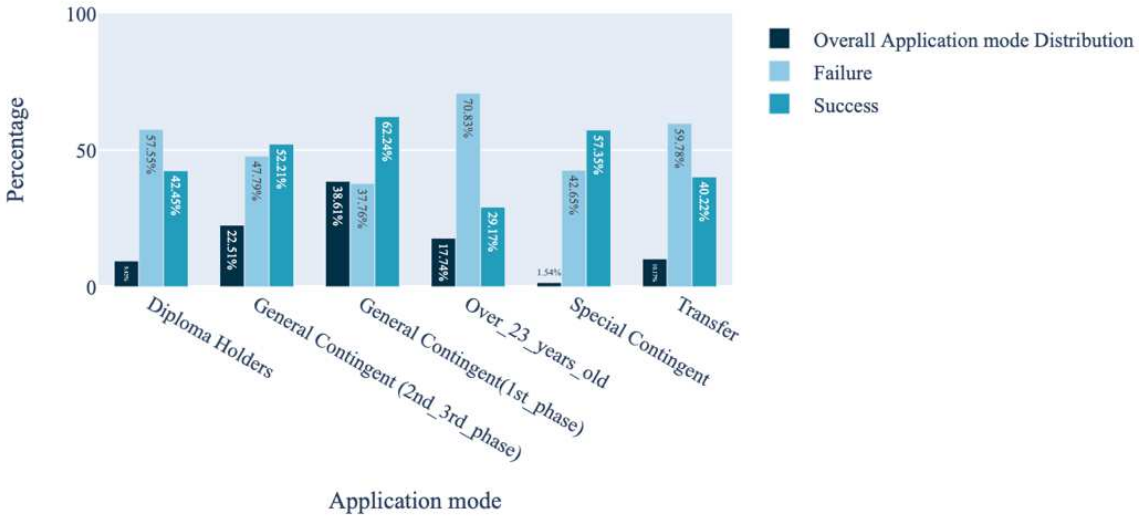

Figure 19: Distribution of target variable by Application mode. The plot reveals a predominance of students that applies in 1st phase of general contingent, yet the dropout rate is higher among students that applies with more than 23 years old.

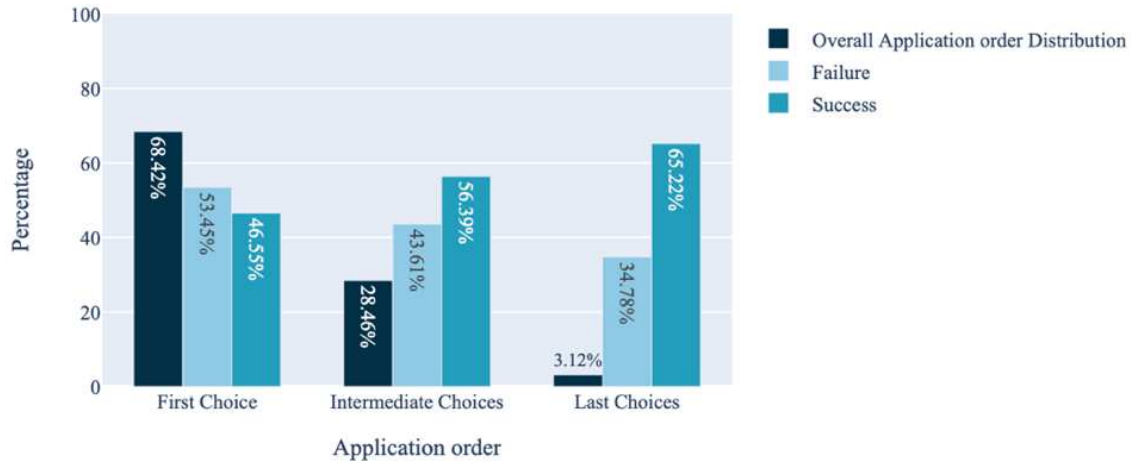

Figure 20: Distribution of target variable by Application order. The plot reveals a predominance of students that enter in their first choice and the dropout rate is higher among these students.

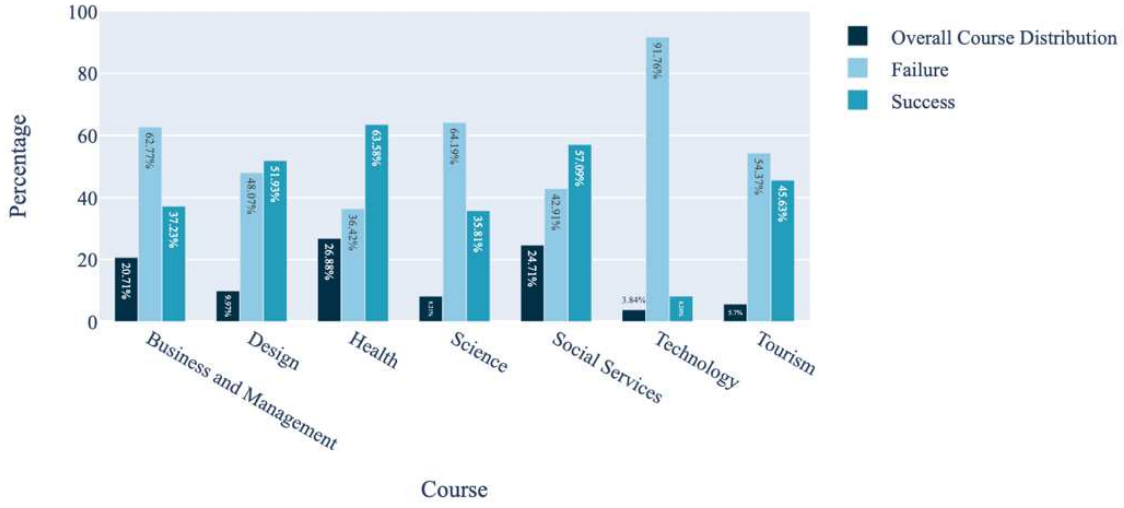

Figure 21: Distribution of target variable by Course. The plot reveals a predominance of students in Health courses, yet the dropout rate is higher among students in technological courses.

Concerning the admission grade to enter the university, it is quite similar for both successful and unsuccessful students, with an institute wide average of 127.0, ranging from 95.0 to 190.0 (Figure 22). Students also prefer daytime classes (89.08%) over evening classes (10.92%), with evening classes showing a higher failure rate (58.39% compared to 49.05%) (Figure 23).

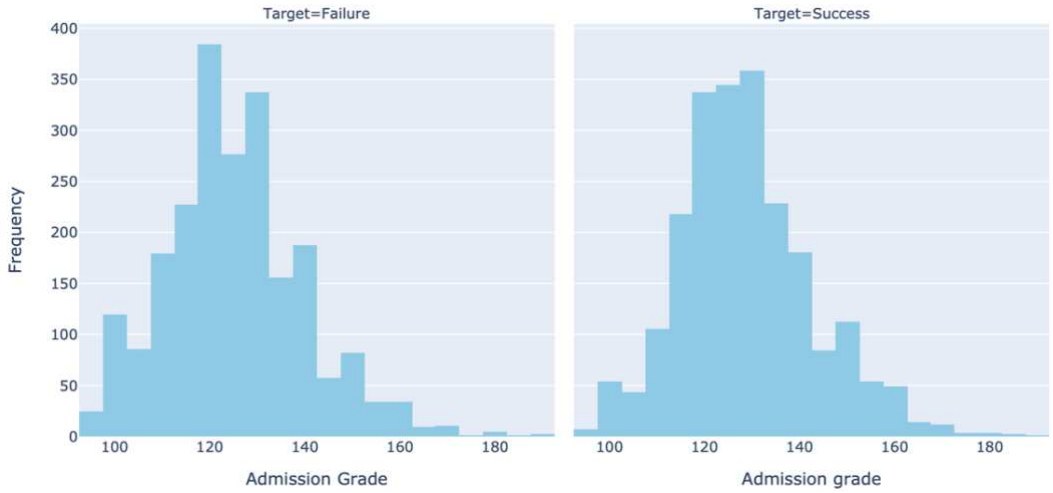

Figure 22: Admission grade distribution by target variable. The plot shows an approximately symmetrical tendency for both classes.

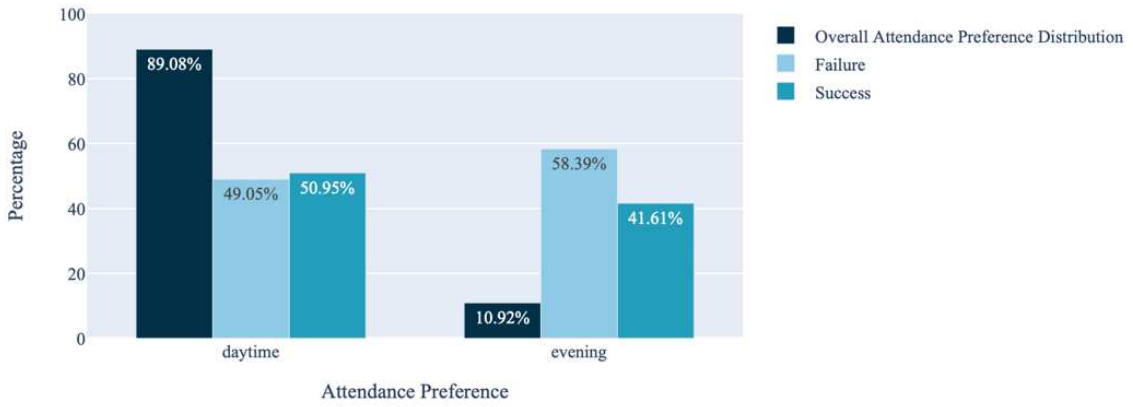

Figure 23: Distribution of target variable by Attendance preference. The plot reveals a predominance of students with daytime classes, yet the dropout rate is higher among the students with evening classes.

In addition to pre-enrolment data, academic records include information from the first two semesters regarding curriculum units. Figure 24 illustrates that both successful and unsuccessful students exhibit a fairly linear behaviour between the first and second semesters. They maintain the same number of enrolled and passed courses, with a slight difference: successful students maintain their semester average of 13, while those facing academic challenges decrease their average from 9 to 8. Another noteworthy point is that successful students, on average, fail at most one course per semester out of the 7 enrolled, whereas unsuccessful students, on average, fail half of the courses per semester out of the 6 enrolled in this case.

Figure 24: Mean for each Unit Curricular variable by target variable.

Modelling

In this section, I will present the performance of the proposed models, Model A and Model B. It is important to note that, upon analysing the confusion matrix, it was noted that despite the model's overall good performance, it exhibited higher False Positive (FP) values, Type I error, than False Negatives (FN) values, Type II error. This observation did not align with my research objective. Given that my thesis goal is to predict student outcomes such as dropout or failure to subsequently create and implement proactive measures and support, predicting FP has a higher cost than predicting FN. I prefer to assist a student who may not need it rather than failing to identify a student at risk. After conducting further research, my solution to address this issue was to adjust the threshold in the model or assign more weight to a specific class in my model using the class weight parameter.

Following another sensitivity analysis, I experimented with various thresholds, various class weights, and the combination of the threshold with class weight, and concluded that the optimal solution differed for the two datasets. Model A, relevant after one year of the course, exhibited improved performance and lower FP values when class weight was implemented, giving greater weight to class 0 (students who fail). On the other hand, Model B, applicable immediately after student enrolment or the start of the course, proved more sensitive to both class weight and threshold. To maintain consistency with my objective, a threshold was implemented.

In this way, two models were created, and their performance was re-evaluated with the respective evaluation metrics.

Firstly, Table 2 shows the best parameters for the models after the grid search.

Table 2: Best Model Parameters. Identification of the best parameters used for the two models.

Best Parameters	Model A	Model B
'border_count'	70	70
'depth'	7	7
'iterations'	50	50
'l2_leaf_reg'	5	5
'learning_rate':	0.1	0.1
‘class_weight’	[{0:2, 1:1}]	-
‘threshold’	-	0.51

Model A: Relevant after one year of the course

The ROC curve and Precision-Recall curve, illustrated in Figure 25 (A) and Figure 25 (B) respectively, depict the performance of the metrics after the addition of the class weight parameter. Additionally, Figure 25 (A) also shows the Area Ander the ROC curve, 0.92, providing an aggregate measure of performance across all possible classification thresholds. After visualizing these graphs, it seems that the model exhibits good performance.

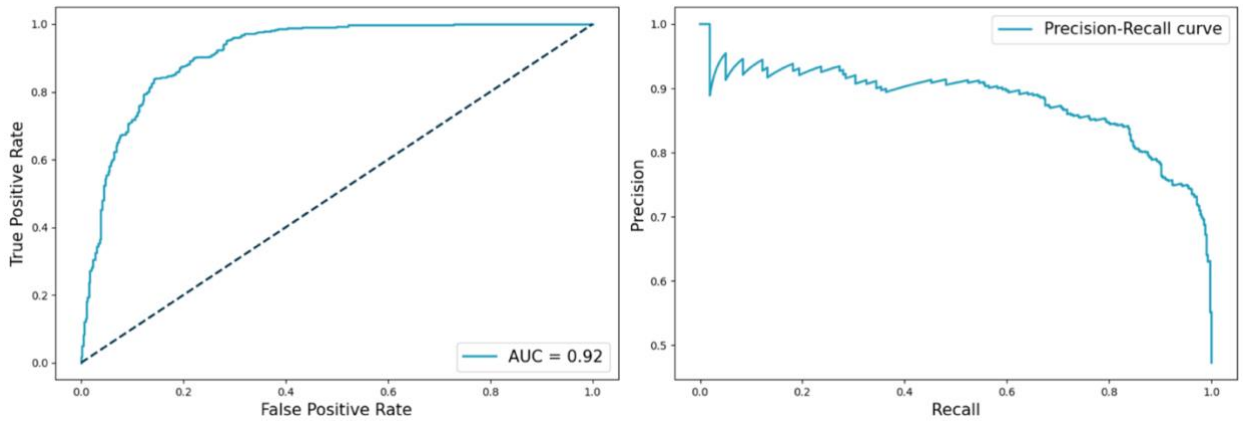

Figure 25: (A) The ROC Curve and AUC for Model A. (B) Precision-Recall Curve for Model A.

Furthermore, the confusion matrix in Figure 26, after the application of the class weight parameter, reveals an interesting pattern in the distribution of prediction errors. Specifically, there is a considerable increase in false negatives (FN) compared to false positives (FP). As a result of the model's sensitivity to the class imbalance introduced by adjusting the class weights, this change in error types is noticeable. The high False Negative count means instances where the model incorrectly predicted that a student would fail (class 0) but succeeded (class 1). In contrast, the False Positive count represents instances where the model predicted success (class 1), but the student failed (class 0).

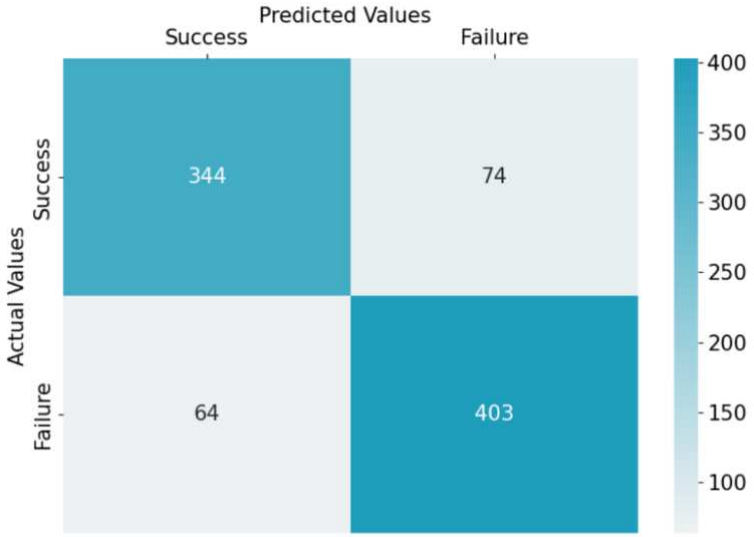

Figure 26: Confusion Matrix - Test Set of Model A.

With the optimized parameters, the performance metrics for the train and test results are represented in Table 3, and the learning curve in Figure 27.

Table 3: Results of the performance metrics for Model A.

	Training Set Metrics:	**Test Set Metrics:**
Accuracy	0.8627	0.8441
Precision	0.8900	0.8431
Recall	0.8314	0.8230
F1 Score	0.8597	0.8329

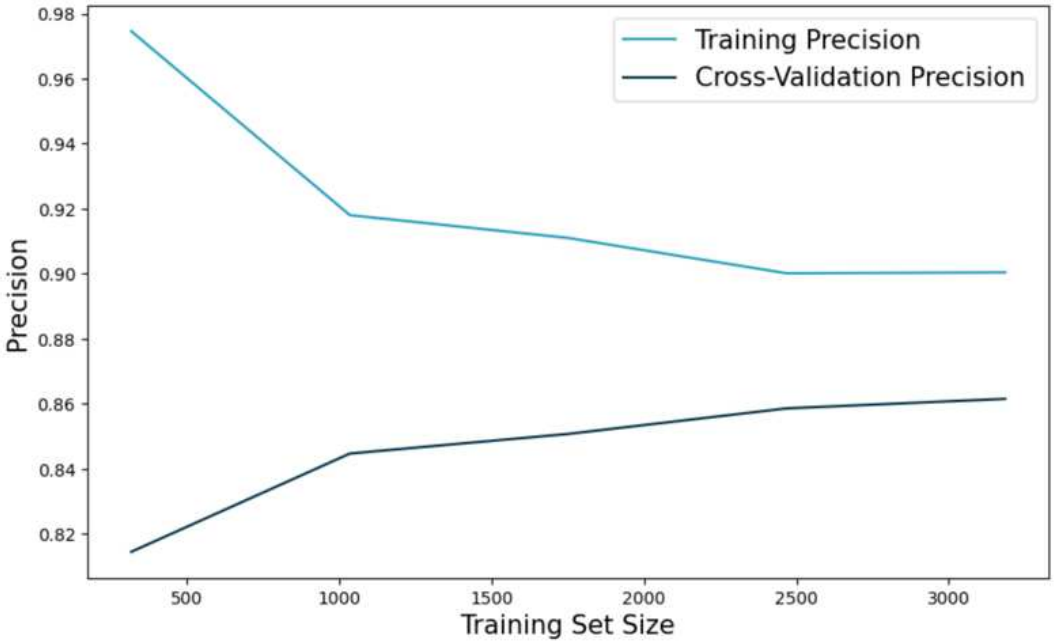

Figure 27: Learning Curve for Model A.

By looking at the learning curve (Figure 27) and the metric values for both sets (Table 3), it was observed that the model can generalise well. Meaning that it not only performs well on training data but also on new and unseen data, with no significant difference between the model's performance on training and test data.

Table 3 includes accuracy results, indicating how well the model can correctly classify whether a student succeeds or not in the course in around 84% of test observations. Precision results show that 84% are correctly predicted success observations out of all predicted success observations. Recall results highlight that 82% are correctly predicted success observations out of all actual success cases. Lastly, a good F1 score suggests low false positives and low false negatives, considering both precision and recall, this measure indicates how well the model is performing accurately. Given the high precision and high recall, the F1 score is also high.

Model B: Applicable immediately after student enrolment or the start of the course

In this model, without the information of the first year in the course, the best parameters after the grid search are the same but with a threshold (Table 2), along with the ROC curve, AUC (Figure 28 (A)), and Precision-Recall curve (Figure 28 (B)) graphics appear to exhibit lower performance compared to the dataset with all the information.

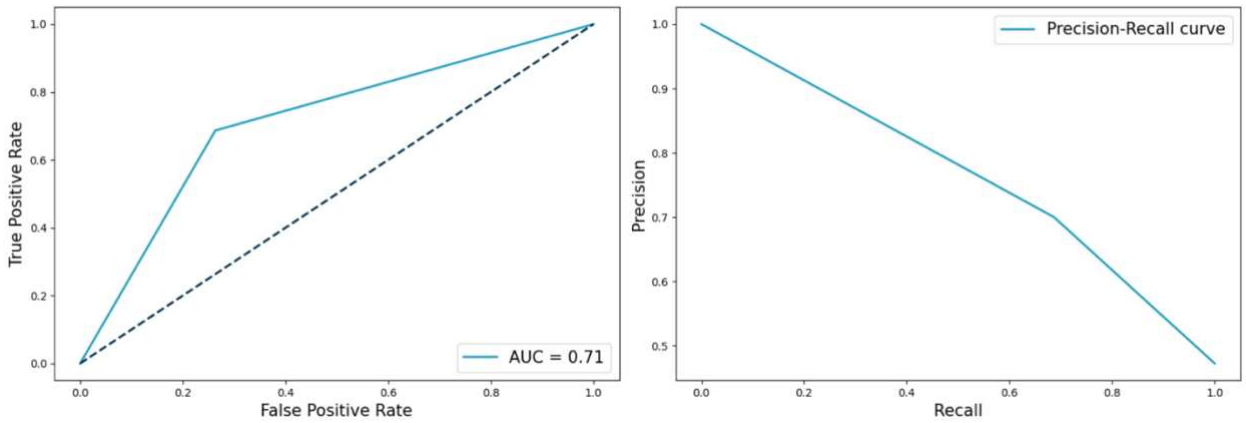

Figure 28: (A) The ROC Curve and AUC for Model B. (B) Precision-Recall Curve for Model B.

The confusion matrix (Figure 29) after applying the threshold. Here, we can also observe an increase in False Negatives (FN) rather than False Positives (FP).

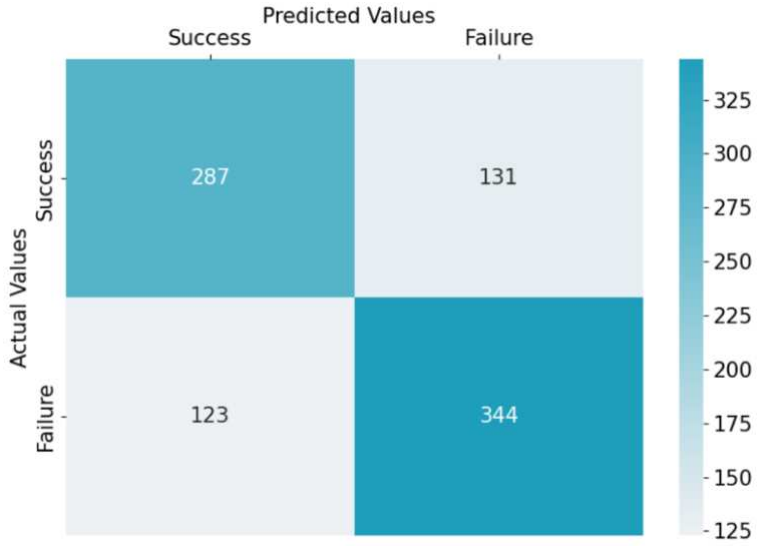

Figure 29: Confusion Matrix - Test Set of Model B.

With the optimized parameters, the performance metrics results are presented in Table 4, and the learning curve in Figure 30.

Table 4: Results of the performance metrics for Model B.

	Training Set Metrics:	**Test Set Metrics:**
Accuracy	0.6982	0.7130
Precision	0.7203	0.7000
Recall	0.6600	0.6866
F1 Score	0.6888	0.6900

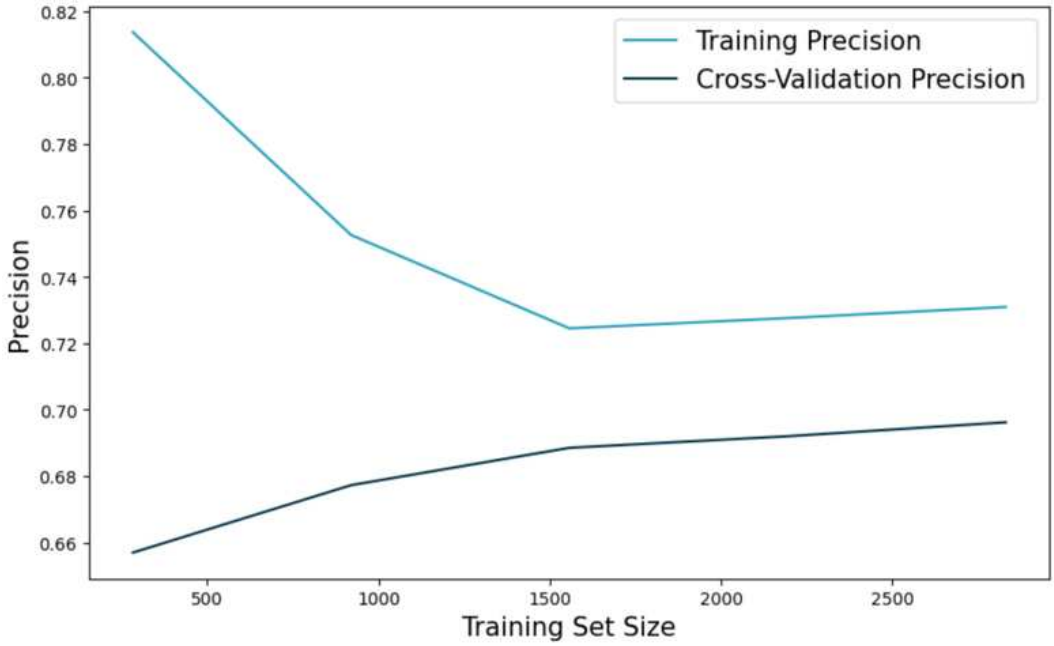

Figure 30: Learning Curve for Model B.

By looking at the learning curve, in Figure 30, and the metric values for both sets, in Table 4, it was observed that the model can generalise well too.

Overall, it is evident that the dataset with all information demonstrates better performance. Through the accuracy results, we observe that the model can correctly classify whether a student succeeds or not in the course in around 71% of test observations. Precision results indicate that from all predicted success observations, the model correctly predicts 70% of those observations.

Recall results highlight that from all actual success cases, the model correctly predicts 69% of those observations. Lastly, the F1 score, considering both precision and recall, is also medium/low emphasizing that this metric evaluates how well the model performs accurately.

Model Explainability

The results from the SHAP method will be at a global and local level for each model.

Model A: Relevant after one year of the course

To gain an overview of the most crucial features of the CatBoost algorithm, I plotted the SHAP values for each feature across all samples. The first plot simply takes the mean absolute value of the SHAP values for each feature, resulting in a standard bar plot. The second plot sorts features by the sum of SHAP value magnitudes across all samples, using SHAP values to illustrate the distribution of the impact each feature has on the model output. Both plots provide a global perspective.

As seen in Figure 31, the most significant variables are the courses approved in the 1st and 2nd semesters, carrying more weight in predicting the absolute probability of student success. On the other hand, less important variables, excluding those with no impact on the predicted absolute outcome probability, include gender and marital status. The variable curricular units approved in the 2nd semester change the prediction outcome, on average, by 1.08%, while the variable curricular units approved in the 1st semester change the prediction outcome, on average, by 0.43%. Additionally, through Figure 32, it is evident that the variables curricular units approved in the 2nd semester, curricular units approved in the 1st semester, curricular units grades in the 2nd semester, and scholarship holder reduce the predicted probability score.

Figure 31: Global Feature Importance plot for Model A.

Figure 32: Global Magnitude Feature Importance plot for Model A.

As mentioned in the introduction, my objective is to gain insights into general trends in student performance. However, beyond understanding the overall trends in the data, it is equally crucial to possess the capability to intervene individually for each student. This is accomplished through localized plots of the SHAP values.

To obtain a detailed perspective, I input a series of SHAP values into the waterfall bar plot function. This process generates a local feature importance plot wherein the bars represent the SHAP values for each feature. The plot visually illustrates the contribution of individual features for a specific student. Features that elevate the prediction are depicted in red, while those diminishing the prediction are shown in blue. It's noteworthy that, by default, SHAP explains the CatBoost classifier model in terms of its margin output before the logloss function. Consequently, the units on the x-axis are in log-odds units, with negative values implying probabilities of less than 0.5 for student success in the course. The grey text accompanying the feature names displays the value of each feature for the given student.

As an example, analysing Figure 33, there is a student with a high probability of success. Having six curricular units approved in the 2nd semester drastically increases the predicted probability of success in the overall course for this student.

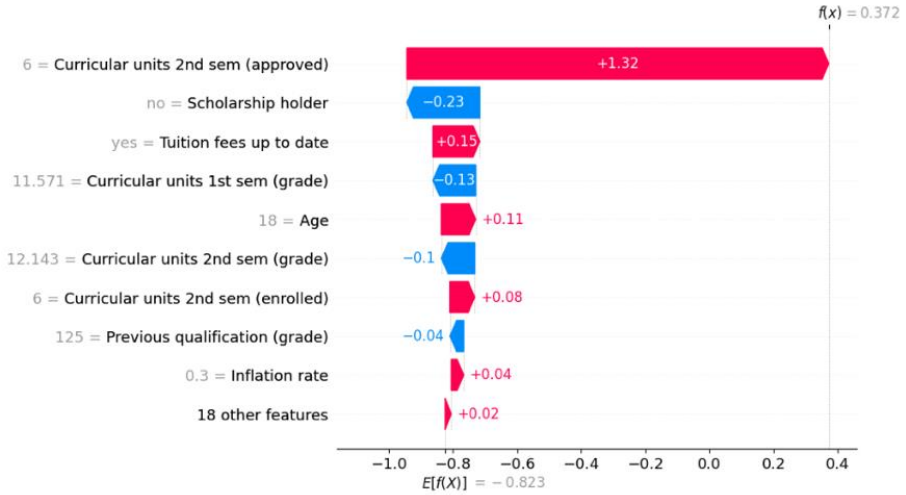

Figure 33: Local Feature Importance plot for Model A. The plot represents a student with high probability of success.

On the other hand, in Figure 34, a student with a high probability of failure registers a significant decrease in the predicted probability by having only one curricular unit approved in the 2nd semester.

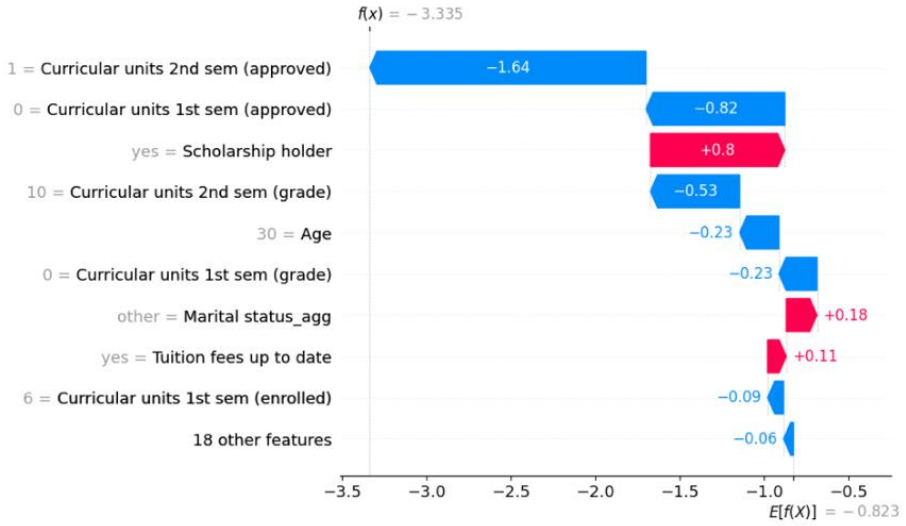

Figure 34: Local Feature Importance plot for Model A. The plot represents a student with high probability of failure.

Model B: Applicable immediately after student enrolment or the start of the course

We observe a distinct behaviour in this model. Beginning with a global overview, as depicted in Figure 35, the most influential variables are age, scholarship holder, and gender, carrying more weight in predicting the absolute probability of student success. Conversely, the less important variables, excluding those with no impact on the predicted absolute outcome probability, include the inflation rate and the unemployment rate of each student. The variable age changes the prediction outcome, on average, by 0.23%; scholarship holder changes the prediction outcome, on average, by 0.22%; and gender changes the prediction outcome, on average, by 0.13%. Additionally, through Figure 36, it is evident that the variables age, admission grade, previous qualification grade, GDP, and unemployment rate reduce the predicted probability score.

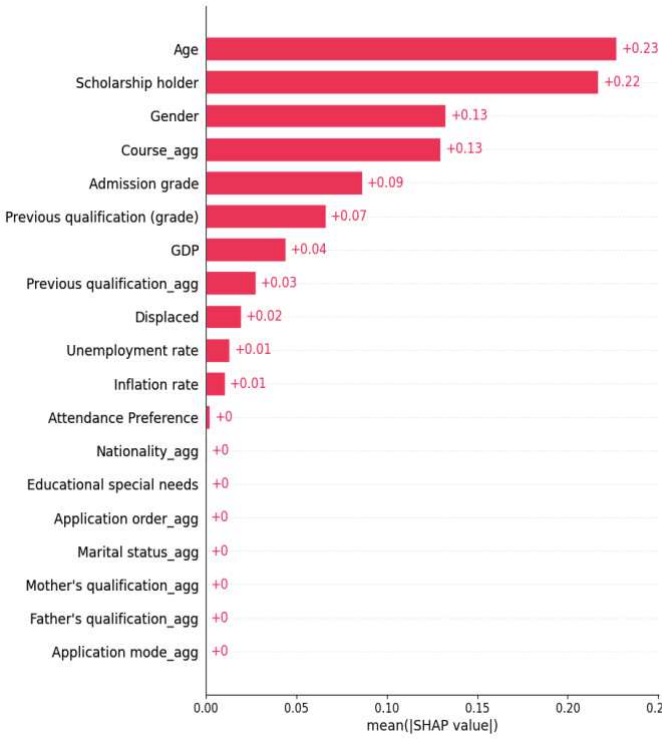

Figure 36: Global Feature Importance plot for Model B.

Figure 35: Global Magnitude Feature Importance plot for Model B.

Moving to a local overview, we can observe in Figure 37, a student with a high probability of success. Having a scholarship and being of a typical age to enter university, 18 years in this case, significantly increases this student's predicted probability of success in the course.

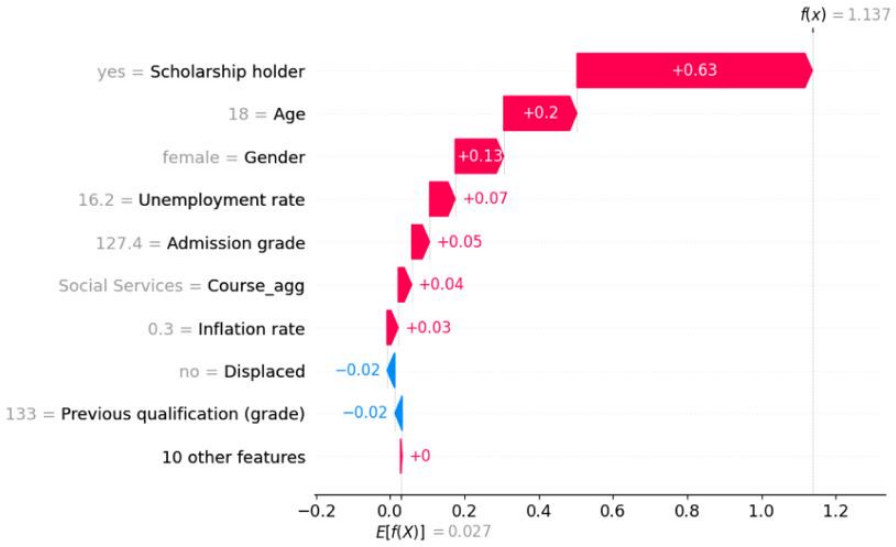

Figure 37: Local Feature Importance plot for Model B. The plot represents a student with high probability of success.

On the other hand, in Figure 38, a student with high probability of failure, having an older age, such as 27 years in this case, and not having a scholarship, significantly reduces this student's predicted probability of success in the course.

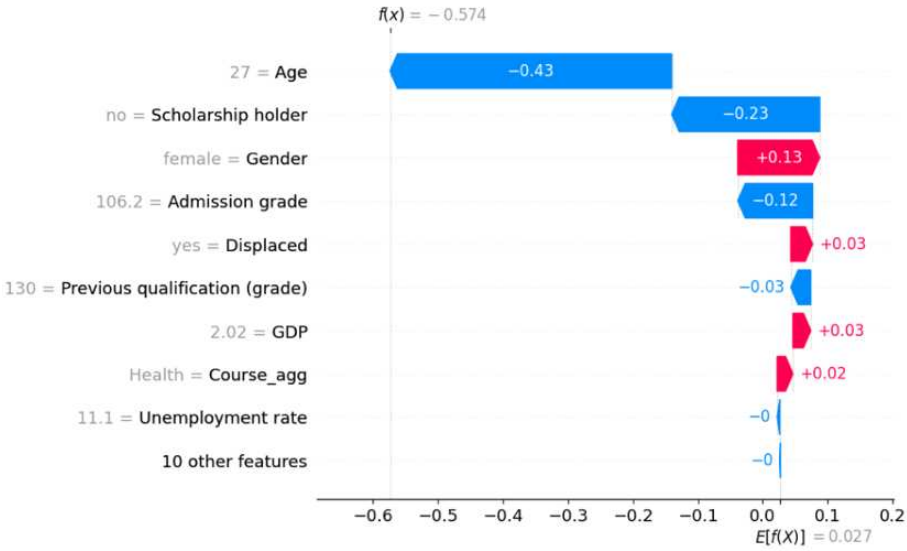

Figure 38: Local Feature Importance plot for Model B. The plot represents a student with high probability of failure.

Discussion

This thesis aims to predict student academic performance and the risk of students dropping out through the implementation of a black box machine learning model. Subsequently, an interpretability model will be applied, providing insights into general trends in student performance as well as individual level assessments for each student. This approach aims to create a transparent and reliable model applicable in real-world scenarios, enhancing the decision-making process and supporting the development of both general and personalized strategies for at risk students.

Firstly, the exploratory data analysis suggests that variables such as gender, age, marital status, previous qualification and grade before enrolling in the course, payment of fees up to date, the opportunity to receive a scholarship, and behaviour during the first two semesters influence the success or failure of a student. However, given the significant impact of the behaviour during the initial semesters, as it provides insights into the student's conduct, I have developed two models. One model is designed to identify at risk students after the first two semesters, utilizing information available before enrolment (Model A). The second model is applied at the student's entry, relying solely on information about the student's background and the behaviour of the previous qualification before entering the course (Model B). The goal is to intervene and provide support as early as possible for a student identified as at risk.

Following the preparation of necessary data and data splitting, the algorithm selected, CatBoost, demonstrated superior results and the significant advantage of automatically handling categorical features. This contrasts with the findings in the study by Martins et al. (2021), which concluded that Random Forest was among the best-performing algorithms using the same dataset. It's worth noting that their study did not incorporate information about students before entering the course, which might have influenced the results. After choosing the algorithm and analysing performance metrics, I observed that, for both models, - whether predicting student outcomes upon course entry or after one year - the rate of false positives (Type I error) was higher than false negatives (Type II error). This implies that the models incorrectly predict a student's success when, in reality, they might fail. As mentioned earlier, this misalignment with my objective is undesirable, as I prefer to assist a student who may not need help rather than miss aiding a student who truly requires support. Consequently, different approaches were employed to enhance the performance of both models. Model A achieved the best result using

the class weight parameter, giving more weight to class 0 (students who fail) than class 1 (successful students). In Model B, a threshold was implemented to increase the minimum probability of a success prediction.

Thus, in this thesis, the developed models were fine-tuned. Overall, based on the metrics used, both models demonstrate strong performance in prediction without showing signs of overfitting or underfitting. However, Model A outperforms Model B, possibly due to the fact that the characteristics a student exhibits during their first two semesters have a more significant impact on generalizing the behaviour the student may display until the end of the course.

Furthermore, SHAP values were employed to determine the magnitude and importance of each variable for the model, both globally (across the student community) and locally (for a specific student). Upon analysing the variables and their respective magnitudes globally, the results suggest that in the model incorporating all available student information after one year in the course, the probability of a student's success increases with a higher number of approved course units in the 1st and 2nd semesters. This aligns with previous findings emphasizing the significant influence of behaviour in first-level courses and the grade point average on students' academic success (Alhazmi, Sheneamer, and Sheneamer 2017; Thammasiri et al. 2014). Additionally, Kersanszki, Holik, and Sanda (2022) assert that the initial year, particularly the first semester, is a critical phase, and success during this period substantially reduces the risk of failure. Delen (2011) also states that over 50% of student attrition can be attributed to the first year of college. In contrast, concerning the model with only pre-course information, the results suggest that age, gender, and the possession of a scholarship impact the probability of a student's success in the course. It has been previously noted that academic failure not only depends on performance in coursework but is also influenced by factors such as vocational issues, economic constraints, and challenges in the integration process (Lopes, Pereira, and Vaz 2023).

In conclusion, at the individual level, the results suggest the possibility of obtaining a detailed perspective when needed. In Model A, the analysis identifies a student with a high probability of success in the course, showcasing, for instance, multiple successfully completed courses in the second semester. Conversely, the result for a student with a high probability of failure reveals a lack of approved curricular units in the first semester and only one in the second semester. In Model B, without academic information, the analysis identifies a student with a

high probability of success, indicating that having the typical university entry age (18 years) and receiving a scholarship influence this specific student's likelihood of success. On the other hand, when pinpointing a specific student with a high probability of failure, the results suggest that being above the average age (27 years) and lacking a scholarship increases the likelihood of failure for this student. These findings suggest the potential for targeted intervention and personalized support or strategies for a specific at risk student. This underscores the value added by explainability and interpretability, providing an advantage for stakeholders in the education sector (Chitti, Chitti, and Jayabalan 2020).

Limitations

While this thesis offered valuable insights into student performance within the higher education system, it was crucial to recognize certain limitations that might have affected the generalizability and robustness of the findings.

A notable constraint was the size of the dataset. The study relied on predicting student performance by categorizing them as successes and failures, utilizing a dataset with 4,424 instances specific to a university in Portugal. This specificity might have resulted in a homogeneous sample, limiting the generalizability of our conclusions to a more diverse population.

Another limitation concerned the characteristics of the dataset, as the majority of variables exhibited high cardinality. Each unique value of certain variables was labelled with specific names and titles from the institute in question. This practice might have led to a lack of understanding or inconsistency in data aggregation for categories. This was particularly evident in variables such as mother and father occupation and certain aspects of the curricular units due to insufficient information.

Throughout the study, it was crucial to acknowledge the presence of assumptions and simplifications inherent in any modelling approach. For example, the interpretation of variables was challenged by the incomplete data dictionary during data collection. Recognizing these assumptions underscored that they might not have fully captured the complexity of real world scenarios.

It was imperative to emphasize that recognizing these limitations signified a commitment to transparency and continuous improvement. Despite these challenges, the study yielded valuable insights into student performance within the higher education system. By addressing these limitations, future research could build upon and refine the contributions made in this study.

Conclusion

Recommendations for Implementation

The findings carry practical implications for the higher education system. By creating a machine learning model with interpretability, it was possible to conclude that it is possible to identify at risk students in advance, thereby preventing potential failures after one year of the course or even upon enrolment. These results and insights can be embraced by any higher education institution aiming to establish general or personalized strategies and support systems to reduce dropout rates.

Future Work

This thesis identifies specific limitations, presenting opportunities for future research.
To improve these results, additional information about the students' academic performance during the first academic semesters, such as class attendance, and details about the integration process, along with more interesting student characteristics, could also be included in the dataset. Efforts can be made to increase the dataset and gather information from students across different universities to achieve the greatest possible diversity. Investigating other machine learning algorithms and models of explainability contributes to refining and expanding the findings.

In summary, the thesis underscores the significance of integrating explanations into a model for predicting student performance. And it encourages professionals in the higher education system to incorporate these approaches. This integration is poised to enhance their outcomes, decrease dropout rates, and ultimately elevate the student experience by offering more personalized, targeted support, and strategies for each type of at risk student.

References

Alhazmi, Essa, Abdullah Sheneamer, and Abdullah M Sheneamer. 2017. "Early Predicting of Students Performance in Higher Education." https://doi.org/10.1109/ACCESS.2017.DO.

Beaulac, Cédric, and Jeffrey S. Rosenthal. 2019. "Predicting University Students' Academic Success and Major Using Random Forest," January.

Cerda, Patricio, and Gaël Varoquaux. 2019. "Encoding High-Cardinality String Categorical Variables," July. https://doi.org/10.1109/TKDE.2020.2992529.

Chitti, Manjari, Padmini Chitti, and Manoj Jayabalan. 2020. "Need for Interpretable Student Performance Prediction." In *Proceedings - International Conference on Developments in ESystems Engineering, DeSE*, 2020-December:269–72. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/DeSE51703.2020.9450735.

Delen, Dursun. 2011. "Predicting Student Attrition with Data Mining Methods." *Journal of College Student Retention: Research, Theory and Practice* 13 (1): 17–35. https://doi.org/10.2190/CS.13.1.b.

DGEEC, and Ministério da Educação. 2021. *Educação e Formação Em Portugal*. Direção-Geral de Estatísticas da Educação e Ciência (DGEEC) Ministério da Educação e Ministério da Ciência, Tecnologia e Ensino Superior.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning."

Esteves, Henrique Rosario Carvalho, Carlos Alberto Dias, Ciro Meneses Santos, and Agnaldo Keiti Higuchi. 2021. "Evasão Escolar No Ensino Superior: Uma Revisão Literária Entre Os Anos de 2014 a 2020." *Research, Society and Development* 10 (3): e21310313210. https://doi.org/10.33448/rsd-v10i3.13210.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases (© AAAI)." Vol. 17. www.ffly.com/.

Hoffait, Anne Sophie, and Michaël Schyns. 2017. "Early Detection of University Students with Potential Difficulties." *Decision Support Systems* 101 (September): 1–11. https://doi.org/10.1016/j.dss.2017.05.003.

INE. 2021. "CENSOS 2021-Resultados Definitivos."

Kersanszki, Tamas, Ildiko Holik, and Istvan Daniel Sanda. 2022. "Causes and Possibilities of Reducing Student Drop-out in Hungarian Technical Higher Education." In *INES 2022 - 26th IEEE International Conference on Intelligent Engineering Systems 2022, Proceedings*, 169–73. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/INES56734.2022.9922627.

Leal, Fátima, Bruno Veloso, Carla Santos Pereira, Fernando Moreira, Natércia Durão, and Natacha Jesus Silva. 2022. “Interpretable Success Prediction in Higher Education Institutions Using Pedagogical Surveys | Enhanced Reader.”

Lopes, Marina Isabel, Ana Paula Pereira, and Paula Marisa Vaz. 2023. “School Dropouts in Higher Education and Concurrent Factors.” https://doi.org/10.34620/eduser.v15i1.236.

Martins, Mónica V., Daniel Tolledo, Jorge Machado, Luís M. T. Baptista, and Valentim Realinho. 2021a. “Predict Students’ Dropout and Academic Success.” UCI Machine Learning Repository. 2021. https://doi.org/10.1007/978-3-030-72657-7_16.

Martins, Mónica V., Daniel Tolledo, Jorge Machado, Luís M.T. Baptista, and Valentim Realinho. 2021b. “Early Prediction of Student’s Performance in Higher Education: A Case Study.” In *Advances in Intelligent Systems and Computing*, 1365 AIST:166–75. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-72657-7_16.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. “A Survey on Bias and Fairness in Machine Learning,” August. http://arxiv.org/abs/1908.09635.

Miguéis, V. L., Ana Freitas, Paulo J.V. Garcia, and André Silva. 2018. “Early Segmentation of Students According to Their Academic Performance: A Predictive Modelling Approach.” *Decision Support Systems* 115 (November): 36–51. https://doi.org/10.1016/J.DSS.2018.09.001.

Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” *Artificial Intelligence*. Elsevier B.V. https://doi.org/10.1016/j.artint.2018.07.007.

Molnar, Christoph. 2023. “Interpretable Machine Learning A Guide for Making Black Box Models Explainable.” http://leanpub.com/interpretable-machine-learning.

Nadaf, Ali, Sebas Eliëns, and Xin Miao. 2021. “Interpretable-Machine-Learning Evidence for Importance and Optimum of Learning Time.” *International Journal of Information and Education Technology* 11 (10): 444–49. https://doi.org/10.18178/ijiet.2021.11.10.1548.

Onose, Ejiro. 2023. “Explainability and Auditability in ML: Definitions, Techniques, and Tools,” August. https://neptune.ai/blog/explainability-auditability-ml-definitions-techniques-tools.

Thammasiri, Dech, Dursun Delen, Phayung Meesad, and Nihat Kasap. 2014. “A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition.” *Expert Systems with Applications* 41 (2): 321–30. https://doi.org/10.1016/j.eswa.2013.07.046.

The CatBoost team, and Andrey Gulin. n.d. “Parameter Tuning.” 2023. Accessed December 26, 2023. https://catboost.ai/en/docs/concepts/parameter-tuning.

Tinto, Vincent. 1975. “Dropout from Higher Education: A Theoretical Synthesis of Recent Research.” *Review of Educational Research* 45 (1): 89. https://doi.org/10.2307/1170024.

Tinto, Vincent. 2010. “From Theory to Action: Exploring the Institutional Conditions for Student Retention.” In , 51–89. https://doi.org/10.1007/978-90-481-8598-6_2.

U-World Blog. 2021. “9 Factos Interessantes Sobre o Ensino Universitário Em Portugal.” U-World Blog. October 19, 2021. https://u-world.pt/blog/9-factos-ensino-universitario-portugal/.

Wang, Ji, and Qinglong Zhan. 2021. “Visualization Analysis of Artificial Intelligence Technology in Higher Education Based on SSCI and SCI Journals from 2009 to 2019.” *International Journal of Emerging Technologies in Learning* 16 (8): 20–33. https://doi.org/10.3991/ijet.v16i08.18447.

Yandex. 2023. “CatBoost Website.” 2023. https://catboost.ai.

Appendix

Table S1: Data Dictionary of the original dataset. Identification and description of the variables and data type.

Variable	Description
Marital Status*[1]	The marital status of the student. [Categorical]
Application Mode*[2]	The method of application used by the student. [Categorical]
Application Order*[3]	The order in which the student applied. [Numerical]
Course*[4]	The course taken by the student. [Categorical]
Daytime/evening attendance*[5]	Whether the student attends classes during the day or in the evening. [Categorical]
Previous Qualification*[6]	The qualification obtained by the student before enrolling in higher education. [Categorical]
Previous Qualification (grade)*[7]	The grade obtained by the student before enrolling in higher education. [Numerical]
Nationality*[8]	The nationality of the student. [Categorical]
Mother's Qualification*[9]	The qualification of the student's mother. [Categorical]
Father's Qualification*[10]	The qualification of the student's father. [Categorical]
Mother's Occupation*[11]	The occupation of the student's mother. [Categorical]
Father's Occupation*[12]	The occupation of the student's father. [Categorical]
Admission grade*[13]	Represents the cumulative academic qualification or score used by higher education institutions to assess an applicant's overall academic suitability. [Numerical]
Displaced*[14]	Whether the student is a displaced person. [Categorical]
Educational special needs*[15]	Whether the student has any special educational needs. [Categorical]
Debtor*[16]	Whether the student is a debtor. [Categorical]
Tuition fees up to date*[17]	Whether the student's tuition fees are up to date. [Categorical]
Gender*[18]	The gender of the student. [Categorical]
Scholarship holder*[19]	Whether the student is a scholarship holder. [Categorical]
Age at enrolment	The age of the student at the time of enrolment.

International*[20]	Whether the student is an international student. [Categorical]
Curricular units 1st sem (credited)	The number of curricular units credited by the student in the first semester. [Numerical]
Curricular units 1st sem (enrolled)	The number of curricular units enrolled by the student in the first semester. [Numerical]
Curricular units 1st sem (evaluations)	The number of curricular units evaluated by the student in the first semester. [Numerical]
Curricular units 1st sem (approved)	The number of curricular units approved by the student in the first semester. [Numerical]
Curricular units 1st sem (grade)*[21]	The average grade of curricular units in the first semester. [Numerical]
Curricular units 1st sem (without evaluations)	The number of curricular units without evaluations in the first semester. [Numerical]
Curricular units 2nd sem (credited)	The number of curricular units credited by the student in the second semester. [Numerical]
Curricular units 2nd sem (enrolled)	The number of curricular units enrolled by the student in the second semester. [Numerical]
Curricular units 2nd sem (evaluations)	The number of curricular units evaluated by the student in the second semester. [Numerical]
Curricular units 2nd sem (approved)	The number of curricular units approved by the student in the second semester. [Numerical]
Curricular units 2nd sem (grade)*[22]	The average grade of curricular units in the second semester. [Numerical]
Curricular units 2nd sem (without evaluations)	The number of curricular units without evaluations in the second semester. [Numerical]
Unemployment rate	From the region. [Numerical]
Inflation rate	From the region. [Numerical]
GDP	From the region. [Numerical]
Target	Enrolled, Graduate, Dropout. [Categorical]

*[1]: 1 – single; 2 – married; 3 – widower; 4 – divorced; 5 – facto union; 6 – legally separated.

*[2]: 1 - 1st phase - general contingent; 2 - Ordinance No. 612/93; 5 - 1st phase - special contingent (Azores Island); 7 - Holders of other higher courses; 10 - Ordinance No. 854-B/99; 15 - International student (bachelor); 16 - 1st phase - special contingent (Madeira Island); 17 - 2nd phase - general contingent; 18 - 3rd phase - general contingent; 26 - Ordinance No. 533-

A/99, item b2) (Different Plan); 27 - Ordinance No. 533-A/99, item b3 (Other Institution); 39 - Over 23 years old; 42 – Transfer; 43 - Change of course; 44 - Technological specialization diploma holders; 51 - Change of institution/course; 53 - Short cycle diploma holders; 57 - Change of institution/course (International).

*[3]: between 0 - first choice; and 9 last choice.

*[4]: 33 - Biofuel Production Technologies; 171 - Animation and Multimedia Design; 8014 - Social Service (evening attendance); 9003 – Agronomy; 9070 - Communication Design; 9085 - Veterinary Nursing; 9119 - Informatics Engineering; 9130 – Equinculture; 9147 – Management; 9238 - Social Service; 9254 – Tourism; 9500 – Nursing; 9556 - Oral Hygiene; 9670 - Advertising and Marketing Management; 9773 - Journalism and Communication; 9853 - Basic Education; 9991 - Management (evening attendance).

*[5]: 1 – daytime; 0 – evening.

*[6]: 1 - Secondary education; 2 - Higher education - bachelor's degree; 3 - Higher education – degree; 4 - Higher education - master's; 5 - Higher education – doctorate; 6 - Frequency of higher education; 9 - 12th year of schooling - not completed; 10 - 11th year of schooling - not completed; 12 - Other - 11th year of schooling; 14 - 10th year of schooling; 15 - 10th year of schooling - not completed; 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv.; 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv.; 39 - Technological specialization course; 40 - Higher education - degree (1st cycle); 42 - Professional higher technical course; 43 - Higher education - master (2nd cycle).

*[7]: Between 0 and 200.

*[8]: 1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26 - Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 – Colombian.

*[9] e *[10]: 1 - Secondary Education - 12th Year of Schooling or Eq.; 2 - Higher Education - Bachelor's Degree; 3 - Higher Education – Degree; 4 - Higher Education - Master's; 5 - Higher Education – Doctorate; 6 - Frequency of Higher Education; 9 - 12th Year of Schooling - Not Completed; 10 - 11th Year of Schooling - Not Completed; 11 - 7th Year (Old); 12 - Other - 11th Year of Schooling; 13 - 2nd year complementary high school course; 14 - 10th Year of Schooling; 18 - General commerce course; 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv.; 20 - Complementary High School Course; 22 - Technical-professional course; 25 - Complementary High School Course - not concluded; 26 - 7th year of schooling; 27 - 2nd cycle of the general high school course; 29 - 9th Year of Schooling - Not Completed; 30 - 8th year of

schooling; 31 - General Course of Administration and Commerce; 33 - Supplementary Accounting and Administration; 34 – Unknown; 35 - Can't read or write; 36 - Can read without having a 4th year of schooling; 37 - Basic education 1st cycle (4th/5th year) or equiv.; 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv.; 39 - Technological specialization course; 40 - Higher education - degree (1st cycle); 41 - Specialized higher studies course; 42 - Professional higher technical course; 43 - Higher Education - Master (2nd cycle); 44 - Higher Education - Doctorate (3rd cycle).

*[11] e *[12]: 0 – Student; 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers; 2 - Specialists in Intellectual and Scientific Activities; 3 - Intermediate Level Technicians and Professions; 4 - Administrative staff; 5 - Personal Services, Security and Safety Workers and Sellers; 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry; 7 - Skilled Workers in Industry, Construction and Craftsmen; 8 - Installation and Machine Operators and Assembly Workers; 9 - Unskilled Workers; 10 - Armed Forces Professions; 90 - Other Situation; 99 - (blank); 101 - Armed Forces Officers; 102 - Armed Forces Sergeants; 103 - Other Armed Forces personnel; 112 - Directors of administrative and commercial services; 114 - Hotel, catering, trade and other services directors; 121 - Specialists in the physical sciences, mathematics, engineering and related techniques; 122 - Health professionals; 123 – teachers; 124 - Specialists in finance, accounting, administrative organization, public and commercial relations; 125 - Specialists in information and communication technologies (ICT); 131 - Intermediate level science and engineering technicians and professions; 132 - Technicians and professionals, of intermediate level of health; 134 - Intermediate level technicians from legal, social, sports, cultural and similar services; 135 - Information and communication technology technicians; 141 - Office workers, secretaries in general and data processing operators; 143 - Data, accounting, statistical, financial services and registry-related operators; 144 - Other administrative support staff; 151 - personal service workers; 152 – sellers; 153 - Personal care workers and the like; 154 - Protection and security services personnel; 161 - Market-oriented farmers and skilled agricultural and animal production workers; 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence; 171 - Skilled construction workers and the like, except electricians; 172 - Skilled workers in metallurgy, metalworking and similar; 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like; 174 - Skilled workers in electricity and electronics; 175 - Workers in food processing, woodworking, clothing and other industries and crafts; 181 - Fixed plant and machine operators; 182 - assembly workers; 183 - Vehicle drivers and mobile equipment operators; 191 - cleaning workers; 192 - Unskilled workers in

agriculture, animal production, fisheries and forestry; 193 - Unskilled workers in extractive industry, construction, manufacturing and transport; 194 - Meal preparation assistants; 195 - Street vendors (except food) and street service providers.

*[13]: Between 0 and 200.

*[14] and *[15] and *[16] and *[17] and *[19] and *[20]: 1 – yes; 0 – no.

*[18]: 1 – male; 0 – female.

*[21] e *[22]: between 0 and 200.

Table S2: Combination of Parameters. Identification of the parameters used for the grid search.

Hyperparameter	**Parameter Grid**
Iterations (iterations)	Np.arrange (30,60,10)
Depth (depth)	Np.arrange (5,10,1)
Learning Rate (learning_rate)	[0.01, 0.1]
Regularization (l2_leaf_reg)	[3,5]
Border (border_count)	Np.arrange (10,80,10)