UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Dual Predictions in Higher Education: Applying Machine Learning to Student Dropout and Enrollment Forecasting

André Nicolás Sarmiento Acuña

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero Industrial

Quito, 2 de diciembre de 2024

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Dual Predictions in Higher Education: Applying Machine Learning to Student Dropout and Enrollment Forecasting

André Nicolás Sarmiento Acuña

Nombre del profesor, Título académico **Danny Orlando Navarrete Chávez, MSc.**

Quito, 2 de diciembre de 2024

© DERECHOS DE AUTOR

Nombres y apellidos: André Nicolás Sarmiento Acuña

Código: 00216220

Cédula de identidad: 0105943542

Lugar y fecha: Quito, 2 de diciembre de 2024

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Esta investigación examina la modelización predictiva para abordar los retos de la deserción estudiantil y la matriculación en las instituciones de educación superior (IES), utilizando datos de una IES ecuatoriana. Se analizan 2.464 estudiantes para la predicción de abandono y 2.671 solicitantes para la predicción de matrícula, cubriendo varios años académicos e incorporando variables académicas, demográficas y económicas. Guiado por la metodología CRISP-DM, el estudio emplea algoritmos de aprendizaje automático como Regresión Logística, Máquinas de Vectores Soporte, Random Forest, XGBoost y Redes Neuronales Artificiales.

El desequilibrio de clases se gestiona mediante técnicas como SMOTE y Random Under-Sampling, que garantizan conjuntos de datos equilibrados. XGBoost combinado con SMOTE logró la mayor precisión en la predicción del abandono (96%), mientras que la regresión logística destacó en la predicción de la matriculación (98%). Los principales factores de predicción del abandono son las ayudas económicas, el rendimiento académico y el programa de estudios, mientras que los factores de predicción de la matriculación se centran en la edad, el año de admisión, las calificaciones de los exámenes y el coste de la matrícula. El estudio subraya la importancia de la interpretabilidad de los modelos para obtener información práctica que ayude a las IES a tomar decisiones estratégicas para mejorar la retención y la captación de estudiantes.

Palabras clave: Abandono estudiantil, Matriculación estudiantil, Instituciones de enseñanza superior, Aprendizaje automático, Modelización predictiva, Metodología CRISP-DM, Minería de datos, Logística, Regresión, XGBoost, SMOTE, Preparación de datos, Interpretabilidad de modelos.

ABSTRACT

This research examines predictive modeling to address student dropout and enrollment challenges in higher education institutions (HEIs), using data from an Ecuadorian HEI. It analyzes 2,464 students for dropout prediction and 2,671 applicants for enrollment forecasting, covering various academic years and incorporating academic, demographic, and economic variables. Guided by the CRISP-DM methodology, the study employs machine learning algorithms like Logistic Regression, Support Vector Machines, Random Forest, XGBoost, and Artificial Neural Networks.

Class imbalance is managed through techniques like SMOTE and Random Under-Sampling, ensuring balanced datasets. XGBoost combined with SMOTE achieved the highest dropout prediction accuracy (96%), while Logistic Regression excelled in enrollment prediction (98%). Key dropout predictors include financial aid, academic performance, and program of study, while enrollment predictors focus on age, admission year, exam scores, and tuition costs. The study highlights the importance of model interpretability for actionable insights, supporting strategic decision-making in HEIs to enhance student retention and recruitment.

Key words: Student dropout, Student enrollment, Higher education institutions, Machine learning, Predictive modeling, CRISP-DM methodology, Data mining, Logistic, Regression, XGBoost, SMOTE, Data preparation, Model interpretability.

TABLE OF CONTENTS

Introduction **12**

Literature review **13**

Concepts **13**

Dropout 13

Enrollment 14

Approach to the problem **15**

Machine Learning **16**

Supervised learning, unsuperviserd learning and reinforcement learning 16

Machine Learningn Algorithms 17

Determining factors, models and interpretability **18**

Dropout 18

Enrollment 21

Development of the Subject **23**

Methodology **23**

CRISP – DM 23

Evaluation Metrics **27**

Performance Measure 27

Modeling **29**

Unbalanced dataset 29

Hyperparameters 30

Grid Search 30

Principal component analysis 31

Interpretability 31

Research Procedure **31**

Difficult of gathering the data **31**

Dropout **32**

Data Understanding 32

Data Value Exploration 32

Data Preparation 33

Enrollment **33**

Data Understanding 33

Data Value Exploration 34

Data Preparation 34

Modeling **34**

Results and discussion **35**

Dropout dataset **35**

Enrolment dataset **38**

Conclusions **40**

Bibliographic References **42**

Appendix A: Categories Dropouts **62**

Appendix B: Venn diagram of machine learning concepts and classes **63**

Appendix C: ML, DL, ANN models used for Dropout **64**

Appendix D: ML, DL, ANN models used for Enrollment **66**
Appendix E: CRISP – DM process model **67**
Appendix F: CRISP – DM process model of data mining **68**
Appendix G: Trajectory through a data science project **69**
Appendix H: CRISP-DM goal-directed **70**
Appendix I: CRISP-DM goal-directed in the study **71**
Appendix J: supervised machine learning algorithm **73**
Appendix K: LOGISTIC REGRESSION **86**
Appendix L: Support Vector Machine **88**
Appendix M: random Forest **90**
Appendix N: XGBOOST **92**
Appendix O: Feedforward Neural Network (FNN) **93**
Appendix P: List of variables and description **95**
Appendix Q: Drop out Data Exploration **97**
Appendix R: drop out data cleaning process **99**
Appendix S: List of enrollment variables and description **104**
Appendix T: enrollment Data Exploration **106**
Appendix U: Enrollment data cleaning process **108**
Appendix V: Hyperparameters **111**

TABLE INDEX

Table 1
Logistic Regression. 35
Table 2
Support Vector Machine. 36
Table 3
Random Forest 36
Table 4
XGBoost. 36
Table 5
Feedforward Neural Network 36
Table 6
Selection of the best fitting model 37
Table 7
Importance of the principal components of the best-fitting model 37
Table 8
Importance of the variables of the best-fitting model 37
Table 9
Logistic Regression. 38
Table 10
Support Vector Machine. 38
Table 11
Random Forest 38
Table 12
XGBoost. 39
Table 13
Feedforward Neural Network 39
Table 14
Selection of the best fitting model 39
Table 15
Importance of the variables of the best-fitting model 39
Table 16
Categories used for the classification of student dropout factors 62
Table 17
ML, DL, ANN models, types of data used and best model performance for dropout 64
Table 18
ML, DL, ANN models, types of data used and best model performance for enrollment 66
Table 19
CRISP-DM process model descriptions (Wirth & Hipp, 2000) 67
Table 20
CRISP-DM goal-directed (Martinez-Plumed et al., 2021) 70
Table 21
CRISP-DM goal-directed based on (Martinez-Plumed et al., 2021) for this study 71
Table 22
Brief description of supervised machine learning models 73

Table 23
List of all variables provided by the HEI with their detailed description. 95
Table 24
Data cleaning process.....99
Table 25
List of all variables provided by the HEI with their detailed description. 104
Table 26
Data cleaning process.....108
Table 27
Model's Hyperparameters.....111

INDEX OF FIGURES

Figure 1
Venn diagram of machine learning concepts and classes (Janiesch et al., 2021) 63
Figure 2
The CRISP-DM process model of data mining (Martinez-Plumed et al., 2021).....68
Figure 3
Trajectory through a data science project (Martinez-Plumed et al., 2021).....69
Figure 4
Unit-step function and logistic function (Z. H. Zhou, 2021). 86
Figure 5
A and B illustrate the principle of the maximum-margin classifier. C and D demonstrate the introduction of the slack variable, which allows the support vector classifier to maximize its margin while disregarding the influence of nearby observations, even when the data is non-separable (Valkenborg et al., 2023).....88
Figure 6
Random Forest (Biau & Scornet, 2016).....90
Figure 7
Boosting - sequential ensemble learning (Ferreira et al., 2021).....92
Figure 8
Forward Propagation (X. Zhou et al., 2022). 93
Figure 9
Percentage of students who drop out.....97
Figure 10
Students by major.....97
Figure 11
Dropout rate by major. 98
Figure 12
Relationship between student's GPA and credits taken by the student. 98
Figure 13
Percentage of students who enroll the HEI. 106
Figure 14
Percentage of enrollment by major 106
Figure 15
Relationship between the grade point average of the school and the grade achieved in the admission exam. 106

INTRODUCTION

Student dropout and enrollment are two pivotal challenges confronting higher education institutions (HEIs), with profound implications for institutional performance and student success. Dropout rates reflect not only the academic and social integration of students but also broader socioeconomic factors. Similarly, enrollment trends are shaped by complex dynamics, including financial constraints, institutional reputation, and students' career aspirations. Understanding these factors and predicting their outcomes is vital for HEIs to optimize resource allocation, develop targeted interventions, and enhance overall student engagement. This study leverages machine learning techniques to address these challenges. Traditional statistical methods have provided foundational insights, but the advent of data-driven approaches, such as machine learning, allows for more nuanced and accurate predictions. By applying models like Logistic Regression, Random Forest, and XGBoost, this research aims to predict student dropout and enrollment patterns effectively. The study also emphasizes the interpretability of these models, ensuring that the insights derived can inform actionable strategies. The research follows the CRISP-DM methodology, a structured approach that guides the entire data mining process, from understanding the business problem to deploying predictive models.

LITERATURE REVIEW

Concepts

Dropout

To comprehensively understand the phenomenon of student dropout and the factors contributing to it, it is crucial to first define the concept of dropout, particularly in the context of Higher Education Institutions (HEIs). According to Kehm et al., (2020), student attrition can be understood as the difference between retention rates and graduation rates, providing a quantitative measure of dropout. However, Søgaard et al., (2013) argue that the term "student dropout" is frequently used to describe the situation in which a student withdraws from university without completing an academic degree. This concept, however, is multifaceted and open to various interpretations.

In many cases, terms such as "*withdraw*", "*fail*", "*incomplete*", or "*not complete*" are used as synonyms for student dropout (Behr et al., 2020). The term "*withdraw*" generally suggests a voluntary dropout. This decision is a disadvantageous action motivated by various causes, such as transfer to another university, career changes, attractive job offers, personal or family economic problems, as well as personal problems (Nicoletti, 2019). On the other hand, the terms "*failing*", "*incomplete*" or "*not finishing*" are more associated with involuntary dropout, classified as "*non-voluntary*" (Nicoletti, 2019). This type of dropout may be related to insufficient academic integration, which manifests itself, for example, in obtaining low grades or in the perception that the HEI does not meet the student's academic expectations.

There are various theoretical perspectives on student dropout, from which several theoretical models have been developed for its study. These theoretical models are divided into four main orientations: sociological, psychological, economic and phase models (Behr et al.,

2020; Guzmán et al., 2021). Sociological models emphasize the importance of student social and academic integration (Nordmann et al., 2019). Psychological models focus on student behavior and attitudes in the dropout process (Korhonen et al., 2019). Economic models emphasize rational and logical decision making, considering the cost-benefit analysis of leaving HEI (Jüttler, 2020). Finally, phase models combine two or more of the above orientations (Aina et al., 2022).

It is important to note that the aforementioned authors classify the factors according to their own criteria, since there is no standardization in this field. Likewise, the same factor may be present in two or more categories, depending on the focus of the study. The categories used by these authors to classify the factors leading to student dropout are presented in Table 16, which provides an overview of the estudies between 2020 and 2024 as can be seen in Appendix A.

Enrollment

The landscape of higher education institutions (HEIs) is undergoing significant transformation due to factors such as heightened competition among institutions, rapid economic changes, and evolving demands for professional skills. This dynamic environment presents considerable challenges for students, particularly those in their final year of secondary education, as they navigate the complex process of selecting a suitable HEI (Ab Ghani et al., 2019). Traditionally, universities have employed two primary methods to manage student enrollment: (1) direct communication with prospective students to confirm their intent to enroll, and (2) verification of enrollment through payment status checks . However, these conventional approaches are increasingly inadequate in addressing the complexities of modern student recruitment and enrollment (L. Yang et al., 2021). A limited number of HEIs have

shown interest in delving deeper into the motivations behind students' choices to pursue higher education and in forecasting enrollment numbers. Gaining insights into these motivations enables institutions to make strategic decisions, such as introducing new academic programs and optimizing the allocation of internal resources, thereby enhancing their appeal to prospective students (Bousnguar et al., 2022).

Student commitment within an educational environment encompasses a combination of cognitive, emotional, and behavioral components (Fredricks et al., 2004). Cognitive engagement refers to the student's efforts toward their education, including goal-setting and belief in the value of the HEI. Emotional engagement involves the student's sense of belonging and identification with the HEI, while behavioral engagement pertains to actions such as attendance and adherence to institutional rules (Christenson & Reschly, 2008; Reschly & Christenson, 2012). Academic engagement has been consistently identified as a critical factor in academic success, influencing students from elementary and middle school through to their selection of postsecondary education (Abbott-Chapman et al., 2014).

Approach to the problem

Early studies have used traditional statistical methods to help institutions understand the reasons behind student dropouts and identify those at risk of leaving (Berger & Milem, 1999; Forsman et al., 2015). Similarly, these methods have been applied to understand why applicants choose to enroll and to predict which applicants may be at risk of not enrolling (Mayer-Foulkes, 2002; Popov, 2019). Numerous studies have developed statistical models to identify the causal factors influencing both student dropout and enrollment (Burtner, 2005; Lee & Choi, 2013). Beyond the traditional literature on the topic, recent years have seen a growing

emphasis on analytical studies employing machine learning and data mining techniques (Álvarez-Pérez et al., 2024; Diaz Lema et al., 2024; Loder, 2024a; Oztekin, 2016).

Machine Learning

Machine Learning refers to the improvement of a computer program's performance over time as it gains experience with respect to specific tasks or performance measures (Jordan & Mitchell, 2015). The primary goal is to automate the process of building analytical models capable of performing cognitive tasks such as object detection, entity classification, or natural language translation (Z. H. Zhou, 2021). This is achieved through the application of algorithms that iteratively learn from training data related to the problem at hand (Bishop & Nasrabadi, 2006). Machine learning is particularly effective for tasks involving high-dimensional data, such as classification, regression, and clustering, due to its ability to learn from past computations and extract patterns from large datasets, this enables the generation of reliable and repeatable decisions (Janiesch et al., 2021). Depending on the problem and the available data, there are three types of Machine Learning: (1) supervised learning, (2) unsuperviserd learning and (3) reinforcement learning (Morales & Escalante, 2022).

Supervised learning, unsuperviserd learning and reinforcement learning

Supervised learning requires a training dataset that includes both input examples and corresponding labeled responses or target values for the output. The input-output pairs from the training set are used to calibrate the open parameters of the machine learning model (Morales & Escalante, 2022). Once the model has been successfully trained, it can predict the target variable from new or unseen input features (X). Within supervised learning, two types of problems are commonly distinguished: regression, where the goal is to predict a numerical

value, and classification, which involves categorizing input data into predefined classes (Tiwari, 2022).

Unsupervised learning occurs when the learning system is tasked with detecting patterns without the guidance of pre-existing labels or specifications (Morales & Escalante, 2022). In this approach, the training data consist solely of input variables (X), and the objective is to discover structural information of interest. This may involve identifying groups of elements that share common properties (a process known as clustering) or generating data representations that project from a high-dimensional space to a lower-dimensional space (referred to as dimensionality reduction) (Naeem et al., 2023).

In a *reinforcement learning* system, instead of providing input-output pairs, the current state of the system is described, a goal is specified, and a list of allowed actions along with their environmental constraints is provided. The machine learning model then learns through its own experience by applying the trial-and-error principle, aiming to reach the goal by maximizing a reward (Morales & Escalante, 2022; Silver et al., 2018).

Machine Learningn Algorithms

Depending on the learning task, the field offers several categories of algorithms, which can be broadly classified into machine learning, artificial neural networks, and deep neural networks, each with multiple specifications and variants, as can be seen in figure 1 in appendix B (Janiesch et al., 2021).

The family of *artificial neural networks* is of particular interest due to its flexible structure, which allows for adaptation across a wide range of contexts within all three types of machine learning. Inspired by the principles of information processing in biological systems, artificial neural networks consist of mathematical representations of interconnected processing

units known as artificial neurons. Similar to synapses in the brain, each connection between neurons transmits signals, the strength of which can be amplified or attenuated by a weight that is continuously adjusted throughout the learning process (Abdolrasol, Suhail Hussain, et al., 2021).

Deep neural networks typically consist of multiple hidden layers, arranged in deeply nested network architectures. In contrast to simpler artificial neural networks, deep neural networks often incorporate more advanced neurons. These neurons can perform complex operations or utilize multiple activation functions, rather than relying on a single activation function (Borisov et al., 2024).

Determining factors, models and interpretability

Dropout

Phan et al., (2023) identify three key aspects fundamental to the development of predictive models for understanding student dropout. These include: (1) the input data, (2) the models employed, and (3) the interpretability of the predictions generated by these models. This study rovides an overview of the literature reviewed from 2020 to 2024 concerning these three dimensions.

The *input data* typically encompass several categories, as outlined above, and are generally derived from structured data sources collected by HEIs. Common categories within these data sets include academic performance, student background, and socioeconomic status (Behr et al., 2020). The *Academic Performance* category includes detailed academic information such as the student's identification, the program in which they are enrolled, the year of entry into the university, any transfers from other institutions, entrance exam scores, subjects taken, highest grades obtained in those subjects, overall grade point average (GPA),

admission format, among other relevant details (Delen et al., 2020; Kemper et al., 2020; Martins et al., 2023; Santos et al., 2024). This data is collected throughout all semesters in which the student remains active at the HEI. Delen et al., (2020) found that students with a GPA equivalent to an A have a 7.3% probability of dropping out, while those with a GPA equivalent to an F face a dropout probability of 87.8%. In a comparative study, Diaz Lema et al., (2024) determined that students who earn fewer than 10 academic credits in their first semester, and fewer than 40 credits in their second semester, have a higher likelihood of dropping out. Similarly, Kemper et al., (2020) reported that the likelihood of a student deciding to drop out increases the longer they delay the decision, with the highest dropout rates occurring during the 1st and 2nd semesters. The *Student Background* category includes sociodemographic information such as age, gender, province of origin, marital status, ethnicity, family situation, and employment status . This data is typically collected when a student first enters the HEI and generally remains stable throughout their academic journey (Diaz Lema et al., 2024; Phan et al., 2023; Realinho et al., 2022; Segura et al., 2022). Santos et al., (2024) found that the average dropout rate for women is 8.81% higher than for men, taking into account the geographic region in which they reside. Matz et al. (2023) demonstrated that ethnicity is a significant factor, with students of African descent showing a higher likelihood of dropout. Additionally, Kemper et al., (2020)concluded that older students tend to exhibit higher dropout rates. The *Economic Situation* category includes socioeconomic information such as access to financial aid, scholarships, the percentage of scholarship coverage, and the economic conditions of both the family and the country (Barramuño et al., 2022; Matz et al., 2023; Olaya et al., 2020; Realinho et al., 2022). This data is collected by the HEI throughout all semesters during which the student remains active. Realinho et al., (2022) assert that the

availability of financial aid, scholarships, and family income levels significantly influence the likelihood of student dropout.

The *models employed* shows that Random Forest emerges as the most frequently used model, appearing in 13.4% of the reviewed studies, followed closely by Logistic Regression at 12.4%. It is important to note that Logistic Regression has demonstrated the most robust predictive performance in several cases. In particular, models such as Artificial Neural Networks (ANN), Bayesian Optimization, Random Forest, and Support Vector Machines (SVM) have achieved high levels of accuracy when predicting student dropout using datasets collected from student surveys over a 10-year period, with reported accuracies of 99% and 93% (Jiménez-Gutiérrez et al., 2024). Ensemble models, which combine multiple algorithms, tend to outperform individual models. For example, an ensemble model comprising Gradient Boosting, Random Forest, and SVM achieved accuracy rates ranging from 90% to 88%, whereas individual models exhibited accuracy rates between 79% and 88%. It is important to consider that the datasets used for these tests were assembled for each evaluation, and four different datasets were generated (Fernandez-Garcia et al., 2021). Notably, many previous studies have grouped students into a single, homogeneous cohort, predicting dropout behavior on a global level, rather than segmenting students into distinct groups for more targeted predictions. The various models used by authors in predicting student dropout are summarized in Table 17, which provides an overview of the literature reviewed between 2020 and 2024 as can be senn in Appendix C.

The *interpretability* of prediction models is a critical factor in making informed decisions for managing student dropout, particularly from the perspective of data scientists. This interpretability is not inherently tied to a mathematical formula, but rather depends on the ability of human beings to understand and contextualize the algorithm's recommendations.

Interpretability is defined as the extent to which a human can comprehend the rationale behind an algorithm's decision and can be categorized into two types: global interpretability and segmental interpretability (Phan et al., 2023). It is important to highlight that many prior studies have treated students as a single, homogeneous cohort, making predictions at a global level. However, this approach may overlook the potential for more accurate predictions through segmentation, where students are grouped into distinct categories, allowing for more targeted and tailored dropout predictions.

Enrollment

In this literature review on student enrollment, we adopt the three key aspects outlined by Phan et al., (2023): (1) input data, (2) the models employed, and (3) the interpretability of the predictions generated by these models. This study presents an overview of the literature reviewed from 2019 to 2023 concerning these three dimensions.

Input data generally encompass multiple categories, as previously mentioned, and are typically derived from structured data sources collected by HEIs. Common categories within these datasets include personal status, financial status, and institutional status (Ab Ghani et al., 2019). The *personal status* category inclufes detailed information about the student such as age, gender, province of origin, ethnicity and family situation. This data is typically collected when a student first enters the HEI (Fernández-García et al., 2020; Fraysier et al., 2020; Ujkani et al., 2021). Fraysier et al., (2020) found that the average enorollment rate for women is 2.5% lower than for men, taking into account the geographic region in which they reside. The *financial status* category includes socioeconomic information such as access to financial aid, scholarships, and the economic conditions of both the family and the country. . This data is typically collected when a student first enters the HEI (Akmanchi et al., 2023; Goldhaber et

al., 2019; Nita et al., 2022). The more financial aid, scholarship percentage or economic stability the family has, the higher the probability that the student will enroll (Nita et al., 2022). Te *institutional statuts* category includes includes detailed academic information such as the program in which they want to be enrolled, the year of entry into the university, entrance exam scores, high school overall grade point average (GPA), admission format, among other relevant details. This data is collected when the student applies to the HEI (Ab Ghani et al., 2019; Alyahyan & Düştegör, 2020; Goldhaber et al., 2019). The probability of a student enrolling in a HEI increases with a higher high school GPA and a stronger performance on the entrance exam (Waldrop et al., 2019).

The *algorithms employed* for predicting student enrollment have varied across studies, with certain models demonstrating higher predictive accuracy than others. The literature review conducted in this study identifies Random Forest and Decision Trees as the most frequently used algorithms, both of which have consistently shown robust performance in several cases. In particular, Decision Trees, Random Forest, and BP Neural Networks were applied to datasets containing information from prospective HEI applicants. The data used for training spanned three years, while data from one additional year was used for testing. Among these, Random Forest emerged as the superior model, achieving an accuracy of 62.78% (S. Yang et al., 2020). In another study, despite the use of multiple algorithms, including Random Forest, Gradient Boosting Classifier, Logistic Regression, Support Vector Machines, k-Nearest Neighbors, and Multilayer Perceptron, Random Forest continued to outperform the others, yielding an accuracy rate of 81.89% (Fernández-García et al., 2020). Furthermore, when structured databases with 10-fold cross-validation were utilized, Decision Trees demonstrated exceptional robustness, achieving an accuracy of 90.2% (Ujkani et al., 2021). Table 18

provides a summary of the predictive models used in the literature between 2019 and 2023, as detailed in Appendix D.

As with student dropout (Phan et al., 2023) rediction, the interpretability of models used for predicting student enrollment is not solely reliant on mathematical formulas but rather on the capacity of individuals to comprehend and contextualize the model's recommendations . This interpretability is crucial for ensuring that the insights generated by the models can be effectively applied in decision-making processes. Additionally, it is important to highlight that most studies treat students as a single homogeneous cohort, making predictions at a global level. This approach, while common, may overlook the potential benefits of segmenting students into distinct groups, which could lead to more accurate and targeted predictions.

DEVELOPMENT OF THE SUBJECT

Methodology

CRISP – DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is an industry-independent process model designed to guide data mining activities. This method, widely recognized for its effectiveness, has two key aspects: (1) as a methodology, it provides detailed descriptions of the phases of a project, outlining the tasks required in each phase and explaining how these phases interrelate; (2) as a process model, it offers a comprehensive overview of the data mining life cycle. The CRISP-DM methodology consists of six iterative phases, beginning with business understanding and culminating in deployment. Table 19 provides a concise summary of the core concepts, tasks, and outcomes associated with each of these phases as can be seen in Appendix E. (Schröer et al., 2021).

The CRISP-DM methodology, now over 20 years old, was initially developed with a focus on data mining as can be seen in Figure 2 in Appendix F. In contrast, contemporary data science places a greater emphasis on data itself and exploratory analysis. CRISP-DM was designed from a goal-oriented perspective, emphasizing the processes, tasks, and functions within those processes. In this approach, data is viewed as an essential component for achieving the objective, but not the central focus. In other words, within data mining, the process is the main focal point, whereas in modern data science, data takes precedence. With this shift in mind, the methodology applied in this study is CRISP-DM, but with a clear focus on objectives (Martinez-Plumed et al., 2021). As can be seen in figure 3 in appendix G the Data Science Trajectories (DST) map illustrates this perspective, with exploratory activities represented in the outer circle, goal-directed activities (such as CRISP-DM) in the inner circle, and data management activities at the core. Table 20 provides a concise summary of the core concepts, tasks, and outcomes associated with each of these phases as can be seen in appendix H. In this study, we applied the CRISP-DM methodology with a focus on objectives, detailing the actions undertaken in each phase, as presented in Table 21 in Appendix I.

Machine Learning Models

The goal of the higher education institution is to predict the number of students likely to drop out and those likely to enroll, in order to better allocate its resources and efforts (Klein et al., 2014; Loder, 2024b). Since this is a classification problem, where the probability of a student being classified into one class or another is also of interest, it is most appropriate to use supervised machine learning, as historical data that has already been classified is available (Jiang et al., 2020). The selection of models was based on a literature review, focusing on how

frequently they were identified as yielding the best results in analyses, as well as their strengths in addressing classification problems, as shown in Table 22 of Appendix J. The selected models are:

- **Logistic Regression. –** This algorithm is a technique used for binary classification, where the sigmoid function maps input variables to a probability value between 0 and 1. The sigmoid function is a mathematical tool that transforms any real-valued input into a range bounded by 0 and 1, creating an S-shaped curve, known as the sigmoid or logistic function. The output of logistic regression is always constrained to fall within this range, ensuring the prediction represents a valid probability. A threshold value is then applied to determine the final classification: values above the threshold are classified as 1, while those below it are classified as 0 (Nusinovici et al., 2020). For more information about the algorithm see appendix K.
- **Support Vector Machine. –** This algorithm distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closest data points from each class. The number of features in the input data determines whether the hyperplane is a line in a 2-D space or a plane in an n-dimensional space. Since multiple hyperplanes could potentially separate the classes, SVM aims to maximize the margin between the points to identify the best decision boundary. This approach helps the algorithm generalize well to new data, making accurate classification predictions. The lines adjacent to the optimal hyperplane are known as support vectors, as they pass through the data points

that define the maximal margin (Pisner & Schnyer, 2020; Valkenborg et al., 2023). For more information about the algorithm see appendix L.

- **Random Forest. –** This algorithm is a powerful tree-based learning technique in machine learning. It works by creating multiple Decision Trees during the training phase. Each tree is constructed using a random subset of the dataset and measures a random subset of features at each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance (Rigatti, 2017). During prediction, the algorithm aggregates the results of all the trees—either by voting (for classification tasks) or by averaging (for regression tasks). This collaborative decision-making process, supported by multiple trees and their individual insights, yields stable and accurate results (Biau & Scornet, 2016). For more information about the algorithm see appendix M.
- **XGBoost. –** This algorithm is considered the gold standard in ensemble learning, particularly in the realm of gradient-boosting algorithms. The model builds a series of weak learners sequentially, with each learner improving on the predictions of the previous one to produce a reliable and accurate predictive model. Fundamentally, XGBoost creates a strong predictive model by aggregating the predictions of several weak learners, typically decision trees. It employs a boosting technique, where each weak learner corrects the mistakes made by its predecessors, resulting in an extremely accurate ensemble model. The optimization method used (gradient) minimizes a cost function by iteratively adjusting the model's parameters in response to the gradients of the errors (Wade, 2020). For more information about the algorithm see appendix N.

- **Artificial Neural Network. –** Artificial Neural Networks (ANNs) can be conceptualized as weighted, directed graphs arranged in layers. Each layer comprises multiple nodes that mimic the function of biological neurons in the human brain. These nodes are interconnected and equipped with activation functions. The first layer receives raw input signals from the external environment, similar to how the optic nerve processes visual information. Subsequent layers process the output from the previous layers, resembling the way neurons deeper in the brain receive signals from those closer to sensory input. The output at each node is referred to as its activation or node value. The final layer generates the system's output. ANNs are mathematical models designed to learn and adapt from data (Abdolrasol, Hussain, et al., 2021). For more information about the algorithm see appendix O.

Evaluation Metrics

Performance Measure

Performance measures are essential for assessing the effectiveness of models, particularly in quantifying their generalization capability. Each task presents distinct requirements, which are reflected in the chosen performance measures. Consequently, model quality is a relative concept that depends not only on the algorithm and the data but also on the specific requirements of the task at hand (Janiesch et al., 2021). In prediction problems, which align with the objectives of this project, we work with a dataset $D = (x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ where y represents the dependent variable and x represents the independent variable (Z. H. Zhou, 2021). Among the most commonly used and accurate

performance measures are accuracy, precision, F1-score, recall, the receiver operating characteristic (ROC) curve and Area Under the Curve (AUC) (Flach, 2019).

The *accuracy* is the proportion of correctly classified samples. Given the dataset D, the accuracy defines as $acc(f;D) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) \neq y_i)$, for example in the dataset 50 students were classified while the model classified 45, that means it will have an accuracy of 90%. (Z. H. Zhou, 2021). The *precision* indicates the proportion of predicted positives that are correctly classified. In the context of this study, the precision metric helps determine, from the predicted number of students who are expected to dropout or enroll, how many were misclassified as dropouts or enrolls when they did not, in fact, drop out or enroll. Precision can defines as $P = \frac{TP}{TP+FP}$ where TP represents true positives and FP represents false positives (Fernandez-Garcia et al., 2021). The *recall* indicates the proportion of actual positives that are correctly predicted as positives. In the context of this study, the recall metric determines how many students who dropped or enrolled out were correctly identified out of the total number of actual dropouts or enrolls. Alternatively, it can be viewed as the number of students who dropped out or enroll but were not identified by the model. Recall can defines as $R = \frac{TP}{TP+FN}$ where TP represents true positives and FN represents false negatives (Gutiérrez-De-Rozas et al., 2022). The *F1 score* measures a model's accuracy. It combines the precision and recall scores of a model. This metric computes how many times a model made a correct prediction across the entire dataset. F1 score can defines as $F1 = \frac{2\times P\times R}{P+R}$ (Chicco & Jurman, 2020). The *ROC curve* is a probability curve, while the *AUC* measures the degree of separability between classes. A higher AUC indicates the model's enhanced ability to distinguish between different classes. In simpler

terms, a higher AUC reflects better model performance in correctly predicting class 0 as 0 and class 1 as 1 (Z. H. Zhou, 2021).

Modeling

Unbalanced dataset

The defining feature of an unbalanced dataset is class imbalance, which occurs when certain data streams, especially those with fewer instances or lower priority, are overlooked in real-world problems (Wang et al., 2021). Typically, minority class instances are labeled as positive, while majority class instances are labeled as negative. This imbalance poses a significant challenge, as minority class instances are often under-represented. As a result, even if the overall classification model achieves high accuracy, its performance on the minority class may be poor (Li et al., 2018).

When the minority class is especially important, special attention must be paid to ensure that it is accurately represented and predicted to solve this problem three techniques were implemented to balance the database:

- **ClassWeight. –** The class_weight parameter automatically adjusts weights inversely proportional to the class frequencies in the input data, helping to balance the model's attention across different classes (Bakirarar & Elhan, 2023).
- **Syntetic Minority Over-sampling Technique (SMOTE). –** This method creates synthetic samples by connecting existing minority class samples to their nearest neighbors in the feature space. The algorithm selects one of the k nearest neighbors at random and introduces a slight perturbation to the feature vector between the original sample and the neighbor. This generates new synthetic

samples that resemble the minority class but are not exact duplicates of any existing data points (Pradipta et al., 2021).

- **Random Under Samppler. –** This method removes instances from the majority class. By reducing the number of majority class samples, the dataset becomes more balanced, improving the model's ability to learn from both classes equally (Hasanin & Khoshgoftaar, 2018).

Hyperparameters

Many parameters of machine learning algorithms are learned during training. However, most modern machine learning algorithms also have hyperparameters that are external configuration variables that must be set before training begins. Hyperparameters manage the training process and can be defined manually before training the model. The performance of an algorithm often depends on how well configured its hyperparameters are for a given task (Probst et al., 2018; Weerts et al., 2020).

Hyperparameter tuning requires defining a search space, which includes the hyperparameters and their possible ranges. This process is computationally expensive, especially as the search space expands. Currently, there is no definitive evidence on which hyperparameters are crucial for tuning or which ones yield comparable performance when set to reasonable default values (van Rijn & Hutter, 2018; Weerts et al., 2020).

Grid Search

Grid search is a brute-force method used to identify the optimal set of hyperparameters. It involves generating all possible combinations of hyperparameters, training the model with each combination, and selecting the one that yields the best results. While grid search is a reliable way to find the optimal configuration, it has significant drawbacks. As the number of

hyperparameters and their possible values increases, the required computational power and processing time grow exponentially, making the process resource-intensive (Agrawal, 2021; Pirjatullah et al., 2021).

Principal component analysis

Principal Component Analysis (PCA) is a multivariate statistical technique that consolidates information from multiple observed variables into a smaller set of variables called principal components (PCs). The total variance of the original variables measures the information retained, with PCs designed to capture most of that variance. The geometric properties of PCs enable a structured and intuitive interpretation of the key features within a complex multivariate data set (Greenacre et al., 2022). PCA's primary objective is to optimize the variance while reducing the dimensionality of the feature space. As an unsupervised learning method, PCA helps simplify data without losing essential information (Salih Hasan & Abdulazeez, 2021).

Interpretability

Global interpretability provides an overarching view of model predictions across all observations, offering insights into the model's structure or internal statistics. This type of interpretation is applicable to all prediction models. In contrast, segmental interpretability focuses on specific groups within the data, revealing characteristics that may not be evident at the global level. Whether segmental interpretation is possible depends on the model's ability to create and analyze distinct segments (Gunning et al., 2019).

RESEARCH PROCEDURE

Difficult of gathering the data

To effectively implement a predictive model, a substantial amount of data is required. HEIs generate significant volumes of data related to student performance and background. When utilized properly, this data can provide valuable insights to inform decision-making in a timely manner. However, a key challenge arises from the fact that, in many instances, HEIs either do not store the data adequately, fail to collect comprehensive datasets, or collect data solely for specific, immediate purposes. Even when data is collected for necessary purposes, sharing it must adhere to strict security protocols, particularly in compliance with data protection regulations. While these regulations are essential for ensuring privacy, they can also complicate the process of accessing and sharing data for broader predictive analysis (Fernández-García et al., 2020).

Dropout

Data Understanding

In this study, we analyzed real data from 2,464 students, provided by a leading higher education institution (HEI) in Ecuador. The dataset spans four consecutive academic years (2020 to 2024) and includes students from three faculties across 12 academic programs. The database comprises 164,446 records and 35 variables, categorized into 17 categorical and 18 numerical variables. The data encompass information about students' sociodemographic characteristics, grouped into three categories: Academic Performance, Student Background, and Economic Situation, as detailed in Table 23 of Appendix P.

Data Value Exploration

392 students (15.91%) dropped out of the IES from the 2,464, while 2,072 (84.09%) remain actively enrolled. Among these students, 850 are enrolled in the Medicine program, making it the largest program in the dataset, followed by Dentistry with 415 students.

Additionally, there are 1,544 female students and 920 male students. This gender disparity becomes more significant when considering that male students have a 2.54% higher dropout rate compared to female students. Furthermore, the risk of dropout increases as the student's age increases, regardless of their GPA. However, when GPA is combined with the number of credits taken, the data shows that students with higher GPAs and more credits taken have a lower probability of dropping out, as detailed in Figures of Appendix Q.

Data Preparation

The data treatment process involves several steps. First, all duplicate values in the database are removed, meaning any identical records are eliminated. Second, missing values are either eliminated or imputed using the mean, depending on the specific variable. Third, outliers observations that differ significantly from others are removed, as their small number does not significantly affect the dataset. Finally, all categorical variables are converted into dummy variables for further analysis. (Ilyas & Chu, 2019; Oluleye, 2023; Sahoo* et al., 2019). Resulting in a dataset of 2464 entries and 26 variables. As detailed in Table 24 in Appendix R.

Enrollment

Data Understanding

In this study, we analyzed real data from 2,671 students, provided by a leading higher education institution (HEI) in Ecuador. The dataset spans three consecutive academic years (2021 to 2023) and includes student applicants from three faculties across 12 academic programs. The database consists of 2671 records and 19 variables, classified into 8 categorical and 11 numerical variables. The data covers information on the sociodemographic characteristics of the students, grouped into three categories: academic performance, student background, and economic status, as detailed in Table 25 in Appendix S.

Data Value Exploration

Out of 2,671 student applicants, 1,528 (57.2%) did not complete their enrollment at the IES, while 1,143 (42.8%) chose to enroll. Among these, 1,460 applied for the Medicine program, making it the most sought-after field, followed by Dentistry with 404 applicants. The majority of applicants were female, with 1,821 female and 848 male applicants. Despite this, a higher percentage of male applicants enrolled (48.34% compared to 40.14% of females). However, the overall number of female applicants remains significantly higher. Additionally, students with a higher grade point average (GPA) and strong entrance exam scores were more likely to enroll, likely due to the additional benefits offered by the IES. Even minimal financial aid appeared to increase the likelihood of enrollment, as shown in the figures in Appendix T.

Data Preparation

The data treatment process involves several steps. First, all duplicate values in the database are removed, meaning any identical records are eliminated. Second, missing values are either eliminated or imputed using the mean, depending on the specific variable. Third, outliers observations that differ significantly from others are removed, as their small number does not significantly affect the dataset. Finally, all categorical variables are converted into dummy variables for further analysis. (Ilyas & Chu, 2019; Oluleye, 2023; Sahoo* et al., 2019). Resulting in a dataset of 2669 entries and 16 variables. As detailed in Table 26 in Appendix U.

Modeling

The following artificial intelligence algorithms were employed in this study: Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and Feedforward Neural Network. Each algorithm was tested with data balanced using ClassWeight, SMOTE, and Random Under Sampler. Additionally, data normalization and the Get Dummies method

were applied. Once the datasets were prepared, the models were executed using Google Colab, generating 10 functional models—5 for the attrition dataset and 5 for the enrollment dataset. The datasets were initially divided into three subsets: training, validation, and test sets. Subsequently, the GridSearchCV function was used to identify the optimal hyperparameters from a predefined list, as detailed in Table 27, Appendix V. It is important to note that all processes—including dataset splitting, balancing, and model training—were conducted using a random seed 42 to ensure the experiment's reproducibility. The following insights were derived from these models.

RESULTS AND DISCUSSION

In this study, we evaluated the performance of five machine learning models: Logistic Regression, Support Vector Machine, Random Forest, XGBoost, and Feedforward Neural Networks. Categorical data in the dataset was transformed using the one-hot encoding (Dummies) method. Additionally, each model was trained on a balanced version of the dataset, resulting in four sets of performance metrics for each model. The goal of this evaluation is to identify the model that best predicts student dropout and enrollment, determining the most effective algorithm for each task.

Dropout dataset

Table 1
Logistic Regression.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.94	0.93	0.94	0.85
Recall	0.80	0.83	0.83	0.85
Precision	0.78	0.73	0.77	0.50

Table 2
Support Vector Machine.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.93	0.92	0.93	0.85
Recall	0.77	0.77	0.78	0.86
Precision	0.76	0.71	0.75	0.51

Table 3
Random Forest

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.89	0.90	0.91	0.91
Recall	0.28	0.35	0.45	0.80
Precision	0.98	0.97	0.97	0.70

Table 4
XGBoost.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.94	0.94	0.96	0.90
Recall	0.65	0.68	0.85	0.85
Precision	0.98	0.94	0.90	0.62

Table 5
Feedforward Neural Network

Evaluation Metrics	No balancing	SMOTE	Random Under Sampler
Accuracy	0.93	0.94	0.77
Recall	0.55	0.74	0.86
Precision	0.97	0.80	0.38

Table 6
Selection of the best fitting model

Evaluation Metrics	Logistic Regression	Support Vector Machine	Random Forest	XGBoost	Feedforward Neural Network (FNN)
Accuracy	0.94	0.93	0.91	**0.96**	0.94
Recall	0.83	0.77	0.80	**0.85**	0.74
Precision	0.77	0.76	0.70	**0.90**	0.80

Once it was determined that the best-fitting model was XGBoost balanced with SMOTE, due to its superior evaluation metric values, the next step involved identifying the variables the model considered important for decision-making. It is important to note that PCA was applied to this dataset, meaning the most significant variables are denoted as principal components (PCs). The analysis then focused on identifying the factors comprising these principal components.

Table 7
Importance of the principal components of the best-fitting model

Principal Component	Importance
PC24	0.029653
PC72	0.020676
PC30	0.020206
PC16	0.020044
PC10	0.019671
PC817	0.015055
PC9	0.014079
PC253	0.012648
PC163	0.011252
PC5	0.010231

Table 8
Importance of the variables of the best-fitting model

Variable	Importance

tipo_ayuda_financiera_Beca	0.007222
periodo_materia_202310_Introducción a la Economía	0.007034
periodo_materia_202230_PASEM FIN MAY 2024	0.006204
etnia_Otra	0.005930
carrera_ingreso_Medicina	0.005230
colegio_ingreso_Escuela de Medicina	0.005230
periodo_materia_202010_Emprendimiento	0.004571
periodo_materia_202320_Cosmos	0.004527
periodo_materia_202120_Digestivo/Hepato-Imagen	0.004461
periodo_materia_202120_Digestivo/Hepato-Prácticas	0.004461

Enrolment dataset

Table 9
Logistic Regression.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.98	0.97	0.96	0.96
F1 - Score	0.98	0.97	0.98	0.96

Table 10
Support Vector Machine.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.81	0.84	0.85	0.83
F1 - Score	0.74	0.80	0.82	0.79

Table 11
Random Forest

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.93	0.93	0.94	0.93
F1 - Score	0.92	0.92	0.93	0.92

Table 12
XGBoost.

Evaluation Metrics	No balancing	ClassWeight	SMOTE	Random Under Sampler
Accuracy	0.95	0.94	0.94	0.95
F1 - Score	0.94	0.94	0.94	0.95

Table 13
Feedforward Neural Network

Evaluation Metrics	No balancing	SMOTE	Random Under Sampler
Accuracy	0.96	0.95	0.95
F1 - Score	0.95	0.95	0.94

Table 14
Selection of the best fitting model

Evaluation Metrics	Logistic Regression	Support Vector Machine	Random Forest	XGBoost	Feedforward Neural Network (FNN)
Accuracy	**0.98**	0.85	0.94	0.95	0.96
F1 - Score	**0.98**	0.82	0.93	0.95	0.95

After determining that the best-fitting model was Logistic Regression without balancing, based on its superior evaluation metrics, the next step was to identify the variables that the model deemed significant for decision-making.

Table 15
Importance of the variables of the best-fitting model

Variable	Coefficient
Edad	5.6572
año_academico_admision	-0.5273

nota_examen_ingreso	0.0057
costo_carrera_semestral	-0.0013

CONCLUSIONS

The phase that required the most development time was undoubtedly the analysis phase, which involved understanding the business data and performing subsequent data cleansing. Most of the execution time was spent in this phase, resulting in a reliable and structured dataframe. It is important to note that the success or failure of projects of this nature largely depends on the quality of the data obtained for analysis. During the implementation of the algorithms, several challenges were encountered and addressed. One notable challenge was the execution of different algorithms. Despite using the same dataset for all tests, each algorithm operates differently, employs distinct methodologies for processing data, consumes varying hardware resources, and required independent development efforts for each implementation. The diversity of algorithms tested led to variations in their evaluation metrics. This is expected, given the differences in their operations, mathematical foundations, and input variable handling. As a result, each algorithm produced unique outcomes.

From the results obtained, it was determined that the XGBoost algorithm best fit the data, achieving a performance score of 0.94 and a recall of 0.85 for the attrition dataset. For the enrollment dataset, Logistic Regression emerged as the best model, with a performance score of 0.98 and an F1 score of 0.98. These metrics represent acceptable values, considering the complexity of the datasets.

For the dropout dataset, the model identified several important variables, including the type of financial assistance provided, particularly scholarships, and the student's chosen field

of study, with medicine being a significant factor. Additionally, the model highlighted the importance of specific courses, underscoring the need for a more in-depth evaluation at the faculty level. For the enrollment dataset, key variables included the student's age, emphasizing the importance of early enrollment, and the academic year of admission, indicating that the broader context of that year may influence students' decisions to enroll. Exam scores also played a crucial role, demonstrating that students with higher scores are more likely to enroll. Finally, the cost of tuition per semester was identified as a factor, suggesting that financial considerations may influence enrollment decisions.

It is important to highlight certain limitations in the dataset, primarily related to how information is stored. For instance, some students are listed under an initial major they never actually pursued. Additionally, percentage updates related to tuition costs are not accurately maintained as tuition rates increase. Variables such as ethnicity and college of origin also present potential biases due to a significant amount of missing data, which the university has not consistently collected.

BIBLIOGRAPHIC REFERENCES

Ab Ghani, N. L., Che Cob, Z., Mohd Drus, S., & Sulaiman, H. (2019). Student Enrolment Prediction Model in Higher Education Institution: A Data Mining Approach. *Lecture Notes in Electrical Engineering*, *565*, 43–52. https://doi.org/10.1007/978-3-030-20717-5_6

Abbott-Chapman, J., Martin, K., Ollington, N., Venn, A., Dwyer, T., & Gall, S. (2014). The longitudinal association of childhood school engagement with adult educational and occupational achievement: findings from an Australian national study. *British Educational Research Journal*, *40*(1), 102–120. https://doi.org/10.1002/berj.3031

Abdolrasol, M. G. M., Hussain, S. M. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., & Milad, A. (2021). Artificial Neural Networks Based Optimization Techniques: A Review. *Electronics*, *10*(21), 2689. https://doi.org/10.3390/electronics10212689

Abdolrasol, M. G. M., Suhail Hussain, S. M., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., & Milad, A. (2021). Artificial neural networks based optimization techniques: A review. In *Electronics (Switzerland)* (Vol. 10, Issue 21). MDPI. https://doi.org/10.3390/electronics10212689

Agrawal, T. (2021). Hyperparameter Optimization Using Scikit-Learn. In *Hyperparameter Optimization in Machine Learning* (pp. 31–51). Apress. https://doi.org/10.1007/978-1-4842-6579-6_2

Ahmed Arafa, A., Radad, M., Badawy, M., & El-Fishawy, N. (2022). Logistic Regression Hyperparameter Optimization for Cancer Classification. *Menoufia Journal of Electronic Engineering Research*, *0*(0), 0–0. https://doi.org/10.21608/mjeer.2021.70512.1034

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. In *Socio-Economic Planning Sciences* (Vol. 79). Elsevier Ltd. https://doi.org/10.1016/j.seps.2021.101102

Akmanchi, S., Bird, K. A., & Castleman, B. L. (2023). *Human versus Machine: Do college advisors outperform a machine-learning algorithm in predicting student enrollment?* https://doi.org/10.26300/gadf-ey53

Álvarez-Pérez, P. R., López-Aguilar, D., González-Morales, M. O., & Peña-Vázquez, R. (2024). Academic Engagement and Dropout Intention in Undergraduate University Students. *Journal of College Student Retention: Research, Theory and Practice*, *26*(1), 108–125. https://doi.org/10.1177/15210251211063611

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. In *International Journal of Educational Technology in Higher Education* (Vol. 17, Issue 1). Springer. https://doi.org/10.1186/s41239-020-0177-7

Ambesange, S., Vijayalaxmi, A., Sridevi, S., Venkateswaran, & Yashoda, B. S. (2020). Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 827–832. https://doi.org/10.1109/WorldS450073.2020.9210404

Andina, D., Vega-Corona, A., Seijas, J. I., & Torres-Garcìa, J. (2007). Neural Networks Historical Review. In *Computational Intelligence* (pp. 39–65). Springer US. https://doi.org/10.1007/0-387-37452-3_2

Arif Ali, Z., H. Abduljabbar, Z., A. Tahir, H., Bibo Sallow, A., & Almufti, S. M. (2023). eXtreme Gradient Boosting Algorithm with Machine Learning: a Review. *Academic*

Journal of Nawroz University, *12*(2), 320–334. https://doi.org/10.25007/ajnu.v12n2a1612

Bakirarar, B., & Elhan, A. H. (2023). Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research. *Turkiye Klinikleri Journal of Biostatistics*, *15*(1), 19–29. https://doi.org/10.5336/biostatic.2022-93961

Barramuño, M., Meza-Narváez, C., & Gálvez-García, G. (2022). Prediction of student attrition risk using machine learning. *Journal of Applied Research in Higher Education*, *14*(3), 974–986. https://doi.org/10.1108/JARHE-02-2021-0073

Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education*, *8*(2), 614–652. https://doi.org/10.1002/rev3.3202

Berger, J. B., & Milem, J. F. (1999). The role of student involvement and perceptions of integration in a causal model of student persistence. *Research in Higher Education*, *40*(6), 641–664. https://doi.org/10.1023/A:1018708813711

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bishop, C., & Nasrabadi, N. (2006). *Pattern recognition and machine learning*. https://link.springer.com/book/9780387310732

Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(6), 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Boumi, S., & Vela, A. E. (2021). Quantifying the impact of student enrollment patterns on academic success using a hidden markov model. *Applied Sciences (Switzerland)*, *11*(14). https://doi.org/10.3390/app11146453

Bousnguar, H., Najdi, L., & Battou, A. (2022). Forecasting approaches in a higher education setting. *Education and Information Technologies*, *27*(2), 1993–2011. https://doi.org/10.1007/s10639-021-10684-z

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). *Data Cleaning, Feature Selection, and Data Transforms in Python*.

Burtner, J. (2005). The Use of Discriminant Analysis to Investigate the Influence of Non-Cognitive Factors on Engineering School Persistence. *Journal of Engineering Education*, *94*(3), 335–338. https://doi.org/10.1002/j.2168-9830.2005.tb00858.x

Camizuli, E., & Carranza, E. J. (2018). Exploratory Data Analysis (EDA) . In *The Encyclopedia of Archaeological Sciences* (pp. 1–7). Wiley. https://doi.org/10.1002/9781119188230.saseas0271

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2023). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory and Practice*, *25*(1), 51–75. https://doi.org/10.1177/1521025120964920

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chopra, A., and, S. M.-T. I. of N. T., & 2021, undefined. (2021). Is Job Shadowing a Panacea for Educational Drop Outs? *SpringerA Chopra, S MenonThe Importance of New Technologies and Entrepreneurship in Business, 2021•Springer*, *194 LNNS*, 1975–1987. https://doi.org/10.1007/978-3-030-69221-6_142

Christenson, S., & Reschly, A. (2008). Best practices in fostering student engagement. *Experts.Umn.Edu*. https://experts.umn.edu/en/publications/best-practices-in-fostering-student-engagement

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Dalal, S., Onyema, E. M., & Malik, A. (2022). Hybrid XGBoost model with hyperparameter tuning for prediction of liver disease with better accuracy. *World Journal of Gastroenterology*, *28*(46), 6551–6563. https://doi.org/10.3748/wjg.v28.i46.6551

Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European*

Journal of Operational Research, *281*(3), 575–587. https://doi.org/10.1016/j.ejor.2019.03.037

Delogu, M., Lagravinese, R., Paolini, D., & Resce, G. (2024). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, *130*. https://doi.org/10.1016/j.econmod.2023.106583

Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2022, September 16). Predicting Students' Performance Using Machine Learning Algorithms. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3564982.3564990

Diaz Lema, M., Vooren, M., Cannistrà, M., van Klaveren, C., Agasisti, T., & Cornelisz, I. (2024). Predicting dropout in Higher Education across borders. *Studies in Higher Education*, *49*(1), 141–156. https://doi.org/10.1080/03075079.2023.2224818

Drucker, H., Burge, C., Kaufman, L., Smola, A., & Vapoik, V. (1996). Support Vector Regression Machines. *Advances in Neural Information Processing Systems*.

Fernandez-Garcia, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sanchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. *IEEE Access*, *9*, 133076–133090. https://doi.org/10.1109/ACCESS.2021.3115851

Fernández-García, A. J., Rodríguez-Echeverría, R., Preciado, J. C., Conejero Manzano, J. M., & Sánchez-Figueroa, F. (2020). Creating a recommender system to support higher education students in the subject enrollment decision. *IEEE Access*, *8*, 189069–189088. https://doi.org/10.1109/ACCESS.2020.3031572

Ferreira, L., Pilastri, A., Martins, C. M., Pires, P. M., & Cortez, P. (2021). A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost. *2021 International*

Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534091

Findeisen, S., Brodsky, A., Michaelis, C., Schimmelpenningh, B., & Seifried, J. (2024). Dropout intention: a valid predictor of actual dropout? *Empirical Research in Vocational Education and Training*, *16*(1). https://doi.org/10.1186/s40461-024-00165-1

Flach, P. (2019). Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9808–9814. https://doi.org/10.1609/aaai.v33i01.33019808

Forsman, J., van den Bogaard, M., Linder, C., & Fraser, D. (2015). Considering student retention as a complex system: a possible way forward for enhancing student retention. *European Journal of Engineering Education*, *40*(3), 235–255. https://doi.org/10.1080/03043797.2014.941340

Fraysier, K., Reschly, A., & Appleton, J. (2020). Predicting Postsecondary Enrollment With Secondary Student Engagement Data. *Journal of Psychoeducational Assessment*, *38*(7), 882–899. https://doi.org/10.1177/0734282920903168

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Furini, M., Galli, G., & Martini, M. C. (2021). On using video lectures data usage to predict university students dropout. *GoodIT 2021 - Proceedings of the 2021 Conference on Information Technology for Social Good*, 313–316. https://doi.org/10.1145/3462203.3475890

Geng, Y., Li, Q., Yang, G., & Qiu, W. (2024). Logistic Regression. In *Practical Machine Learning Illustrated with KNIME* (pp. 99–132). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-3954-7_4

Goldhaber, D., Long, M. C., Person, A. E., Rooklyn, J., & Gratz, T. (2019). Sign Me Up: The Factors Predicting Students' Enrollment in an Early-Commitment Scholarship Program. *AERA Open*, *5*(2). https://doi.org/10.1177/2332858419857703

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Gutiérrez-De-Rozas, B., Molina, E. C., & López-Martín, E. (2022). Academic Failure and Dropout: Untangling Two Realities. *European Journal of Educational Research*, *11*(4), 2275–2289. https://doi.org/10.12973/eu-jer.11.4.2275

Guzmán, A., Barragán, S., & Cala Vitery, F. (2021). Dropout in Rural Higher Education: A Systematic Review. In *Frontiers in Education* (Vol. 6). Frontiers Media S.A. https://doi.org/10.3389/feduc.2021.727833

Hasanin, T., & Khoshgoftaar, T. (2018). The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 70–79. https://doi.org/10.1109/IRI.2018.00018

Hemachandran, K., Tayal, S., George, P. M., Singla, P., & Kose, U. (2022). *Bayesian Reasoning and Gaussian Processes for Machine Learning Applications*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003164265

Hieu, N. D., Ho, N. C., & Lan, V. N. (2020). ENROLLMENT FORECASTING BASED ON LINGUISTIC TIME SERIES. *Journal of Computer Science and Cybernetics*, *36*(2), 119–137. https://doi.org/10.15625/1813-9663/36/2/14396

Ilyas, I., & Chu, X. (2019). Data Cleaning. In *Data Cleaning*.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). *Machine learning and deep learning*. https://doi.org/10.1007/s12525-021-00475-2/Published

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E., & Alvarado-Corona, R. (2024). Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-53576-1

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jüttler, M. (2020). Predicting economics student retention in higher education: The effects of students' economic competencies at the end of upper secondary school on their intention to leave their studies in economics. *PLoS ONE*, *15*(2). https://doi.org/10.1371/journal.pone.0228505

Kalita, D. J., Singh, V. P., & Kumar, V. (2020). *A Survey on SVM Hyper-Parameters Optimization Techniques* (pp. 243–256). https://doi.org/10.1007/978-981-15-2071-6_20

Kehm, B. M., Larsen, M. R., & Sommersel, H. B. (2020). Student dropout from universities in Europe: A review of empirical literature. *Hungarian Educational Research Journal*, *9*(2), 147–164. https://doi.org/10.1556/063.9.2019.1.18

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Klein, D. J., Bershteyn, A., & Eckhoff, P. A. (2014). Dropout and re-enrollment. *AIDS*, *28*(Supplement 1), S47–S59. https://doi.org/10.1097/QAD.0000000000000081

Kocsis, Á., & Molnár, G. (2024). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*. https://doi.org/10.1080/03054985.2024.2316616

Korhonen, V., Educational, J. R.-S. J. of, & 2019, undefined. (2019). Identifying problematic study progression and "at-risk" students in higher education in Finland. *Taylor & FrancisV Korhonen, J RautopuroScandinavian Journal of Educational Research, 2019•Taylor & Francis*, *63*(7), 1056–1069. https://doi.org/10.1080/00313831.2018.1476407

Lee, Y., & Choi, J. (2013). A structural equation model of predictors of online learning retention. *The Internet and Higher Education*, *16*, 36–42. https://doi.org/10.1016/j.iheduc.2012.01.005

Li, F., Zhang, X., Zhang, X., Du, C., Xu, Y., & Tian, Y.-C. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. *Information Sciences*, *422*, 242–256. https://doi.org/10.1016/j.ins.2017.09.013

Loder, A. K. F. (2024a). Comparing Student Performance in Multiple Enrollments and Single Enrollments: Possible Target Groups for University Management. *Journal of College Student Retention: Research, Theory and Practice*. https://doi.org/10.1177/15210251241262435

Loder, A. K. F. (2024b). University system & multiple enrollment policy: dropout and graduation clusters. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2406574

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13084702

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-32484-w

Mayer-Foulkes, D. (2002). On the dynamics of quality student enrollment at institutions of higher education. *Economics of Education Review*, *21*(5), 481–489. https://doi.org/10.1016/S0272-7757(01)00036-X

Morales, E. F., & Escalante, H. J. (2022). A brief introduction to supervised, unsupervised, and reinforcement learning. In *Biosignal Processing and Classification Using Computational Learning and Intelligence* (pp. 111–129). Elsevier. https://doi.org/10.1016/B978-0-12-820125-1.00017-8

Morelli, M., Chirumbolo, A., Baiocco, R., Educativa, C. E.-P., & 2021, undefined. (2021). Academic failure: Individual, organizational, and social factors. *Iris.Uniroma1.ItM*

Morelli, A Chirumbolo, R Baiocco, C ElenaPsicología Educativa, 2021•iris.Uniroma1.It. https://doi.org/10.5093/psed2021a8

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

Nicoletti, M. do C. (2019). Revisiting the Tinto's Theoretical Dropout Model. *Higher Education Studies*, *9*(3), 52. https://doi.org/10.5539/hes.v9n3p52

Nikolaidis, P., Ismail, M., Shuib, L., Khan, S., & Dhiman, G. (2022). Predicting Student Attrition in Higher Education through the Determinants of Learning Progress: A Structural Equation Modelling Approach. *Sustainability (Switzerland)*, *14*(20). https://doi.org/10.3390/su142013584

Nita, B., Nowosielski, K., Kes, Z., Sidor, O., Oleksyk, P., Walaszczyk, E., Golec, P., Zaniewska, A., Turek, T., & Król, R. (2022). Machine learning in the enrolment management process: a case study of using GANs in postgraduate students' structure prediction. *Procedia Computer Science*, *207*, 1350–1359. https://doi.org/10.1016/j.procs.2022.09.191

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*. https://doi.org/10.1016/j.caeai.2022.100066

Nordmann, E., Calder, C., Bishop, P., Irwin, A., Education, D. C.-H., & 2019, undefined. (2019). Turn up, tune in, don't drop out: The relationship between lecture attendance, use of lecture recordings, and achievement at different levels of study. *SpringerE*

Nordmann, C Calder, P Bishop, A Irwin, D ComberHigher Education, 2019•Springer, *77*(6), 1065–1084. https://doi.org/10.1007/s10734-018-0320-8

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Okoye, K., Nganji, J. T., Escamilla, J., & Hosseini, S. (2024). Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education. *Computers and Education: Artificial Intelligence*, *6*. https://doi.org/10.1016/j.caeai.2024.100205

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems*, *134*. https://doi.org/10.1016/j.dss.2020.113320

Oluleye, A. (2023). Exploratory Data Analysis with Python Cookbook. In *Exploratory Data Analysis with Python Cookbook: Over 50 recipes to analyze, visualize, and extract insights from structured and unstructured data*.

Oppong, S. O. (2023). Predicting Students' Performance Using Machine Learning Algorithms: A Review. *Asian Journal of Research in Computer Science*, *16*(3), 128–148. https://doi.org/10.9734/ajrcos/2023/v16i3351

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). *How Many Trees in a Random Forest?* (pp. 154–168). https://doi.org/10.1007/978-3-642-31537-4_13

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors. *Industrial Management & Data Systems*, *116*(8), 1678–1699. https://doi.org/10.1108/IMDS-09-2015-0363

P, A. (2020). Higher Education Institution (HEI) Enrollment Forecasting Using Data Mining Technique. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2060–2064. https://doi.org/10.30534/ijatcse/2020/179922020

Parmar, A., Katariya, R., & Patel, V. (2019). *A Review on Random Forest: An Ensemble Classifier* (pp. 758–763). https://doi.org/10.1007/978-3-030-03146-6_86

Peng, C.-Y. J., & Nichols, R. N. (2003). Using Multinomial Logistic Models To Predict Adolescent Behavioral Risk. *Journal of Modern Applied Statistical Methods*, *2*(1), 177–188. https://doi.org/10.22237/jmasm/1051748160

Phan, M., De Caigny, A., & Coussement, K. (2023). A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems*, *168*. https://doi.org/10.1016/j.dss.2023.113940

Pirjatullah, Kartini, D., Nugrahadi, D. T., Muliadi, & Farmadi, A. (2021). Hyperparameter Tuning using GridsearchCV on The Comparison of The Activation Function of The ELM Method to The Classification of Pneumonia in Toddlers. *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, 390–395. https://doi.org/10.1109/IC2IE53219.2021.9649207

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Popov, K. (2019). Factors, Affecting Students' Decision to Enroll in a University. *Pedagogia : Jurnal Pendidikan*, *8*(2), 201–210. https://doi.org/10.21070/pedagogia.v8i2.2231

Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., & Ismail, M. (2021). SMOTE for Handling Imbalanced Data Problem : A Review. *2021 Sixth International*

Conference on Informatics and Computing (ICIC), 1–8. https://doi.org/10.1109/ICIC54025.2021.9632912

Prasanth, A., & Alqahtani, H. (2023). Predictive Modeling of Student Behavior for Early Dropout Detection in Universities using Machine Learning Techniques. *International Conference on Engineering Technologies and Applied Sciences: Shaping the Future of Technology through Smart Computing and Engineering, ICETAS 2023*. https://doi.org/10.1109/ICETAS59148.2023.10346531

Probst, P., Bischl, B., & Boulesteix, A.-L. (2018). *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11). https://doi.org/10.3390/data7110146

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. *Handbook of Research on Student Engagement*, 3–19. https://doi.org/10.1007/978-1-4614-2018-7_1

Revathy, M., Kamalakkannan, S., & Kavitha, P. (2022). *Machine Learning based Prediction of Dropout Students from the Education University using SMOTE*. 1750–1758. https://doi.org/10.1109/icssit53264.2022.9716450

Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

Rodríguez-Pineda, M., Uniciencia, J. Z.-A.-, & 2021, undefined. (n.d.). College student dropout: cohort study about possible causes. *Scielo.Sa.CrM Rodríguez-Pineda, JA*

Zamora-ArayaUniciencia, 2021•scielo.Sa.Cr. Retrieved September 16, 2024, from https://www.scielo.sa.cr/scielo.php?pid=S2215-34702021000100019&script=sci_arttext&tlng=en

Sahoo*, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 4727–4735. https://doi.org/10.35940/ijitee.L3591.1081219

Salih Hasan, B. M., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, *02*(01). https://doi.org/10.30880/jscdm.2021.02.01.003

Santos, G., Souza, A., Mantovani, R., Cruz, R., Cordeiro, T., & Souza, F. (2024, May 20). An Exploratory Analysis on Gender-Related Dropout Students in Distance Learning Higher Education using Machine Learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3658271.3658323

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? *Mathematics*, *10*(18). https://doi.org/10.3390/math10183359

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. *IEEE Access*, *10*, 6470–6482. https://doi.org/10.1109/ACCESS.2022.3141992

Søgaard, M., Kasper, L., Kornbeck, P., Müller, R., Malene, K., Larsen, R., & Sommersel, H. B. (2013). *Clearinghouse-research series 2013 number 15 A systematic review*. http://edu.au.dk/en/research/research-areas/danish-clearinghouse-

Stefos, E. (2019). Los Estudiantes de Pregrado en Ecuador: Un Análisis de Datos. *Revista Scientific*, *4*(14), 85–100. https://doi.org/10.29394/Scientific.issn.2542-2987.2019.4.14.4.85-100

Suyal, M., & Goyal, P. (2022). A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. *International Journal of Engineering Trends and Technology*, *70*(7), 43–48. https://doi.org/10.14445/22315381/IJETT-V70I7P205

Thanh Ngoc, T., Van Dai, L., & Minh Thuyen, C. (2021). Support Vector Regression based on Grid Search method of Hyperparameters for Load Forecasting. *Acta Polytechnica Hungarica*, *18*(2), 143–158. https://doi.org/10.12700/APH.18.2.2021.2.8

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169–190. https://doi.org/10.3233/AIC-170729

Tiwari, A. (2022). Supervised learning: From theory to applications. In *Artificial Intelligence and Machine Learning for EDGE Computing* (pp. 23–32). Elsevier. https://doi.org/10.1016/B978-0-12-824054-0.00026-5

Torok, E., & Angeli, E. (2022). Analysis of dual and non-dual student learning outcomes and student dropout data. *IEEE Global Engineering Education Conference, EDUCON, 2022-March*, 1303–1309. https://doi.org/10.1109/EDUCON52537.2022.9766648

Ujkani, B., Minkovska, D., & Stoyanova, L. (2021, September 15). A Machine Learning Approach for Predicting Student Enrollment in the University. *2021 30th International Scientific Conference Electronics, ET 2021 - Proceedings*. https://doi.org/10.1109/ET52713.2021.9579795

Ujkani, B., Minkovska, D., & Stoyanova, L. (2022). Application of Logistic Regression Technique for Predicting Student Dropout. *2022 31st International Scientific Conference Electronics, ET 2022 - Proceedings*. https://doi.org/10.1109/ET55967.2022.9920280

Valkenborg, D., Rousseau, A.-J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754–757. https://doi.org/10.1016/j.ajodo.2023.08.003

van Rijn, J. N., & Hutter, F. (2018). Hyperparameter Importance Across Datasets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2367–2376. https://doi.org/10.1145/3219819.3220058

Vapnik, V. N., & Chervonenkis, A. Ya. (2015). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In *Measures of Complexity* (pp. 11–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-21852-6_3

Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python: Vol. First*.

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*. https://doi.org/10.1016/j.chb.2019.106189

Waldrop, D., Reschly, A. L., Fraysier, K., & Appleton, J. J. (2019). Measuring the Engagement of College Students: Administration Format, Structure, and Validity of the Student Engagement Instrument–College. *Measurement and Evaluation in Counseling and Development*, *52*(2), 90–107. https://doi.org/10.1080/07481756.2018.1497429

Wang, L., Han, M., Li, X., Zhang, N., & Cheng, H. (2021). Review of Classification Methods on Unbalanced Data Sets. *IEEE Access*, *9*, 64606–64628. https://doi.org/10.1109/ACCESS.2021.3074243

Weerts, H. J. P., Mueller, A. C., & Vanschoren, J. (2020). *Importance of Tuning Hyperparameters of Machine Learning Algorithms*.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Xiong, X., Guo, X., Zeng, P., Zou, R., & Wang, X. (2022). A Short-Term Wind Power Forecast Method via XGBoost Hyper-Parameters Optimization. *Frontiers in Energy Research*, *10*. https://doi.org/10.3389/fenrg.2022.905155

Yang, L., Feng, L., Zhang, L., & Tian, L. (2021). Predicting freshmen enrollment based on machine learning. *Journal of Supercomputing*, *77*(10), 11853–11865. https://doi.org/10.1007/s11227-021-03763-y

Yang, S., Chen, H. C., Chen, W. C., & Yang, C. H. (2020). Student Enrollment and Teacher Statistics Forecasting Based on Time-Series Analysis. *Computational Intelligence and Neuroscience*, *2020*. https://doi.org/10.1155/2020/1246920

Zhou, X., Liu, H., Shi, C., & Liu, J. (2022). The basics of deep learning. In *Deep Learning on Edge Computing Devices* (pp. 19–36). Elsevier. https://doi.org/10.1016/B978-0-32-385783-3.00009-0

Zhou, Z. H. (2021). Machine Learning. In *Machine Learning*. Springer Nature. https://doi.org/10.1007/978-981-15-1967-3

Zong, C., & Davis, A. (2022). Modeling University Retention and Graduation Rates Using IPEDS. *Journal of College Student Retention: Research, Theory and Practice*. https://doi.org/10.1177/15210251221074379

APPENDIX A: CATEGORIES DROPOUTS

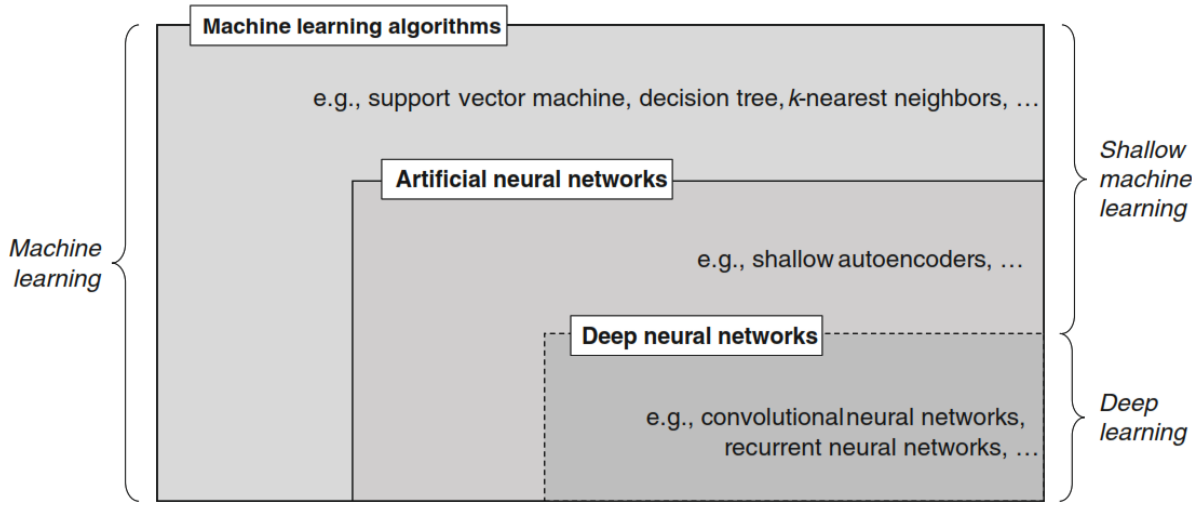
Table 16
Categories used for the classification of student dropout factors

Authors	Categories
(Álvarez-Pérez et al., 2024)	Individual Sociodemographics (student profile), academic profile and related to academic engagement.
(Barramuño et al., 2022)	Personal profile, academic profile, Socioeconomic profile.
(Chopra et al., 2021)	Student, Institutional, Family, Community, Social Networks.
(Dervenis et al., 2022)	Demographics, Socioeconomic, Academic Profile, Interpersonal Relationships, Psychological Profile.
(Diaz Lema et al., 2024)	Student's personal profile, academic profile in school, academic profile in university.
(Fernandez-Garcia et al., 2021)	Personal and admission data, Qualifications, Scholarships.
(Jiménez-Gutiérrez et al., 2024)	Individual Sociodemographic (student profile), related to academic profile, basic services, work, etc.
(Martins et al., 2023)	Demographics, Socioeconomic, Academic profile, Macroeconomic, etc.
(Matz et al., 2023)	Individual Sociodemographic (student profile), Academic Profile.
(Morelli et al., 2021)	Psychological profile, Organizational profile University profile and related profiles
(Nikolaidis et al., 2022)	Student-university, student-teacher interaction interests, student drop-out beliefs
(Niyogisubizo et al., 2022)	Academic profile
(Realinho et al., 2022)	Demographic, Socioeconomic, Academic Profile, Macroeconomic, Academic Profile 1st and 2nd Semester.
(Rodríguez-Pineda et al., 2021)	Academic profile, environmental, socio-economic and voational data.
(Santos et al., 2024)	Geographical location, related to the student's profile, related to the academic profile.
(Segura et al., 2022)	Career, field of study, gender, academic level, scholarship, type of scholarship, reason for admission.
Singh & Alhulail, 2022	Personal profile, academic profile, Socioeconomic profile
(Ujkani et al., 2022)	Individual Sociodemographic (student profile), Academic Profile
(Zong & Davis, 2022)	Institutional Achievement, Student Engagement, Academic Record, Student Finances, Academic Environment, Social Environment

APPENDIX B: VENN DIAGRAM OF MACHINE LEARNING CONCEPTS AND CLASSES

Figure 1

Venn diagram of machine learning concepts and classes (Janiesch et al., 2021)



APPENDIX C: ML, DL, ANN MODELS USED FOR DROPOUT

Table 17
ML, DL, ANN models, types of data used and best model performance for dropout

Authors	Data Type	ML/DL/NN Models Used	Best Performing Model
(Findeisen et al., 2024)	Quantitative (survey)	Logistic Regression, Decision Trees	Logistic Regression
(Phan et al., 2023)	Textual data (from institutional sources)	Support Vector Machines, Random Forest	Support Vector Machines
(Fernandez-Garcia et al., 2021)	Academic performance data	Gradient Boosting, Random Forest, Support Vector Machine	Ensemble (Gradient Boosting + Random Forest + SVM)
(Santos et al., 2024)	Enrollment and academic records	Decision Trees, Logistic Regression	Logistic Regression
(Torok & Angeli, 2022)	Dual and non-dual student performance data	Bayesian Networks, Logistic Regression, Decision Trees	Decision Trees
(Ujkani et al., 2022)	Survey and academic records	Logistic Regression, Random Forest	Logistic Regression
(Jiménez-Gutiérrez et al., 2024)	Academic data	Neural Networks, Support Vector Machines, Logistic Regression	Neural Networks
(Cardona et al., 2023)	Retention-related institutional data	Neural Networks, Decision Trees, SVM, Logistic Regression	Support Vector Machines
(Delen et al., 2020)	Freshman student dropout data	Bayesian Belief Networks, Logistic Regression, Decision Trees	Bayesian Belief Network
(Furini et al., 2021)	Video lecture usage data	KNN, Random Forest, Decision Tree	Random Forest + KNN
(Oppong, 2023)	Various educational data sources	Neural Networks, Decision Trees, Naive Bayes	Neural Networks
(Kocsis & Molnár, 2024)	Various data from academic performance	Decision Trees, Logistic Regression	Logistic Regression

(Diaz Lema et al., 2024)	High school metrics, academic performance	GLM, Lasso, RF, Classification Trees	Random Forest
(Revathy et al., 2022)	Survey data	SMOTE, Logistic Regression, KNN, PCA	KNN with PCA-SMOTE
(Okoye et al., 2024)	Student retention and academic performance data	SVM, Decision Tree, MLP, RF	SVM
(Segura et al., 2022)	Demographic and academic data	Neural Networks, SVM, Decision Trees	Neural Networks
(Martins et al., 2023)	Demographic, socio-economic, and academic data	Random Forest, SMOTE, SVM, RusBoost	SMOTE + Random Forest
(Waheed et al., 2020)	Big data from virtual learning environment	Deep Learning Models, Logistic Regression, SVM	Random Forest + Deep Learning
(Delogu et al., 2024)	Enrollment and institutional data	RF, GBM, Neural Networks, LASSO	Random Forest
(Realinho et al., 2022)	Institutional data	Random Forest, KNN, SVM	Random Forest + KNN
(Niyogisubizo et al., 2022)	Institutional data	Random Forest, Gradient Boosting, Decision Trees	Random Forest + Gradient Boosting
(Dervenis et al., 2022)	Academic performance data	KNN, Naive Bayes, Neural Networks	Neural Networks
(Singh & Alhulail, 2022)	Data from student-teachers	Logistic Regression, Decision Trees	Logistic Regression
(Barramuño et al., 2022)	University student data	KNN, Decision Trees, Neural Networks	Random Forest + KNN
(Prasanth & Alqahtani, 2023)	Academic performance data	SVM, Random Forest	SVM + Random Forest
(Olaya et al., 2020)	Socio-demographic characteristics and engagement	SMOTE, Random Forest, XGBoost	Random Forest
(Matz et al., 2023)	Institutional data	Random Forest, XGBoost, GBM	Random Forest + XGBoost

APPENDIX D: ML, DL, ANN MODELS USED FOR ENROLLMENT

Table 18
ML, DL, ANN models, types of data used and best model performance for enrollment

Authors	Data Type	Models Used	Best Performing Model
(Nita et al., 2022)	Case study	GAN-based models, Deep Learning	ICGAN-DSVM algorithm
(L. Yang et al., 2021)	HEI data	Decision Tree, Random Forest, BP Neural Network	Random Forest
(Ujkani et al., 2021)	HEI data	Naive Bayes, KNN, Decision Tree, Logistic Regression	Decision Tree
(Boumi & Vela, 2021)	HEI data	Hidden Markov Model	Hidden Markov Model
(S. Yang et al., 2020)	HEI data	WOASVR (Whale Optimization Algorithm Support Vector Regression)	WOASVR
(Ab Ghani et al., 2019)	Case Study	Logistic Regression, Decision Tree, Naïve Bayes	Decision Tree
(Fernández-García et al., 2020)	HEI data	Random Forest, Gradient Boosting Classifier, Logistic Regression, Support Vector Machine, k Nearest Neighbors, Multilayer Perceptron	Random Forest
(Hieu et al., 2020)	Enrollment Forecasting	Fuzzy Time Series	Fuzzy Time Series
(P, 2020)	Enrollment Forecasting	ARIMA	ARIMA (0,2,1)

APPENDIX E: CRISP – DM PROCESS MODEL

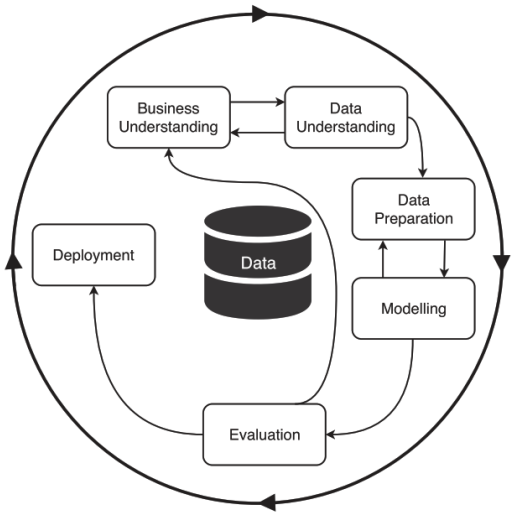
Table 19
CRISP-DM process model descriptions (Wirth & Hipp, 2000)

Phase	Description
Business Understanding	In this phase, it is essential to obtain a comprehensive overview of the business current situation, including an assessment of the resources required and those already available. The primary objective of this phase is to clearly define the goal of the data mining process, specifying the type of data mining to be performed and the criteria for measuring its success. Additionally, it is crucial to establish a well-defined project plan to guide the subsequent steps.
Data Understanding	In this phase, the collection of data from its source, along with its exploration, description, and quality assessment, are critical tasks. Employing statistical analysis, determining relevant attributes, and ensuring proper data matching are essential steps to ensure the integrity and usefulness of the data.
Data Preparation	In this phase, data selection should be carried out by establishing clear inclusion and exclusion criteria. Poor-quality data can be addressed during the cleaning process. Based on the model defined in the initial phase, derived attributes should be constructed as needed. Various data cleansing techniques and model types can be applied to ensure the data is properly prepared to fit the chosen models.
Modeling	This phase involves selecting the appropriate modeling technique, constructing the test case, and developing the model. Any data mining technique may be utilized during this process. Specific parameters must be configured to build the model effectively. To evaluate the model's performance, it should be assessed against predefined evaluation criteria, with the best-performing models being selected for further analysis.
Evaluation	In this phase, the results are compared with the predefined objectives to ensure alignment. The outcomes must be carefully interpreted, and new actions should be defined based on these findings. Additionally, a comprehensive review of the entire process is necessary to ensure consistency and to identify areas for improvement.
Deployment	In this phase, a final report or software component must be produced, marking the deployment of the solution. Additionally, plans for monitoring and maintenance should be established to ensure the ongoing functionality and effectiveness of the implemented model.

APPENDIX F: CRISP – DM PROCESS MODEL OF DATA MINING

Figure 2

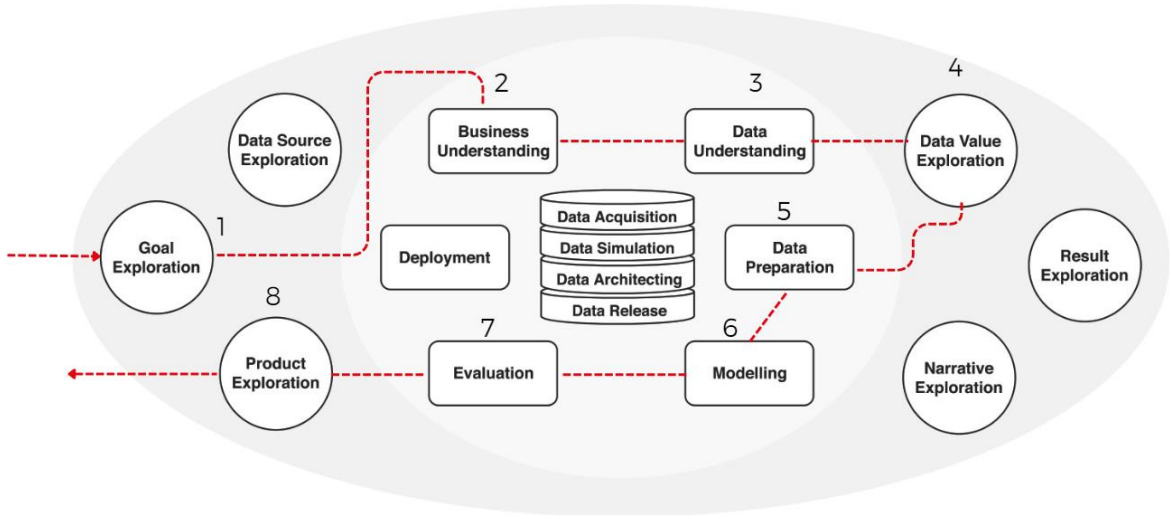
The CRISP-DM process model of data mining (Martinez-Plumed et al., 2021)



APPENDIX G: TRAJECTORY THROUGH A DATA SCIENCE PROJECT

Figure 3

Trajectory through a data science project (Martinez-Plumed et al., 2021)



APPENDIX H: CRISP-DM GOAL-DIRECTED

Table 20
CRISP-DM goal-directed (Martinez-Plumed et al., 2021)

Phase	Description
Goal Exploration	In this phase, the objectives of the project are defined, including the goals to be achieved and the specific characteristics of the project. A thorough understanding of the problem is essential, as it establishes the foundation for the subsequent phases of the process.
Business Understanding	In this phase, it is essential to obtain a comprehensive overview of the business current situation, including an assessment of the resources required and those already available.
Data Understanding	In this phase, data collection from its source, along with its exploration, description, and quality assessment, are critical tasks. Conducting statistical analysis, identifying relevant attributes, and ensuring proper data matching are essential steps to guarantee the integrity and usability of the data.
Data Value Exploration	This phase focuses on exploring how the data can generate value for the organization or the research being conducted. The goal is to identify key variables or data segments that are most likely to yield actionable insights and inform decision-making processes.
Data Preparation	In this phase, data selection should be carried out by establishing clear inclusion and exclusion criteria. This process involve cleaning, transforming, integrating, and structuring the data to ensure it is suitable for modeling.
Modeling	This phase involves selecting the appropriate modeling technique, constructing the test case, and developing the model. To evaluate the model's performance, it should be assessed against predefined evaluation criteria, with the best-performing models being selected for further analysis
Evaluation	In this phase, the results are compared with the predefined objectives to ensure alignment. The outcomes must be carefully interpreted, and new actions should be defined based on these findings.
Product Exploration	This phase refers to how data-driven insights or models can be applied to develop or enhance business. The findings from the proyect will inform to the people who make the decisions on what to improve or what path to take with respect to the project.

APPENDIX I: CRISP-DM GOAL-DIRECTED IN THE STUDY

Table 21
CRISP-DM goal-directed based on (Martinez-Plumed et al., 2021) for this study

Phase	Description
Goal Exploration	The objective of this project is to predict the probability of student desertion and enrollment in a Higher Education Institute using Machine Learning and Deep Learning models. The aim is to provide critical information that supports decision-making to reduce the desertion rate and increase the enrollment rate. Among the specific goals is the use of six different classification models to label students and identify which model best fits the data for future analyses.
Business Understanding	The HEI is aware and considers that its dropout rate is high and that the student enrollment rate could be improved, they are clear that they have a somewhat structured database since 2020.
Data Understanding	The data were collected by a designated department and subsequently anonymized to ensure confidentiality during utilization. Following this, data cleaning was performed according to the steps outlined by Brownlee in (2020), and an exploratory data analysis was conducted according to the steps detailed by Camizuli & Carranza in (2018).
Data Value Exploration	This phase focuses on exploring how the data can generate value for the organization or the research being conducted. The goal is to identify key variables or data segments that are most likely to yield actionable insights and inform decision-making processes.
Data Preparation	In this phase, data selection should be carried out by establishing clear inclusion and exclusion criteria. This process involve cleaning, transforming, integrating, and structuring the data to ensure it is suitable for modeling.
Modeling	This phase involves selecting the appropriate modeling technique, constructing the test case, and developing the model. To evaluate the model's performance, it should be assessed against predefined evaluation criteria, with the best-performing models being selected for further analysis
Evaluation	In this phase, the results are compared with the predefined objectives to ensure alignment. The outcomes must be carefully interpreted, and new actions should be defined based on these findings.
Product Exploration	This phase refers to how data-driven insights or models can be applied to develop or enhance business. The findings from the proyect will inform to

	the people who make the decisions on what to improve or what path to take with respect to the project.

APPENDIX J: SUPERVISED MACHINE LEARNING ALGORITHM

Table 22
Brief description of supervised machine learning models

Algorithm	Description	Strengths	Limitations	Reference
Naive Bayes	Naïve Bayes belongs to a family of generative learning algorithms, meaning it models the distribution of inputs within a given class or category. Unlike discriminative classifiers, such as logistic regression, it does not focus on identifying the most important features for distinguishing between classes. Naïve Bayes assumes that the predictors are conditionally independent, meaning each feature is unrelated to the others in the model. It also assumes that all features contribute equally to the outcome. Although these assumptions are often violated in real-world scenarios (e.g., the likelihood of a word in an email	Lower complexity: Is considered a simpler classifier compared to others, as its parameters are easier to estimate. Good scalability: It is a fast and efficient classifier, offering high accuracy when the assumption of conditional independence is met. Additionally, it requires minimal storage.	Zero-frequency issue: This occurs when a categorical variable is absent from the training set. For example, if we attempt to estimate the maximum likelihood for the word "sir" in the class “spam,” but "sir" is not present in the training data, the likelihood would be zero. Since Naïve Bayes multiplies all conditional probabilities, this would result in a posterior probability of zero. To address this issue, techniques such as Laplace smoothing or discounting can be applied. Unrealistic assumption of independence: Although the assumption of conditional independence often performs well in practice, it is not always	(Hemachandran et al., 2022)

	depends on the preceding word), they simplify classification problems, making them more computationally efficient. This simplification reduces the problem to calculating a single probability for each variable, which, in turn, facilitates model computation. It is based on Bayes' theorem, which can be expressed as: $P(A \mid B)=(P(A \mid B)P(A))/P(B)$	Handles high-dimensional data: Naïve Bayes can effectively manage high-dimensional datasets, such as those encountered in document classification, where other classifiers might struggle.	satisfied, which can lead to incorrect classifications.	
Linear discriminant analysis	Linear Discriminant Analysis (LDA, also known as Normal Discriminant Analysis (NDA) or Discriminant Function Analysis (DFA), follows a generative model framework. This means LDA models the data distribution for each class and applies Bayes' theorem to classify new data points. Bayes' theorem calculates conditional probabilities—that is, the probability of an event given that another event has occurred. LDA uses this approach to predict	Simplicity and computational efficiency: LDA is a straightforward yet powerful algorithm, making it easy to understand and implement, especially for those new to machine learning. Its efficient computation allows for quick results. Effective in high-dimensional spaces: LDA performs well when the number of features exceeds the number of training samples. This makes it particularly valuable in applications such as	Shared mean distributions: LDA faces difficulties when class distributions have the same mean. In such cases, it struggles to find a linear axis that effectively separates the classes, making it less effective at distinguishing between groups with overlapping statistical properties. Limited to labeled data: LDA is a supervised learning algorithm, meaning it requires labeled data for classification or separation. In contrast, Principal Component Analysis (PCA), another	(Tharwat et al., 2017)

	the likelihood that a given input belongs to a specific class. LDA identifies a linear combination of features that best separates or distinguishes two or more classes. It achieves this by projecting data from a higher-dimensional space onto a single dimension, making classification simpler and more effective.	text analysis, image recognition, and genomics, where high-dimensional data is common. Handles multicollinearity: LDA addresses multicollinearity—when features are highly correlated—by transforming the data into a lower-dimensional space while preserving essential information.	dimensionality reduction technique, does not rely on class labels and instead focuses on preserving variance in the data.	
Logistic Regression	Logistic regression is a technique used for binary classification, where the sigmoid function maps input variables to a probability value between 0 and 1. The sigmoid function is a mathematical tool that transforms any real-valued input into a range bounded by 0 and 1, creating an S-shaped curve, known as the sigmoid or logistic function. The output of logistic regression is always constrained to fall within this range, ensuring the	Ease of implementation and interpretation: Logistic regression is simple to implement, easy to interpret, and highly efficient to train. No distributional assumptions: It does not assume any specific distribution of classes in the feature space. Extension to multiple classes: Logistic regression can be extended to handle multiple classes through multinomial regression and provides a natural	Limitations with high-dimensional data: Logistic regression should be avoided when the number of observations is smaller than the number of features, as this can lead to overfitting. Linear decision boundaries: It constructs linear boundaries, which limits its ability to model complex relationships. Linearity assumption: A major limitation of logistic regression is the assumption of a linear relationship between the independent variables and the	(Nusinovici et al., 2020)

	prediction represents a valid probability. A threshold value is then applied to determine the final classification: values above the threshold are classified as 1, while those below it are classified as 0.	probabilistic interpretation of class predictions. Interpretability of coefficients: The model not only estimates the magnitude of predictors (coefficients) but also indicates the direction of their association (positive or negative). Speed and efficiency: It classifies new data quickly, making it suitable for large datasets. Performance: Logistic regression achieves good accuracy for many simple datasets and performs well when the data is linearly separable. Feature importance: Model coefficients can be interpreted as indicators of feature importance. Overfitting considerations: While logistic regression is less prone to overfitting, it may overfit in high-	log-odds of the dependent variable. Discrete outcomes: Logistic regression is designed to predict discrete outcomes, meaning the dependent variable must belong to a finite set of categories. Challenges with non-linear data: Logistic regression cannot solve non-linear problems because its decision surface is linear. However, linearly separable data is rare in real-world scenarios. Multicollinearity: Logistic regression performs best when there is little to no multicollinearity among the independent variables. Limited modeling of complex relationships: It struggles to capture complex relationships in data. More advanced models, such as neural networks, often outperform logistic regression in these scenarios.	

		dimensional datasets. To address this, regularization techniques (L1 and L2) can be applied.		
k-nearest neighbors (KNN)	The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier that uses proximity to classify or predict the grouping of an individual data point. For classification tasks, the class label is assigned based on a majority vote—i.e., the label most frequently represented among the nearest neighbors of a given data point. Although this is technically referred to as "plurality voting," the term "majority vote" is more commonly used in the literature. The difference between these terms is that "majority voting" technically requires more than 50% of the votes, which is most relevant when there are only two categories.	Easy to implement: Due to its simplicity and accuracy, KNN is often one of the first classifiers that a new data scientist learns. Adaptable: As new training samples are added, the algorithm easily adjusts to accommodate the new data, since all training data is stored in memory. Few hyperparameters: KNN requires only the selection of a k value and a distance metric, making it less complex in terms of hyperparameter tuning compared to other machine learning algorithms.	Scalability issues: Since KNN is a lazy algorithm, it requires significant memory and data storage compared to other classifiers. This can be costly in terms of both time and resources. Curse of dimensionality: KNN is susceptible to the curse of dimensionality, meaning its performance deteriorates when handling high-dimensional data. Prone to overfitting: Due to the curse of dimensionality, KNN is more prone to overfitting. Although techniques like feature selection and dimensionality reduction can help mitigate this, the choice of k also plays a crucial role. Smaller values of k tend to overfit the data, while	(Suyal & Goyal, 2022)

			larger values of k smooth the predictions by averaging over a larger neighborhood. However, if k is too large, the model may underfit the data.	
Support vector machine (SVM)	Support Vector Machines (SVMs) are widely used for classification tasks. They distinguish between two classes by finding the optimal hyperplane that maximizes the margin between the closest data points from each class. The number of features in the input data determines whether the hyperplane is a line in a 2-D space or a plane in an n-dimensional space. Since multiple hyperplanes could potentially separate the classes, SVM aims to maximize the margin between the points to identify the best decision boundary. This approach helps the algorithm generalize well to new data, making accurate classification predictions. The lines adjacent to the optimal hyperplane are known	High-dimensional performance: SVM excels in high-dimensional spaces, making it particularly well-suited for tasks such as image classification and gene expression analysis. Nonlinear capability: By using kernel functions like RBF and polynomial kernels, SVM can effectively model nonlinear relationships. Outlier resilience: The soft margin feature allows SVM to tolerate outliers, improving its robustness in tasks like spam detection and anomaly detection. Binary and multiclass support: SVM is effective for both binary classification and multiclass classification, making it a	Slow training: SVM can be slow when dealing with large datasets, which can impact its performance in data mining tasks. Parameter tuning difficulty: Selecting the right kernel and adjusting parameters, such as C, requires careful tuning, which can be challenging and affect the algorithm's performance. Noise sensitivity: SVM struggles with noisy datasets and overlapping classes, which can limit its effectiveness in real-world applications. Limited interpretability: The complexity of the hyperplane in higher-dimensional spaces makes SVM less interpretable compared to other models.	(Pisner & Schnyer, 2020)

	as support vectors, as they pass through the data points that define the maximal margin.	versatile choice for applications such as text classification. Memory efficiency: Since SVM focuses on support vectors, it is more memory-efficient compared to other algorithms.	Feature scaling sensitivity: Proper feature scaling is crucial for SVM; without it, the model's performance can degrade significantly.	
Decision Tree	A decision tree is commonly used to model and predict outcomes based on input data. It consists of a tree-like structure, where each internal node tests an attribute, each branch corresponds to a possible attribute value, and each leaf node represents the final decision or prediction. The process of constructing a decision tree involves recursively partitioning the data based on the values of different attributes. At each internal node, the algorithm selects the best attribute to split the data, using criteria such as information gain or Gini impurity. This splitting process continues until a stopping criterion is met,	Easy to interpret: The Boolean logic and visual representation of decision trees make them easier to understand and interpret. The hierarchical structure also makes it clear which attributes are most important, a feature that is not always as obvious with other algorithms, such as neural networks. Minimal data preparation required: Decision trees are highly flexible, handling various data types (discrete or continuous). Continuous values can be converted into categorical values through thresholds. Additionally, decision trees can manage missing values, which	Prone to overfitting: Complex decision trees are prone to overfitting and may not generalize well to new data. This issue can be mitigated through pre-pruning and post-pruning techniques. Pre-pruning stops tree growth when there is insufficient data, while post-pruning removes subtrees that lack adequate data after the tree is constructed. High variance estimators: Small variations in the data can lead to significantly different decision trees. Bagging, or averaging multiple estimates, can help reduce the variance of decision trees. However, this approach has	(Charbuty & Abdulazeez, 2021)

	such as reaching a maximum depth or having a minimum number of instances in a leaf node.	can be problematic for other classifiers. Flexible: Decision trees can be used for both classification and regression tasks, offering more versatility than some other algorithms. They are also less sensitive to the underlying relationships between attributes—if two variables are highly correlated, the algorithm will choose only one to split on.	its limitations, as it may lead to highly correlated predictors. Higher computational cost: Because decision trees use a greedy search approach during construction, they can be more computationally expensive to train compared to other algorithms.	
Random Forest	The Random Forest algorithm is a powerful tree-based learning technique in machine learning. It works by creating multiple Decision Trees during the training phase. Each tree is constructed using a random subset of the dataset and measures a random subset of features at each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and	Reduced risk of overfitting: Decision trees are prone to overfitting, as they tend to fit all samples in the training data. However, in a random forest, the risk of overfitting is reduced because averaging uncorrelated trees decreases overall variance and prediction error. With a large number of trees, the classifier is less likely to overfit the model. Provides flexibility: The random forest algorithm excels at both regression and classification tasks	Time-consuming process: Although random forest algorithms can handle large datasets and provide more accurate predictions, they can be slow to process the data, as they compute information for each individual decision tree. Requires more resources: Since random forests process larger	(Rigatti, 2017)

	improving overall prediction performance. During prediction, the algorithm aggregates the results of all the trees—either by voting (for classification tasks) or by averaging (for regression tasks). This collaborative decision-making process, supported by multiple trees and their individual insights, yields stable and accurate results. Random forests are widely used for both classification and regression tasks, known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in diverse environments.	with a high degree of accuracy, making it a popular choice among data scientists. Additionally, its ability to cluster features makes it an effective tool for estimating missing values, maintaining accuracy even when a portion of the data is missing. Easy-to-determine feature importance: The random forest simplifies the process of assessing the importance or contribution of variables to the model. There are several methods to evaluate feature importance. Gini importance and the mean decrease in impurity (MDI) are commonly used metrics to determine how much the model's accuracy decreases when a particular variable is excluded.	datasets, they require more resources to store that data. More complex: Predicting with a single decision tree is easier to interpret compared to using a forest of trees.	
Extreme Gradient Boosting	XGBoost is a state-of-the-art machine learning algorithm renowned for its exceptional predictive performance. It is considered the gold standard in ensemble learning, particularly in	User-friendly implementation: Gradient boosting decision trees are relatively easy to implement. Many implementations support handling categorical features, require minimal data	Requires careful parameter tuning to achieve optimal performance: Proper tuning of parameters is essential to ensure the best performance of the model.	(Wade, 2020)

	the realm of gradient-boosting algorithms. The model builds a series of weak learners sequentially, with each learner improving on the predictions of the previous one to produce a reliable and accurate predictive model. Fundamentally, XGBoost creates a strong predictive model by aggregating the predictions of several weak learners, typically decision trees. It employs a boosting technique, where each weak learner corrects the mistakes made by its predecessors, resulting in an extremely accurate ensemble model. The optimization method used (gradient) minimizes a cost function by iteratively adjusting the model's parameters in response to the gradients of the errors. The algorithm introduces the concept of "gradient boosting with decision trees," where the objective function is minimized	preprocessing, and streamline the process of handling missing data.	

	by calculating the importance of each decision tree added to the ensemble. Additionally, by incorporating a regularization term and utilizing a more advanced optimization algorithm, XGBoost further enhances both accuracy and efficiency.		
		Bias reduction: In machine learning, bias refers to systematic errors that can lead models to make inaccurate or unfair predictions. Boosting algorithms, including gradient boosting, sequentially add multiple weak learners to the larger predictive model. This technique is highly effective at reducing bias, as each additional weak learner iteratively improves the model.	Can be prone to overfitting if not properly regularized: Without proper regularization, the model can overfit the training data, leading to poor generalization to new data.
		Improved accuracy: Boosting enables decision trees to learn sequentially, with new trees compensating for the errors made by previous ones. This iterative process results in more accurate predictions than any individual weak learner could achieve.	May not perform as well with high-dimensional sparse data: The model might struggle with high-dimensional sparse datasets, where many features have little or no meaningful information.

		Additionally, decision trees can handle both numerical and categorical data types, making them versatile for a wide range of problems. Faster training on large data sets: Boosting methods prioritize features that increase the model's predictive accuracy during training. This selectivity reduces the number of data attributes, resulting in computationally efficient models capable of handling large datasets. Boosting algorithms can also be parallelized to further speed up model training.	Training can be computationally expensive, especially with large datasets: Training the model can be resource-intensive, particularly with large datasets, requiring significant computational power. Interpreting the model can be challenging due to its complexity: The complexity of the model can make it difficult to interpret and understand the individual contributions of features to the final prediction.	
Artificial Neural Networks (ANNs)	Artificial Neural Networks (ANNs) can be best viewed as weighted directed graphs, commonly organized in layers.	Artificial neural networks have the ability to process data in parallel, meaning they can handle multiple tasks at the same time.	Artificial neural networks have the ability to process data in parallel, meaning they can handle multiple tasks at the same time.	(Abdolrasol, Hussain, et al., 2021)

	These layers consist of numerous nodes that imitate the biological neurons of the human brain. These nodes are interconnected and contain an activation function. The first layer receives the raw input signal from the external world, analogous to the optic nerves in human visual processing. Each successive layer receives the output from the preceding layer, similar to the way neurons further from the optic nerve receive signals from those closer to it. The output at each node is called its activation or node value. The final layer produces the system's output. ANNs are mathematical models capable of learning.	They are resistant to failure, which means that the loss of one or more neurons does not significantly affect the performance of the network. Artificial neural networks are designed to store information within the network, so even in the absence of a data pair, the network can still generate results. Additionally, artificial neural networks are robust and gradually degrade over time, meaning they do not suddenly stop working. We can train ANNs to learn from past events and make decisions.	They are resistant to failure, which means that the loss of one or more neurons does not significantly affect the performance of the network. Artificial neural networks are designed to store information within the network, so even in the absence of a data pair, the network can still generate results. Additionally, artificial neural networks are robust and gradually degrade over time, meaning they do not suddenly stop working. We can train ANNs to learn from past events and make decisions.

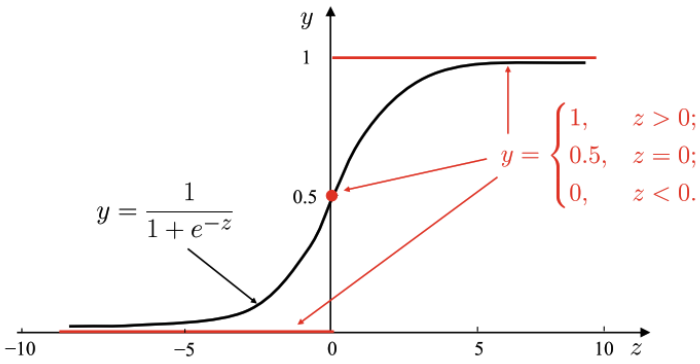
APPENDIX K: LOGISTIC REGRESSION

The first mathematical model of logistic regression was proposed in the 1940s as a technique to address the limitations of ordinary least squares (OLS) regression when dealing with dichotomous variables. By the 1980s, logistic regression became available in statistical software packages, facilitating its widespread adoption (Geng et al., 2024). It has been extensively used in epidemiological research and continues to gain prominence in the social sciences, particularly in higher education (Peng & Nichols, 2003; Singh & Alhulail, 2022; Ujkani et al., 2022).

The logistic regression model converts the continuous output of a linear regression function into a categorical output by using a sigmoid function. This function maps any set of real-valued independent variables to a value between 0 and 1 and is commonly referred to as the logistic function (Geng et al., 2024).

Log odds can be challenging to interpret in logistic regression analyses. To simplify this, it is common to exponentiate the beta estimates, transforming them into odds ratios (OR). The OR represents the likelihood of an outcome occurring given a specific event, compared to the likelihood of the outcome in the absence of that event. An OR greater than 1 indicates that the event is associated with higher odds of the outcome occurring, while an OR less than 1 suggests lower odds of the outcome (Geng et al., 2024; Z. H. Zhou, 2021).

Figure 4
Unit-step function and logistic function (Z. H. Zhou, 2021).


y
1
0.5
$y = \frac{1}{1+e^{-z}}$
$y = \begin{cases} 1, & z > 0; \\ 0.5, & z = 0; \\ 0, & z < 0. \end{cases}$
−10
−5
0
5
10
z

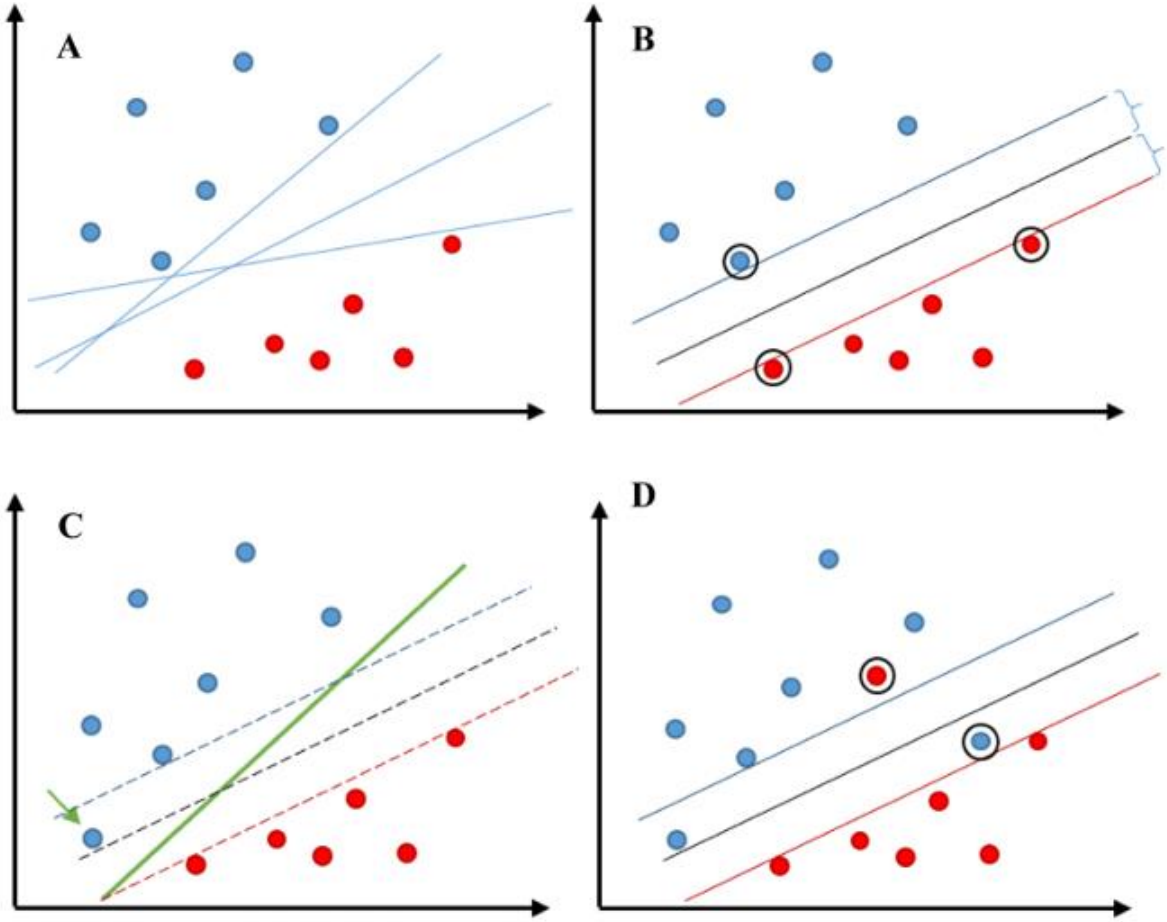
APPENDIX L: SUPPORT VECTOR MACHINE

The fundamentals of Support Vector Machines (SVM) were introduced in the 1960s (Vapnik & Chervonenkis, 2015), but the theory continued to evolve over the following decades. In the 1990s, SVM gained significant attention from the scientific community due to two major advancements: the development of the kernel trick (Boser et al., 1992), which enabled SVM to handle non-linear classification problems, and the extension of SVM to solve regression tasks (Drucker et al., 1996). The mathematical foundation of SVM is complex, requiring a solid understanding of optimization theory, linear algebra, and learning theory (Valkenborg et al., 2023).

Support Vector Machine (SVM) aims to maximize the margin of separation between two classes by identifying the optimal hyperplane. This hyperplane, also known as the *maximum margin hyperplane* or *hard margin*, is chosen to maximize the distance between itself and the closest data points from each class, known as support vectors (Pisner & Schnyer, 2020). When such a hyperplane exists, it ensures the greatest separation between classes. Additionally, SVM has the capability to handle outliers effectively, as it focuses on maximizing the margin while ignoring data points that do not significantly impact the decision boundary, making the algorithm robust to outliers (Valkenborg et al., 2023).

Figure 5

A and B illustrate the principle of the maximum-margin classifier. C and D demonstrate the introduction of the slack variable, which allows the support vector classifier to maximize its margin while disregarding the influence of nearby observations, even when the data is non-separable (Valkenborg et al., 2023).


A
B
C
D

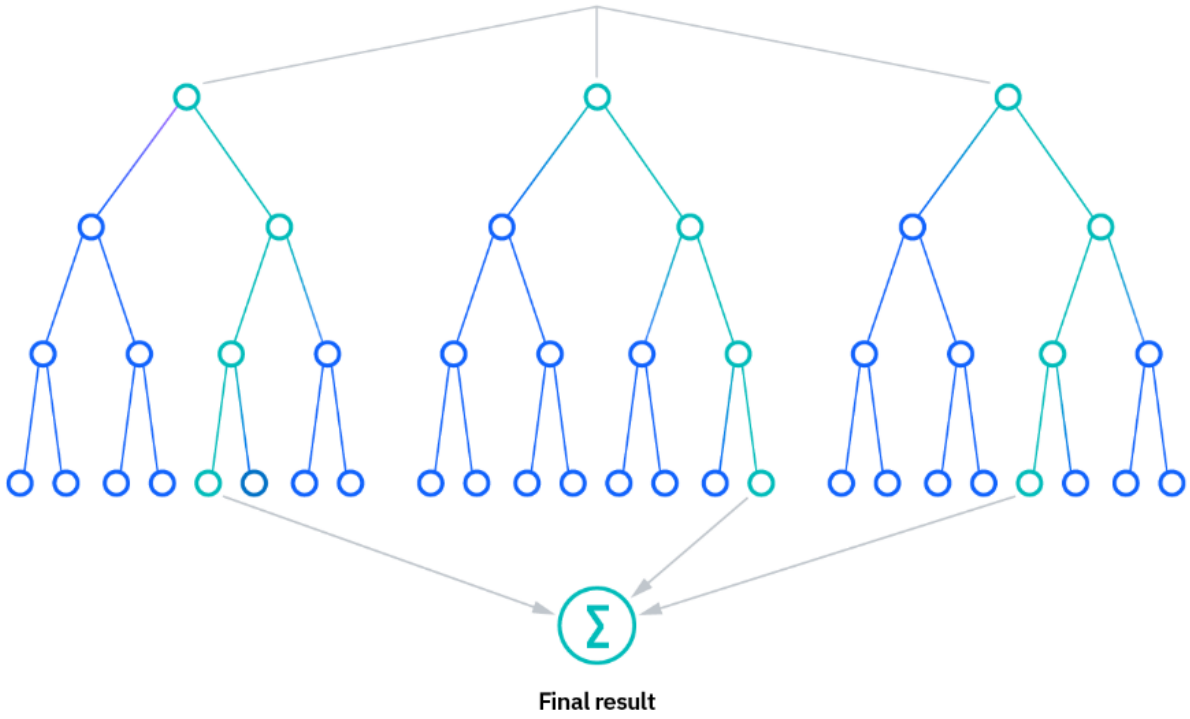
APPENDIX M: RANDOM FOREST

This algorithm is a collaborative ensemble of decision trees working together to produce a single outcome, first introduced in 2001 by Breiman. Since Random Forest consists of multiple decision trees, it is helpful to briefly explain decision trees. A decision tree begins with a basic question, branching into additional questions to guide the decision-making process. These questions form the nodes of the tree, splitting the data at each step. The final decision is represented by the leaf nodes (Costa & Pedreira, 2023).

The Random Forest algorithm is made up of a collection of decision trees, each constructed from a randomly drawn sample of the training data, using a technique called bootstrapping. Approximately one-third of each bootstrap sample is reserved as test data, known as the out-of-bag (OOB) sample, which plays a crucial role in model evaluation. Another layer of randomness is introduced through feature bagging, which selects random subsets of features for each split, enhancing model diversity and reducing correlations between trees (Parmar et al., 2019).

Depending on the problem type, the prediction method varies. For regression tasks, the predictions from individual trees are averaged to obtain the final result. In classification tasks, a majority vote—where the most frequently predicted class is chosen—determines the final output (Biau & Scornet, 2016).

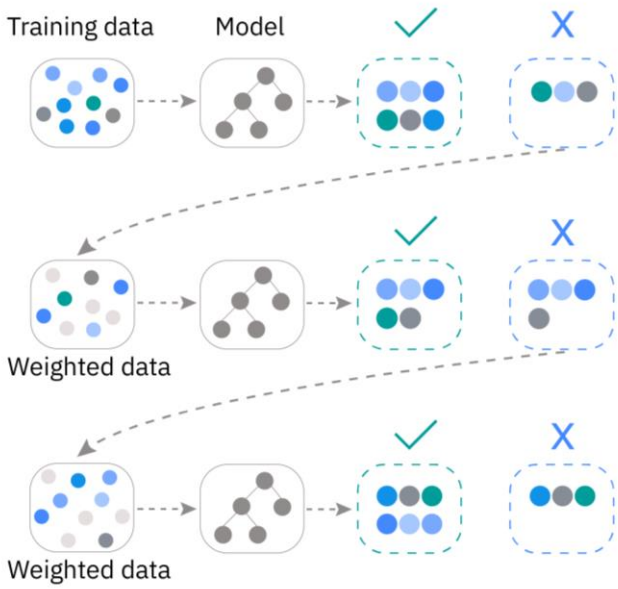
Figure 6
Random Forest (Biau & Scornet, 2016).


Σ
Final result

APPENDIX N: XGBOOST

This algorithm is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger, more accurate prediction. Introduced in 2015 by Chen & Guestrin, XGBoost stands for "Extreme Gradient Boosting." In this approach, decision trees are built sequentially, with each tree correcting the errors of the previous one. Weights play a crucial role in XGBoost. Initially, weights are assigned to all independent variables fed into the decision tree. If the tree incorrectly predicts certain outcomes, the weights of those variables are increased, emphasizing their importance in the next tree. This iterative process results in a collection of individual classifiers or predictors that, when combined, create a robust and accurate model (Arif Ali et al., 2023). XGBoost is highly versatile, capable of handling user-defined regression, classification, and prediction tasks, making it a powerful tool across a wide range of applications (Ferreira et al., 2021).

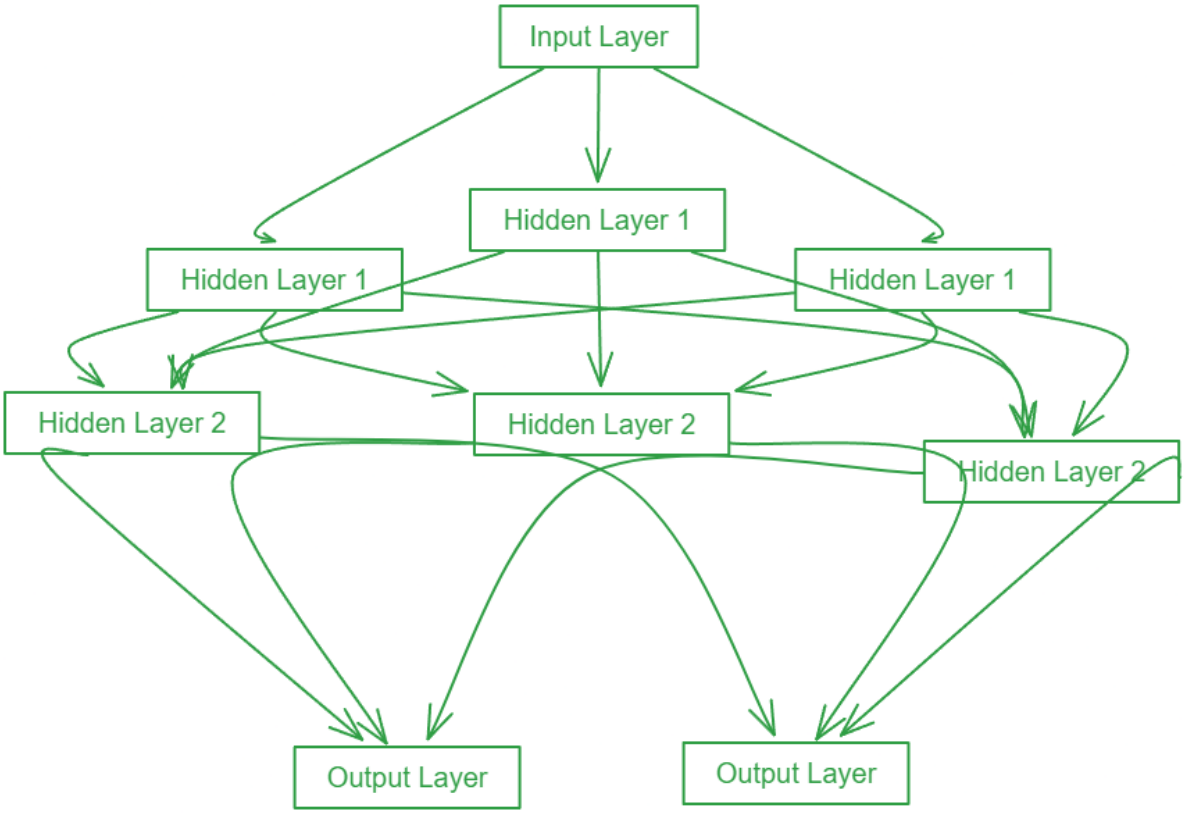
Figure 7
Boosting - sequential ensemble learning (Ferreira et al., 2021).



APPENDIX O: FEEDFORWARD NEURAL NETWORK (FNN)

Artificial neural networks (ANNs) emerged in the 1940s as scientists attempted to replicate the functions of the human brain through physical models and program simulations. One of the simplest types of ANNs is the feedforward neural network, characterized by its unidirectional data flow—from the input nodes, through hidden nodes (if present), and finally to the output nodes, without any loops (Andina et al., 2007). A feedforward neural network comprises three types of layers: input, hidden, and output. Each layer contains units known as neurons, which are interconnected by weights. The network operates in two main phases: the Feedforward Phase and the Backpropagation Phase (X. Zhou et al., 2022). In the feedforward phase, data enters the network and propagates through it. The inputs pass through the hidden layers, where weighted sums are calculated, until they reach the output layer, where a prediction is made. In the backpropagation phase, after the prediction is generated, the network calculates the error. This error is propagated backward, and the weights are adjusted to minimize the error, improving the network's accuracy over time (Abdolrasol, Hussain, et al., 2021).

Figure 8
Forward Propagation (X. Zhou et al., 2022).


Input Layer
Hidden Layer 1
Hidden Layer 1
Hidden Layer 1
Hidden Layer 2
Hidden Layer 2
Hidden Layer 2
Output Layer
Output Layer

APPENDIX P: LIST OF VARIABLES AND DESCRIPTION

Table 23
List of all variables provided by the HEI with their detailed description.

Category	Variable Name	Description	DataType
Academic Performance	Number of Major Transfers	Number of times the student has transferred majors within the university.	int64
	Credits Taken	Total credits completed by the student up to the year of analysis.	int64
	Dropped Out	Student's major or status after the year of analysis.	int64
	Course Grade Equivalency	Approval equivalency based on the course grade obtained.	object
	Semester GPA	Student's GPA for the semester mentioned in the "Registration Period" field.	float64
	Course Name	Name of the course the student took during the "Registration Period" semester.	object
	Course Grade	Grade received in the course taken during the "Registration Period" semester.	object
	Major Transfer Period	Semester in which the student changed majors.	float64
	Transferred Major	Major to which the student transferred.	float64
	Transferred University	University from which the student transferred.	float64
Economic Situation	Semester Major Cost	Cost associated with the student's major during the year of analysis.	int64
	Financial Aid Percentage	Percentage of financial aid received by the student during the "Registration Period" semester.	int64
	Scholarship Percentage	Percentage of the scholarship received by the student during the "Registration Period" semester.	int64
	Future Payment Percentage	Percentage of future payments for the student during the "Registration Period" semester.	int64

Student Background	Total Financial Aid Percentage	Total percentage of financial aid received by the student during the "Registration Period" semester.	int64
	Type of Financial Aid	Type of financial aid the student received during the "Registration Period" semester.	object
	Academic Year of Admission	Academic year in which the student was admitted to the university.	object
	Academic Year of Enrollment	Academic year in which the student registered for courses.	object
	Year of Analysis	Academic year being analyzed.	object
	Major	Student's major during the year of analysis.	object
	Admission Major	Major in which the student was admitted during the academic year of admission.	object
	School	School the student is attending during the year of analysis.	object
	High School of Origin	High school where the student completed their secondary education.	object
	Admission School	University school in which the student was admitted during the academic year of admission.	object
	Ethnicity	Ethnicity to which the student belongs.	object
	Date of Birth	Student's date of birth.	
	Gender	Reported gender of the student.	date
	High School GPA	Final GPA with which the student graduated from high school.	int64
	ID	Student's unique identifier.	int64
	Analysis Period	Semester analyzed in the study.	object
	Admission Period	Semester in which the student was admitted to the university.	object
	Registration Period	Semester in which the student registered for courses.	object
	Province	Province to which the student belongs.	object

APPENDIX Q: DROP OUT DATA EXPLORATION

Figure 9
Percentage of students who drop out.

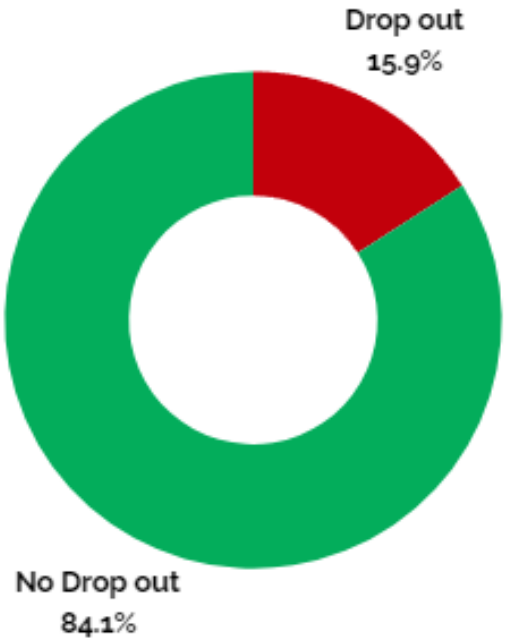


Figure 10
Students by major.

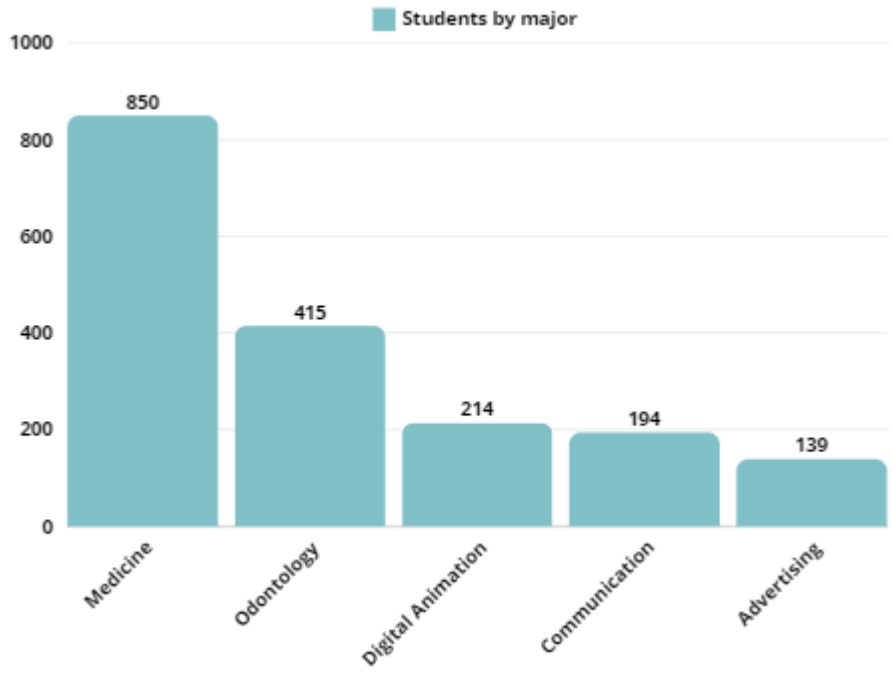


Figure 11
Dropout rate by major.

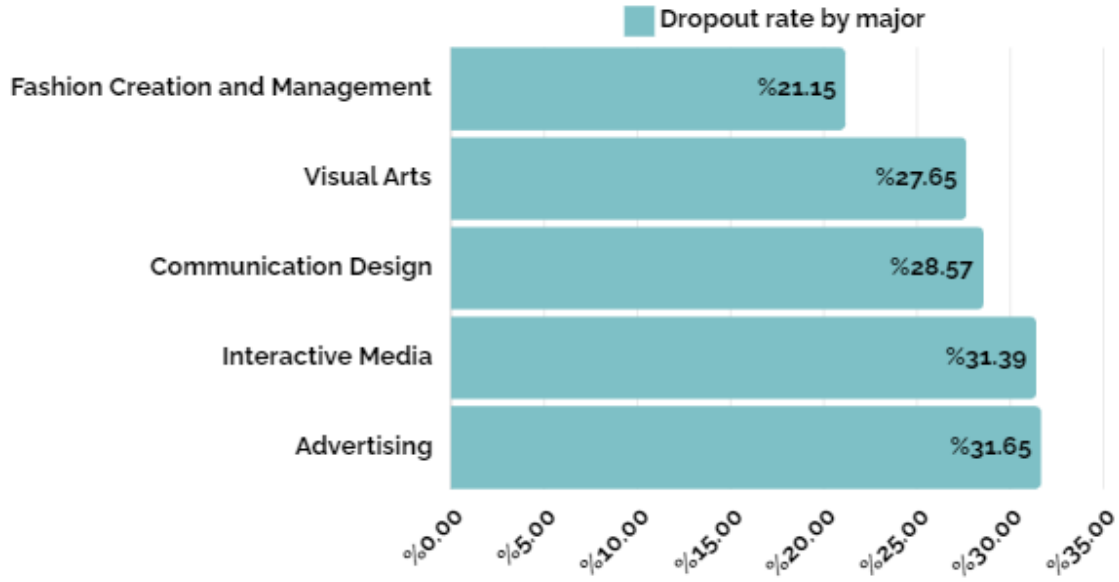
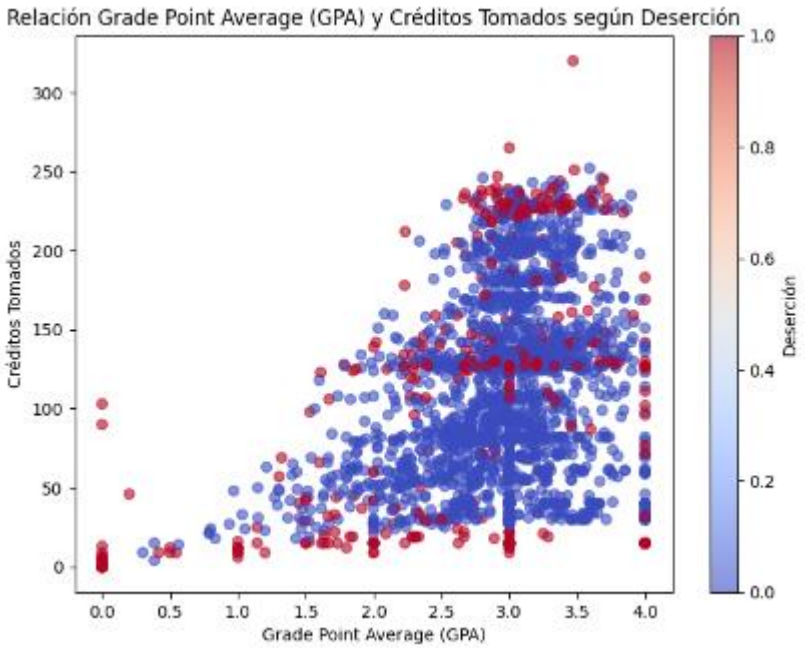


Figure 12
Relationship between student's GPA and credits taken by the student.



APPENDIX R: DROP OUT DATA CLEANING PROCESS

Table 24
Data cleaning process.

Process activity	Activity description
Elimination of duplicate data	All duplicate data were removed.
Processing of missing data from the column “nota_materia”	It was decided to eliminate the Nan (Empty) values since they constitute 8.80% of the data, so they will not significantly affect the representativeness of the data set even though they are close to 10%. The second reason for eliminating them is that the data are from students who are in the current period of analysis and therefore do not yet have all the necessary information.
Processing of missing data from the column “fecha_nacimiento”	It was determined to eliminate the Nan (Empty) values since they constitute 0.11% of the data, so they will not significantly affect the representativeness of the data set.
Processing of missing data from the column “provincia”	It was determined to fill the missing values with “Other” since it is 0.05% of the data and filling it with “Other” will not significantly affect the representativeness of the data set.
Processing of missing data from the column “colegio_procedencia”	It was determined to fill the missing values with “Other College” since it is 0.26% of the data and filling it with “Other College” will not significantly affect the representativeness of the data set.
Processing of missing data from the column “transferido_universidad”	It was determined to fill the Nan (Empty) values with “Not transferred” since according to the database having a transfer name means that it was transferred.
Processing of missing data from the column “gpa_semestral”	It was determined to fill the missing values with 0 since it is 0.28% of the data and filling them with 0 will not significantly affect the representativeness of the data set. In addition, when reviewing the database, it was noted that the empty gpa values come from students who withdrew from the entire semester, but not from students who failed all of their subjects. (The gpa values of 0 will not be taken into account when the average semester gpa is taken).
Processing of missing data from the column “transferido_carrera”	It was determined to fill the Nan (Empty) values with “Not transferred” since according to the database having a transfer name means that it was transferred.
Elimination the column “periodo_transferido_carrera”	It was determined to eliminate the column since the empty data in this column constitutes 90.34% of all its data and cannot be linked to any type of data. In the

	same way, it is not considered crucial since it has the same information in quantity_transferred.
Processing of missing data from the column "nota_examen_ingreso"	It was determined that the empty values of the exam grade are 0.84%, since it is a crucial value for the investigation, it was decided to treat the data, for this purpose it was joined with the outliers adding up to 3.06% of the data. 06% of the data, these outliers became empty and once empty were filled with the average of the data according to the school of entry, this is done because they are a small amount of data and will not significantly affect the representativeness of the data set, in addition to reviewing the graphs can be noted that they follow a distribution close to normal so the filling of the average is adequate.
Processing of missing data from the column "etnia"	It was determined that the amount of empty data is 90.15%, this is a large amount of empty data, these data are filled with "Other" to validate that if they belong to province, only that the university does not have that data.
Processing of missing data from the column "costo_carrera_semestral"	It was determined that the amount of empty data is 0.57%, these data were eliminated since the career to which they were associated no longer exists, i.e., they are data that should not be in the database, so they were eliminated.
Processing of missing data from the column "gpa_colegio"	It was determined that the empty values of the test score are 0.75%, being a crucial value for the research, it was decided to treat the data, for this they were joined with the outliers adding 0.80% of the data. 80% of the data, these outliers became empty and once empty were filled with the mean of the data, this is done because they are a small amount of data and will not significantly affect the representativeness of the data set, in addition to reviewing the graphs can be noted that they follow a distribution close to normal so the filling of the average is adequate.
Processing of missing data from the column "porcentajes"	All 0 values in this column or missing values are correct since they should denote an absence.
Processing of missing data from the column "tipo_ayuda_financiera"	It was determined to fill in the values Nan (Empty) with "No Financial Aid" since according to the base of having this data empty means that the student is not receiving any type of financial aid.
Treatment of outliers in the column "gpa_colegio"	It was treated in the previous section, the outliers, being so few and of human input error, were changed to the correct values.
Treatment of outliers in the column "fecha_nacimiento"	Being 0.04% of the data, the decision was taken to change the age of the type values to 18, since

	according to Stefos in 2019 this is the age at which they are in college.
Treatment of outliers in the column “nota_examen”	It was treated in the previous section, the outliers, being so few, were transformed to the average of all the non-outliers.
Treatment of outliers in the column “porcentaje_total”	It was treated in the previous section, the outliers were eliminated as they gave results greater than 100% and when reviewing the database from which the data was extracted, it was noted that they were erroneous.
Treatment of outliers in the column “cantidad_transferido”	According to the creation of the base it shows us that each time it is within a run it counts so when taking the amount of data from amount_transferred, 2 would be equal to 1 and 3 would be equal to 2.
Treatment of outliers in the columns “materia”, "nota_materia", "equivalencia_materia"	Within the database, there are subjects that have the grade equivalence as “Pass if the subject is from general school / Fail if it is not from general school” so a function was created after doing a manual search of which general school subjects appear and the students obtained a grade of “D”. Those with more letters other than “A - B - C” are considered as “Failed”.
Treatment of outliers in the column “transferido_universidad”	This column does not have incorrect values in its writing, it has many shortened words so we proceeded to write the complete text and replace it with the previous one, all this to have a better handling of the data.
Treatment of outliers in the column “carrera”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “colegio”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “provincia”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “etnia”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “genero”	This column does not have incorrect values in its writing, it only has a letter indicating the value, so it is changed to the full name to have a better data management.
Treatment of outliers in the column “colegio_ingreso”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “carrera_ingreso”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “tipo_ayuda_financiera”	This column does not have incorrect values in its writing, it has many shortened words so we proceeded to write the complete text and replace it with the previous one, all this to have a better handling of the data.

Treatment of outliers in the column “colegio_procedencia”	This column has incorrect values in its writing, many of them for some letter or abbreviation, a manual review is made, followed by the use of the fuzzy function that allows to see the similarity in the whole column and thus to be able to fix the values that are wrongly entered.
Elimination of the columns	"año_academico_registro", "nota_materia", "fecha_nacimiento", "estatus_estudiante", "porcentaje_asistencia_financiera", "porcentaje_beca", "porcentaje_pago_futuro" These columns are eliminated because they contain the same information as other columns and become redundant.
Integration of the columns "equivalencia_materia" a "materia_aprobado" y "materia_reprobado"	The subject equivalence column was eliminated since the aim is to reduce the dimensionality and give more weight to the numerical variables, so the number of subjects passed and the number of subjects not passed were created.
Integration of the column "total_pagado"	In order to better understand how the economic part affects the student, a column was created to see how much he/she has actually paid in comparison to the value of the career.
Integration of the column "cantidad_periodos"	In order to better understand whether it affects the time students are studying within the university, we thus have an additional numerical variable.
Columns "deserto", "transferido_universidad", "transferido_carrera", "genero" value settings	The values of “Yes” are changed to 1 and “No” to 0
Elimination of the columns	"periodo_analisis", "año_academico_ingreso", "año_analisis" These columns are eliminated because they contain the same information as other columns and become redundant.
The column "periodo_materia" is created	This column is created to make effective the creation of dummy variables and also to be able to aggregate the data by period.
Elimination of the columns	The "periodo_registro" column is eliminated , since it served as an anchor for grouping the data.
Grouping data by id	The records are grouped according to the id of the student to have now one record per student.
Expansion of columns	The period_subject column is expanded and now we will have a column for each period and subject that has been taken in it.
Dummy variables are created	Dummy variables of the categorical variables are created.
Data standardization	The numerical values are normalized since they follow a normal distribution and this will allow the

	models to give the same weight to each numerical variable regardless of the extent of the variable.
PCA is performed	A principal component analysis is performed to reduce the dimensionality of the database, thus making it ready for analysis.

APPENDIX S: LIST OF ENROLLMENT VARIABLES AND DESCRIPTION

Table 25
List of all variables provided by the HEI with their detailed description.

Category	Variable Name	Description	DataType
Academic Performance	ID	Previously coded student identifier	int64
	Admission Period	Semester in which the student applied to the University.	int64
	Admission Career	Career to which the student applied within the University.	object
	Admission College	College to which the student applied to	object
	Admission Score	Student's entrance exam score	int64
	Transfer University	University from which the student transferred	object
	Academic Year of Admission	Academic year associated with the student's admission period.	float64
	Completed Registration	If the student completed the registration process, that is, if the student entered the university.	object
Economic Situation	AF Percentage	Percentage of Financial Attendance of the student in the semester "Registration Period".	int64
	Scholarship Percentage	Student's Scholarship Percentage for the "Registration Period" semester.	int64
	Future Payment Percentage	Student's Future Payment Percentage for the "Period of Record" semester.	int64
	Total Financial Aid Percentage	Percentage of total financial aid (Scholarship + Financial Aid + Future Payment) of the student in the "Registration Period" semester.	int64
	Semester Career Cost	Cost associated with the student's career for the year of analysis.	int64

Student Background	School of origin	School where the student attended high school.	object
	High School Grade	Final grade with which the student graduated from high school.	int64
	Sex	Reported gender of the student	object
	Province	Province of residence of the student	object
	Date of Birth	Date of birth of the student	date
	Ethnicity	Ethnicity to which the student belongs.	object

APPENDIX T: ENROLLMENT DATA EXPLORATION

Figure 13
Percentage of students who enroll the HEI.

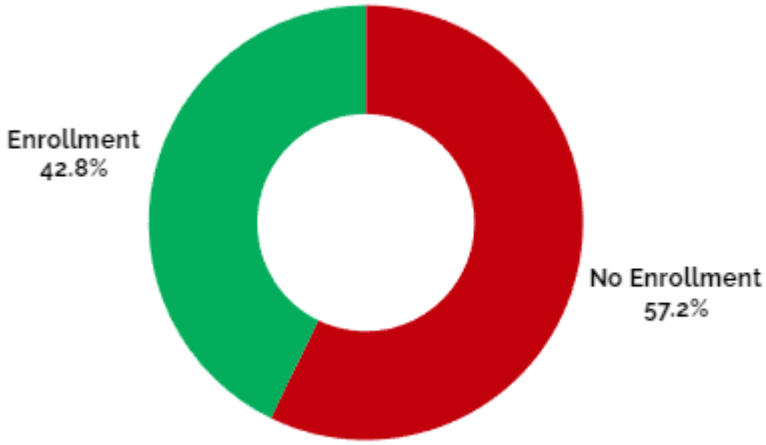


Figure 14
Percentage of enrollment by major

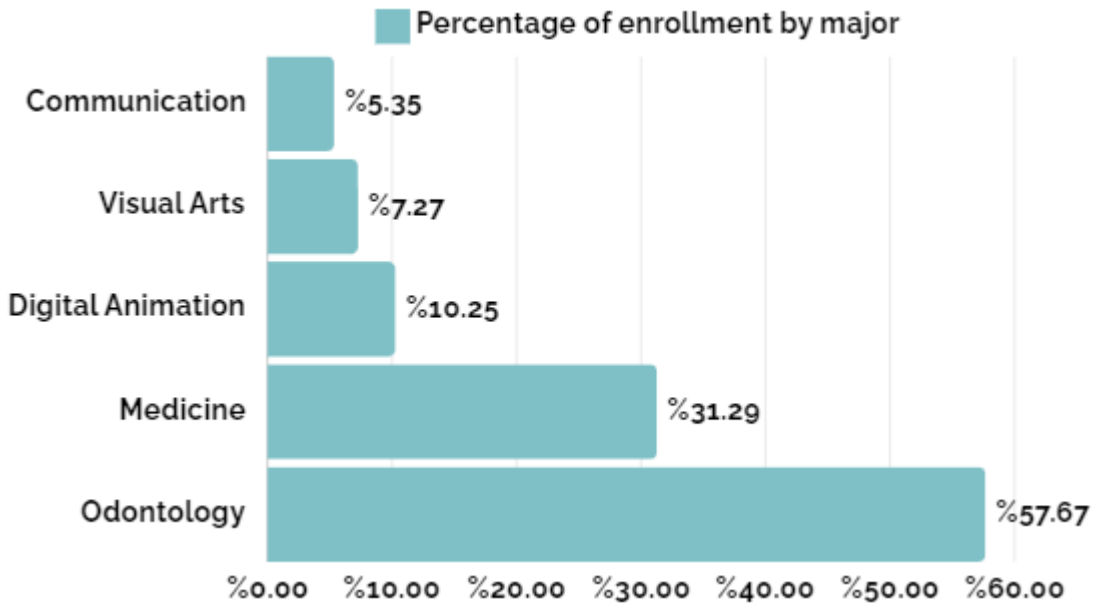
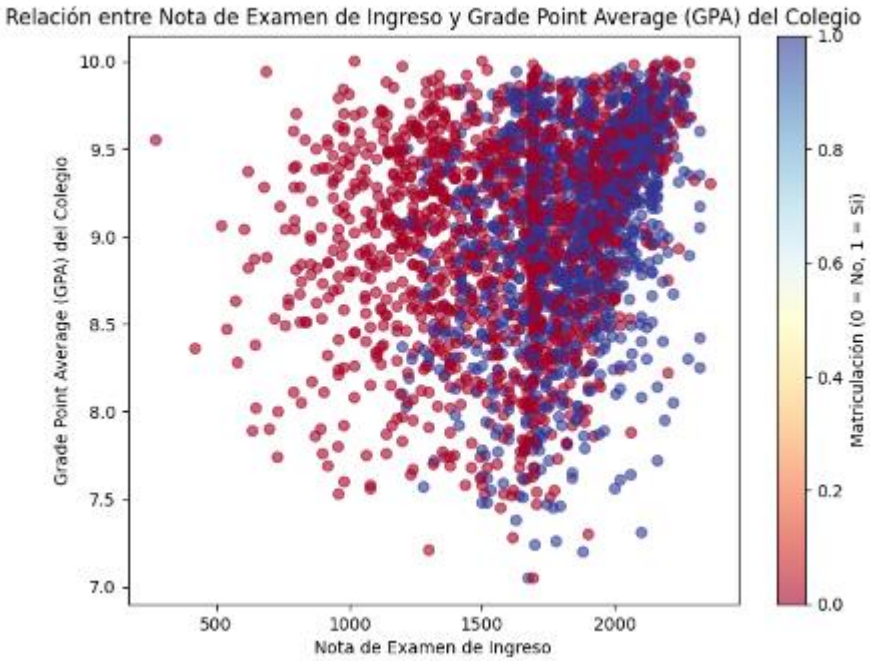


Figure 15
Relationship between the grade point average of the school and the grade achieved in the admission exam.


Relación entre Nota de Examen de Ingreso y Grade Point Average (GPA) del Colegio
Grade Point Average (GPA) del Colegio
Nota de Examen de Ingreso
Matriculación (0 = No, 1 = Sí)
10.0
9.5
9.0
8.5
8.0
7.5
7.0
500
1000
1500
2000
1.0
0.8
0.6
0.4
0.2
0.0

APPENDIX U: ENROLLMENT DATA CLEANING PROCESS

Table 26
Data cleaning process.

Process activity	Activity description
Elimination of duplicate data and rename columns	All duplicate data has been removed and columns have been renamed to be similar to the dropout database.
Processing of missing data from the column "nota_examen_ingreso"	It was determined that the empty values of the exam grade are 6.40%, since it is a crucial value for the investigation, it was decided to treat the data, for this purpose it was joined with the atypical data adding up to 7.60% of the data. 60% of the data, these outliers became empty and once empty were filled with the average of the data according to the school of entry, this is done because they are a small amount of data and will not significantly affect the representativeness of the data set, in addition to reviewing the graphs can be noted that they follow a distribution close to normal so the filling of the average is adequate.
Processing of missing data from the column "transferido_universidad"	It was determined to fill the Nan (Empty) values with "Not transferred" since according to the database having a transfer name means that it was transferred.
Processing of missing data from the column "gpa_colegio"	It was determined that the empty values of gpa_colegion are 0.11% these data were filled with the mean of the data, this is done because they are a small amount of data and will not significantly affect the representativeness of the data set, also when reviewing the graphs it can be noted that they follow a distribution close to normal so the filling of the average is adequate.
Processing of missing data from the column "provincia"	It was determined to fill the missing values with "Other" since it is 0.41% of the data and filling it with "Other" will not significantly affect the representativeness of the data set.
Processing of missing data from the column "fecha_nacimiento"	This column was used as an anchor to create a new column called "edad".
Processing of missing data from the column "edad"	It was determined that the empty values of age are 57.28%, being a crucial value for the research, it was decided to treat the data, taking into account that according to a study conducted in undergraduate students in Ecuador, the average age at which they enter university is 18 years, it was decided to fill the empty data with 18.
Processing of missing data from the column "etnia"	It was determined that the amount of empty data is 90.15%, this is a large amount of empty data, these data are filled with "Other" to validate that if they belong to province, only that the university does not have that data.
Processing of missing data from the column "porcentajes"	All 0 values in this column or missing values are correct since they should denote an absence. It was determined to eliminate outliers (greater than 100%) since they constitute 0.03% of the data, so they will not significantly affect the representativeness of the data set.

Treatment of outliers in the column “gpa_colegio”	It was determined that there is a value that is incorrectly entered, with the value of 5.57, since in Ecuador the minimum grade for graduation is 7/10, the value was changed to 7.57.
Treatment of outliers in the column “edad”	It was determined that the empty values of age are 0.22%, being a crucial value for the research, it was decided to treat the data, for this they were filled with the average of the same, this because they are a small amount of data and will not significantly affect the representativeness of the data set, in addition to reviewing the graphs it can be noted that they follow a distribution close to normal so that filling the average is adequate.
Treatment of outliers in the column “nota_examen”	It was treated in the previous section, the outliers, being so few, were transformed to the average of all the non-outliers.
Treatment of outliers in the column “porcentaje_total”	It was treated in the previous section, the outliers were eliminated as they gave results greater than 100% and when reviewing the database from which the data was extracted, it was noted that they were erroneous.
Treatment of outliers in the column “transferido_universidad”	This column does not have incorrect values in its writing, it has many shortened words so we proceeded to write the complete text and replace it with the previous one, all this to have a better handling of the data.
Treatment of outliers in the column “carrera_admision”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “colegio_admision”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “provincia”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “etnia”	This column does not have incorrect values in its writing.
Treatment of outliers in the column “genero”	This column does not have incorrect values in its writing, it only has one letter indicating the value, so it is changed to the full name to have a better data management.
Treatment of outliers in the column “colegio_procedencia”	This column has incorrect values in its writing, many of them for some letter or abbreviation, a manual review is made, followed by the use of the fuzzy function that allows to see the similarity in the whole column and thus to be able to fix the values that are wrongly entered.
Elimination of the columns	"fecha_nacimiento", "porcentaje_asistencia_financiera", "porcentaje_beca", "porcentaje_pago_futuro" These columns are eliminated because they contain the same information as other columns and become redundant.
Columns "matriculo", "transferido_universidad" value settings	The values of “Yes” are changed to 1 and “No” to 0

Data standardization	The numerical values are normalized since they follow a normal distribution and this will allow the models to give the same weight to each numerical variable regardless of the extent of the variable.
Dummy variables are created	Dummy variables of the categorical variables are created. thus making it ready for analysis.

APPENDIX V: HYPERPARAMETERS

Table 27
Model's Hyperparameters

Algorithm	Hyperparameters	Values	Reference
Logistic Regression	C	0.001, 0.01, 0.1, 1, 10, 100	(Ahmed Arafa et al., 2022; Ambesange et al., 2020)
	Penalty	l1', 'l2'	
	Solver	liblinear', 'saga'	
Support Vector Machine	C	0.001, 0.01, 0.1, 1, 10, 100	(Kalita et al., 2020; Thanh Ngoc et al., 2021)
	Kernel	linear', 'rbf'	
	Gamma	scale', 'auto'	
Random Forest	n_estimators	50, 64, 100, 128, 150, 200	(Oshiro et al., 2012; Probst et al., 2019)
	max_depth	None, 10, 15, 20, 30	
XGBoost	learning_rate	0.05, 0.1, 0.15	(Dalal et al., 2022; Xiong et al., 2022)
	n_estimators	100, 128, 150	
	max_depth	5, 7, 9	