UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Predictive Models for University Dropout at UNIRIO

Henrique Soares Rodrigues

Orientadora

Laura Moraes

Co-orientador

Reinaldo Alvares

RIO DE JANEIRO, RJ - BRASIL
JANEIRO DE 2025

Predictive Models for University Dropout at UNIRIO

Henrique Soares Rodrigues

DISSERTAÇÃO DE MESTRADO APRESENTADA COMO REQUISITO OBRIGATÓRIO ESTIPULADO PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA (PPGI) DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Laura O. Moraes, D.Sc. — (UNIRIO)

Reinaldo Alvares, D.Sc. — (UNIRIO)

Jefferson Elbert, D.Sc. — (UNIRIO)

Carla Delgado, D.Sc. — (UFRJ)

RIO DE JANEIRO, RJ - BRASIL
JANEIRO DE 2025.

Catalogação informatizada pelo autor

Rodrigues, Henrique.
R696 Predictive Models for University Dropout at UNIRIO/ Henrique Soares Rodrigues, 2025.
73p.

Orientadora: Laura O. Moraes
Coorientador: Reinaldo Alvares
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, Programa de Pós-Graduação em Informática, 2025

1. Dropout I 2. University 3. Artificial Intelligence 4. Data Science 5. Student I. Moraes, Laura, orient. II. Alavares, Reinaldo. III Predictive Models for University Dropout at UNIRIO

"Educating the mind without educating the heart is not education."
(Aristotle)

Acknowledgements

I would like to primarily thank God. Secondly to my family: my parents Vladimir Rodrigues and Márcia Soares, my grandmother Wanda Bonifácio, my aunt Simone Soares, my sisters Liliana Rodrigues and Rosana Rodrigues and my nephews Enzo Rodrigues and Daniel Rodrigues. In third place to my advisor Laura Moraes and my co-advisor Reinaldo Alvares, to professors Rodrigo Santos, Ana Cristina Bicharra, and Carlos Eduardo Mello, to my co-authors Gabriel Xará, João Porto, Elmo Júnior and Eduardo Santiago and my researcher peers Herick Henrique and Rafael Carrion. Fourthly to my friends, especially André Felipe and Thiago Parracho, who are the not-blooded siblings. Last but not least to Universidade Federal do Estado do Rio de Janeiro (UNIRIO) and Coordenação de Aperfeiçoamento de Ensino Superior (CAPES).

Henrique Soares Rodrigues.

Rodrigues, Henrique **Predictive Models for University Dropout at UNIRIO**. UNIRIO, 2025. 73 páginas. Dissertação de mestrado. Programa de Pós-Graduação em Informática, UNIRIO.

RESUMO

Instituições de Ensino Superior (IES) almejam o sucesso acadêmico e profissional de seus alunos e egressos, porém, estas instituições enfrentam o problema da evasão de alunos, isto é, quando os alunos abandonam em definitivo os seus cursos. A evasão de alunos do ensino superior pode causar problemas no âmbito pessoal, universitário e social. No âmbito social a evasão pode ocasionar uma menor renda para o evadido. Na esfera universitária pode ocasionar perda de recursos, e em nível social pode ocasionar escassez de profissionais. O objetivo desta pesquisa é usar técnicas de ciência de dados e Inteligência Artificial (IA) preditiva para compreender o fenômeno da evasão universitária nos cursos de ciências exatas da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Para alcançar tais objetivos, foram conduzidos três estudos: um Mapeamento Sistemático da Literatura (MSL) sobre algoritmos de IA para prever evasão de estudantes de ensino superior, um estudo primário para prever evasão no curso de Bacharelado em Sistemas de Informação (BSI) usando Decision Tree e apenas dados do Sistema de Informações para o Ensino (SIE) e um estudo primário sobre previção de evasão nos cursos de ciências exatas, oferecidos pelo Centro de Ciências Exatas e Tecnologia (CCET) da UNIRIO, que além do BSI engloba também os cursos de Licenciatura em Matemática e o Bacharelado em Engenharia de Produção, usando Gradient Boosting com os dados do Sistema de Informações para o Ensino (SIE), da Receita Federal e da Relação Anual de Informações Sociais (RAIS) para entender se há influência de presença de emprego formal, excluindo estágio, e empreendedorismo até o quarto período na evasão. Resultados demonstram que os principais fatores preditivos para a evasão ou a graduação são o desempenho acadêmico e que não há influência da presença de emprego e empreendedorismo na evasão.

Palavras-chave: Evasão, Universidade, Inteligência Artificial, Ciência de Dados, Estudante.

ABSTRACT

Higher Education Institutions (HEIs) aim for the academic and professional success of their students and graduates; however, these institutions face the problem of student dropout, which is when students permanently abandon their courses. Higher education student dropout can cause problems in the student's personal sphere, in the university sphere, and in the social sphere. In the social sphere, dropout can lead to lower income for students, a loss of resources for universities, and a shortage of professionals in society. The objective of this research is to use data science and predictive Artificial Intelligence (AI) techniques to help understand the phenomenon of university dropout in exact sciences courses at the Federal University of the State of Rio de Janeiro (UNIRIO). To achieve these objectives, three studies were conducted: a Systematic Mapping Study (SMS) on AI algorithms to predict higher education student dropout, a primary study on predicting dropout in the Bachelor of Information Systems course (BSI) using Decision Tree and only data from the Academic System (SIE) and a primary study on predicting dropout rates in STEM (Science, Technology, Engineering and Maths) courses at the Center for Exact Sciences and Technology (CCET) at UNIRIO, which in addition to BSI also encompasses the Degree courses in Mathematics and the Bachelor's Degree in Production Engineering, using Gradient Boosting with data from the Academic System (SIE), the Federal Revenue Service and the Annual Social Information List (RAIS) to understand whether there is an influence on the presence of formal employment, excluding internships, and entrepreneurship up to the fourth period in the dropout. Results demonstrate that the main predictive factors for dropout or graduation are academic performance and that there is no influence of the presence of employment and entrepreneurship on dropout.

Keywords: Dropout, University, Artificial Intelligence, Data Science, Student.

Contents

1 Introduction **1**

1.1 Context 1

1.2 Motivation 2

1.3 Objective 2

1.4 Contributions 3

1.5 Planned studies 4

1.6 Organization 5

2 Background **6**

2.1 Ontology of student's dropout 6

2.2 Machine Learning algorithms 8

2.2.1 Algorithms to predict categorical features 9

2.2.1.1 Decision Tree 9

2.2.1.2 Random Forest 10

2.2.1.3 Gradient Boosting 11

2.2.2 Algorithms to predict numerical features 13

2.2.2.1 Linear Regression 14

2.3 Main metrics of AI prediction algorithms 14

2.3.1 Metrics for models that predict categorical features 14

2.3.1.1 Accuracy 15

2.3.1.2 Precision 16

2.3.1.3 Recall 16

2.3.1.4 F1 Score 16

2.3.2 Metrics for models that predict numerical features 16

2.3.2.1 R-Value 17

2.3.2.2 Root Mean Square Error (RMSE) 17

2.4 Chi Square Test 18

3 Systematic Mapping Study **19**

3.1 Systematic mapping study about dropout prediction 19

3.1.1 Introduction 19

3.1.2 Related Work 19

3.1.3 Research Method 20

3.1.3.1 Search Strategy and Data Source 20

3.1.3.2 Search String 21

3.1.3.3 Selection Criteria 21

3.1.3.4 Study Selection Process 21

3.1.3.5 Data Extraction 21

3.1.4 Results 22

3.1.4.1 Sources of Studies 22

3.1.4.2 Filtering 23

3.1.4.3 Country of Origin of the Studies 23

3.1.4.4 Algorithms in the Studies 23

3.1.5 (Sub-Q1): What are the difficulties in using AI to predict university dropout rates? 28

3.1.6 (Sub-Q2): How do AI algorithms use features to predict higher institutions' dropout? 28

3.1.7 Discussion 29

3.1.8 Threats to validity 30

3.1.9 Conclusion 30

4 Method **32**

4.1 Ethical Assessment 32

4.2 Cross Industry Standard Process for Data Mining 32

4.3 Method of the two following chapters 33

4.3.1 Chapter 5: exploratory analysis and preliminary study 33

4.3.1.1 Chi-Square Tests' Hypothesis 33

4.3.2 Chapter 6: follow-up and final study 34

5 Data Understanding and Exploratory Analysis **35**

5.1 Database 35

5.2 Dropout rate formula 37

5.3 Data analysis in Information System course using only academic data . . 38

5.3.1 Model: Decision Tree 41

5.3.2 (Sub-RQ1): What are the most determining variables to predict dropout in BSI at UNIRIO? 42

5.3.3 (Sub-RQ2): In which years was there the highest dropout rate in BSI at UNIRIO? 42

5.3.4 (Sub-RQ3): Which curricular activities are the most decisive for dropout in BSI at UNIRIO? 43

5.4 Data analysis in STEM courses using academic and financial data 44

5.4.1 (sub-RQ4) Is there an association between full-time employment and dropout rates at CCET? 45

5.4.2 (Sub-RQ5): Is there an association between entrepreneurship and dropout rates at CCET? 46

5.4.3 (Sub-RQ6): Do university scholarships help reduce dropout rates at CCET? 47

6 Modeling **49**

6.1 Gradient Boosting Models 49

6.2 Process of training the models 50

6.3 Results 50

6.3.1 (Sub-RQ7): Does a model that uses financial factors along with academic data have greater predictive power than a model that considers only academic data? 57

6.3.2 Gender fairness in the model 58

6.4 Discussion 58

7 Deployment **60**

7.1 Architecture of the Dropout Predictor System 60

7.2 Use Examples of the Dropout Predictor System 61

8 Final Remarks **63**

8.1 Conclusion 63

8.2 Limitations 63

8.3 Future Works 64

References **65**

APPENDIX A – Dropout Prediction System Use Cases Description 73

List of Figures

1.1 Process of the research 5

2.1 The ontology of dropout. Source: the author 6

2.2 Dropout decision. Reference: (Tinto, 1975) 8

2.3 The workflow of machine learning. Reference: (Osman, 2019) 9

2.4 Random Forest. Reference: (Brital, 2021) 11

2.5 The workflow of Gradient Boosting algorithm. Reference: (Hemashreekilari, 2023) 13

2.6 Linear regression formula. Reference: (PennState, 2018) 14

2.7 Possible correlations. Reference: the author 17

3.1 Studies' filtering process 23

3.2 Studies' countries 24

3.3 Algorithms explored in the selected studies 25

4.1 Cross Industry Standard Process for Data Mining. Reference: (Provost; Fawcett, 2016) 33

5.1 The database 35

5.2 Accumulated GPA by outcome 39

5.3 Semester GPA by outcome 40

5.4 Curricular activities grades by outcome 40

5.5 Average semester GPA by outcome 40

5.6 Correlation: academic performance X status (graduation/dropout) 41

5.7 Decision tree of the model focused on the former student 41

5.8 Dropout of students per years since the enrollment 43

5.9 Top 10 curricular activities with the most student's failures 44

5.10 Dropout rates on CCET 45

5.11 Accumulated GPA of employee-students at CCET 45

5.12 Accumulated GPA of students who become company owners at CCET . . 45

5.13 Category of the company owners of CCET 47

5.14 Dropout rates by admission methods and scholarships 48

5.15 Accumulated GPA per students that have scholarship and students that don't have scholarship 48

6.1 Study process 50

6.2 F1 Score across the models 56

7.1 Dropout Predictor system: examples of prediction 60

7.2 Dropout Predictor system: examples of prediction 60

7.3 Dropout Predictor system: example of prediction of graduation 61

7.4 Dropout Predictor system: examples of prediction of dropout 62

7.5 Dropout Predictor system: examples of CSV input page 62

7.6 Dropout Predictor system: examples of prediction logs 62

List of Tables

2.1 Theoretical Confusion Matrix. 15

3.1 PIO structure to formulate the research question. 20

3.2 Selection Criteria. 22

3.3 Number of studies by source. 22

3.4 Results extracted from the studies 26

3.5 Results extracted from the studies 27

5.1 Database features 36

5.2 Database features: Information Systems Curricular Activities 37

5.3 Database features: Production Engineering Curricular Activities 37

5.4 Database features: Mathematics Curricular Activities 37

5.5 Graduation or dropout by gender 38

5.6 Graduation or dropout by admission method 39

5.7 Model accuracy and classification report 42

5.8 Graduation or dropout by full-time work 46

5.9 Graduation or dropout by company ownership 46

5.10 Graduation or dropout by presence of scholarship 47

6.1 Information System's model's accuracy and classification report 52

6.2 Production Engineering's models accuracy and classification report 53

6.3 Mathematics's models accuracy and classification report 54

6.4 CCET's general models accuracy and classification report 55

6.5 Analysis of the importance of variables for the CCET model 1 55

6.6 Overall accuracy and classification report without financial data: CCET model 1 57

6.7 Overall accuracy and classification report without financial data: CCET model 4 57

6.8 CCET Model 1 accuracy and classification report for male students 58

6.9 CCET Model 1 accuracy and classification report for female students . . . 58

1. Introduction

1.1 Context

Higher Education Institutions (HEIs) aim for their students to experience academic and professional success, as this contributes to economic growth and social justice. However, one of the most problematic issues that HEIs face is student dropout (Realinho *et al.*, 2022). 48% of the students who enrolled in federal universities in 2015 in Brazil dropped out, filtering by only STEM courses, the dropout rate reached 55% (Brasil, 2023).

According to (Bardagi; Hutz, 2005), reducing dropout rates in HEIs is not only an educational issue, but also an economic and political one, causing social and economic losses for students, society, and HEIs (Prestes; Fialho, 2018), also causing a shortage of professionals in several areas, compromising an entire necessary ecosystem (Saccaro *et al.*, 2019). Therefore, reducing dropout rates can have a positive impact on students' professional and financial trajectories, as well as reducing the waste of resources at HEIs.

To address the problem of student dropout, especially in Higher Education Institutions (HEIs), Artificial Intelligence (AI) algorithms have been recognized as potential tools (Silva; Roman, 2021; Tete *et al.*, 2022). They can identify students at risk of leaving educational institutions, allowing them to develop policies that support students in continuing their studies until graduation.

By identifying patterns and risk indicators, AI algorithms can anticipate situations in which students are likely to drop out of their studies (Tete *et al.*, 2022). This early detection provides HEIs with the opportunity to intervene proactively, implementing personalized strategies to support struggling students. Such strategies may include tutoring programs, academic advising, specific pedagogical interventions, workshops to prepare high school students for university life, and even financial support measures when necessary (Cruz-Campos *et al.*, 2023).

According to (Foerster, 2003), First-Order Cybernetics is the study of relationships between entities in a system, such as the educational system, where it can be classified as entities of the educational system both the HEI and the student. One important aspect of such a relationship is the concept of feedback, where one entity's behaviour is influenced by what happens with other entities. Foerster introduces the concept of Second-Order

Cybernetics, where each entity is aware of how its own actions influence the system based of a feedback of feedback.

In the educational system, the final status of a student (if they graduated successfully or dropped out) can be seen as feedback to the HEI, which can give the students feedback of feedback in the form of new educational policies. In order for the HEIs to be able to give this feedback of feedback in the form of new educational policies and curricular reforms, it is important that the HEI know the data of the success or failure of its students as the feedback from the students to the HEIs.

Therefore, this research can be seen as a form to collect feedback from former students to make it possible for the university to give feedback to those feedbacks to current students, trying to minimize the dropout rates.

1.2 Motivation

The motivation of this research is to help academic managers of HEIs, specially in STEM courses, to understand the phenomenon of students' dropout and to offer a technological solution to identify students at risk of dropout aiming to let the institutions think in strategies to prevent such students of abandoning their studies.

1.3 Objective

The objective of this research is to explore how AI algorithms can be used to predict and better understand the factors that influence university dropout in STEM (Science, Technology, Engineering, and Mathematics) courses.

This research focuses on early dropout prediction, defined in this work as prediction during the first two years of the course. To achieve this objective, this research also aims to identify other factors that can influence dropout beyond academic performance because, in the first years, there is not enough data related to academic performance. Therefore, beyond academic performance, this study also focuses on financial-related factors.

The scope of this research is the undergraduate programs of the Center of Exact Sciences (CCET) of the Federal University of the State of Rio de Janeiro (UNIRIO), located in the city of Rio de Janeiro, Brazil, which is Information Systems (BSI), Production Engineering and Mathematics.

BSI and Production Engineering have their classes on afternoon and night for the first half of the courses, and only at night for the second half. Mathematics has classes at night throughout the course.

This research aims to answer the following research question (RQ): *"What are the factors and characteristics of academic dropout in STEM courses of UNIRIO?"* To help answer this research question, the following sub-questions were created:

(Sub-RQ1): "What are the most determining variables to predict dropout in BSI at UNIRIO?"

(Sub-RQ2): "In which years was there the highest dropout rate in BSI at UNIRIO?"

(Sub-RQ3): "Which curricular activities are the most decisive for dropout in BSI at UNIRIO?"

(Sub-RQ4): "Is there an association between full-time employment and dropout rates at CCET?"

(Sub-RQ5): "Is there an association between entrepreneurship and dropout rates at CCET?"

(Sub-RQ6): "Do university scholarships help reduce dropout rates at CCET?"

(Sub-RQ7): "Does a model that uses financial factors along with academic data have greater predictive power than a model that considers only academic data?"

1.4 Contributions

This research aims to offer the following contributions to academic research and to the higher education:

1. AI models to predict HEIs's student's dropout using academic and financial data.
2. An analysis of the impact of financial factors on the phenomenon of dropout on UNIRIO's STEM courses.
3. A web system that uses the AI models to identify UNIRIO's STEM courses' students at risk of dropout.

1.5 Planned studies

According to (Baker *et al.*, 2011), Educational Data Mining (EDM) is the application of Data Mining and Machine Learning (ML), which is a subset of Artificial Intelligence to discover knowledge in the context of education (Santos *et al.*, 2021). The main features used to predict dropout are related to academic performance (Moseley; Mead, 2008), personal factors, such as physical and mental health (Osorio; Santacoloma, 2023), socioeconomic (Realinho *et al.*, 2022), interaction with Learning Management Systems (LMS) (Freire *et al.*, 2024) and institutional, such as satisfaction with HEIs (Oliveira; Medeiros, 2024) and if the student has a scholarship (Alvim *et al.*, 2024)) (Alban; Mauricio, 2019). This research uses EDM to achieve the objective of predicting university dropout and students' academic performance. It was planned to conduct four studies:

1. A Systematic Mapping Study (SMS) to understand the status quo of the research in the field of AI algorithms to predict dropout at HEIs. Published in proceedings of the International Conference on Agents and Artificial Intelligence 2024 (ICAART) (Rodrigues *et al.*, 2024a).

2. A primary study to predict dropout of students of Information Systems of UNIRIO using only data from the academic system SIE. Published in Proceedings of the Workshop of Education on Computing (WEI) on Congress of the Brazilian Computer Science Society 2024 (CSBC) (Rodrigues *et al.*, 2024b).

3. A primary study to predict student dropout of UNIRIO STEM courses using data from SIE, CNPJ RAIS and scholarships.

Figure 1.1 represents the process of the research, which uses the Business Process Modelling Notation (BPMN) (White, 2004).

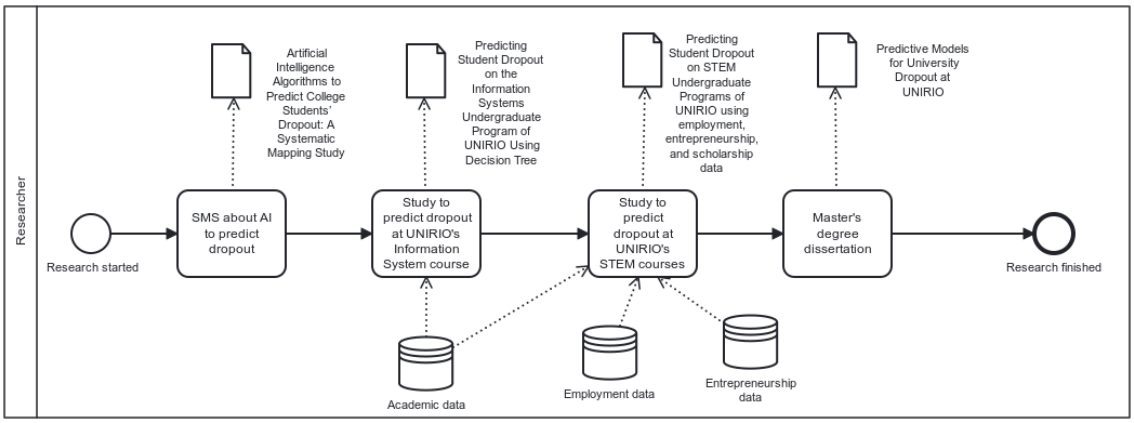


Figure 1.1: Process of the research

1.6 Organization

This research is structured as follows:

- Chapter 2 details the theoretical background of this work.
- Chapter 3 details the systematic mapping studies conducted to understand how AI is used to predict dropout and predict academic performance.
- Chapter 4 details the method used in this research.
- Chapter 5 details the database and the exploratory analysis on the data.
- Chapter 6 details how the final dropout prediction model was built and provides its results.
- Chapter 7 details the web system that was built using the final prediction model
- Chapter 8 provides final remarks, limitations, and future works.

2. Background

2.1 Ontology of student's dropout

Figure 2.1 represents the ontology of dropout of HEI students that was based on systematic mapping studies presented in Chapter 3 (Rodrigues *et al.*, 2024a). The dimensions of the factors that can influence dropout were proposed by (Tete *et al.*, 2022).

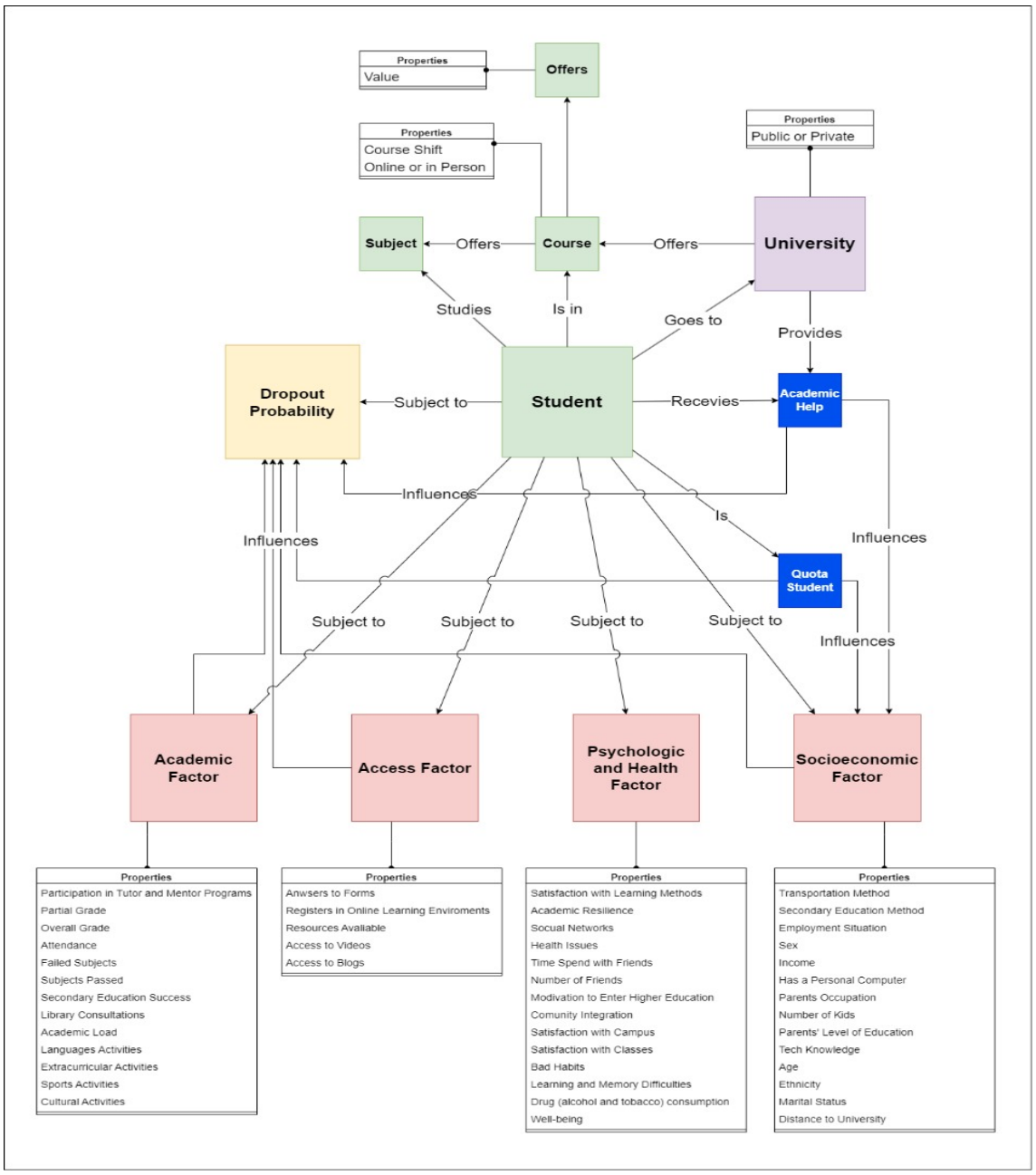


Figure 2.1: The ontology of dropout. Source: the author

The ontology describes the conceptual structure and key mapped elements by the literature (Silva; Roman, 2021; Tete *et al.*, 2022; Rodrigues *et al.*, 2024a) that are essential for the analysis of student dropout in higher education institutions through the use of Artificial Intelligence algorithms.

Clarity and the absence of ambiguity in objects and concepts are crucial to ensure understanding of the ontology. "HEIs" represent educational organizations, such as universities, colleges, and institutes, where students seek higher education. "Student dropout" refers to the phenomenon of students leaving their study programs before completing the course and obtaining a degree. Artificial Intelligence algorithms in this context are restricted to computational methods that use machine learning techniques to predict, in this research, student dropout based on various variables, known as Predictive Variables.

Predictive Variables are used as inputs in Artificial Intelligence algorithms to predict "Student Dropout," and the results of the algorithm predictions determine whether a student is identified as "at risk of dropout" or not. The Predictive Variables include categories such as "social economic", "academic," "psychological and health," and "accessibility," each with their input attributes.

Key concepts include "Student," "University," "Course," "Probability of Dropout," "Academic Factor," "Access Factor," "Psychological and Health Factor," and "Socioeconomic Factor." The relationships and connections include "Subject to," "Influence," "Attends," "Offered," "Providides," "Receives," and "Is in." The ontology shows that the "Student" is "subject to" factors of the categories, which "influence" the probability of "Student Dropout." "Students" "attend" the "University" in which they are part of a "Course" that holds "Value" and is "Offered" by the "University." The "Student" receives "Student Assistance," and being a "Scholarship Recipient" "influences" the "Socioeconomic Factor."

This ontology is used to model the conceptual structure of the research; it is used to understand how concepts are interconnected and how AI algorithms are applied to predict student dropout. Furthermore, the ontology guides the analysis and interpretation of the results, allowing us to identify key factors that influence the dropout of students.

Constructing this ontology is essential to provide a solid conceptual framework guiding our study on the prediction of student dropout in HEIs. It helps to clarify the complexity of the problem and facilitates the analysis of results, contributing to a deeper understanding of the phenomenon of student dropout.

Figure 2.2 represents a theoretical framework describing the factors regarding the decision of a student to drop out. Factors presented in the ontology can be found as decision influences on academic dropout in the framework, such as socioeconomic factors (family background), and psychological and health factors (individual attributes and social integration) and academic factors (pre-college schooling), which can influence goal commitment to academic performance, which influences the students' decision to drop out or not.

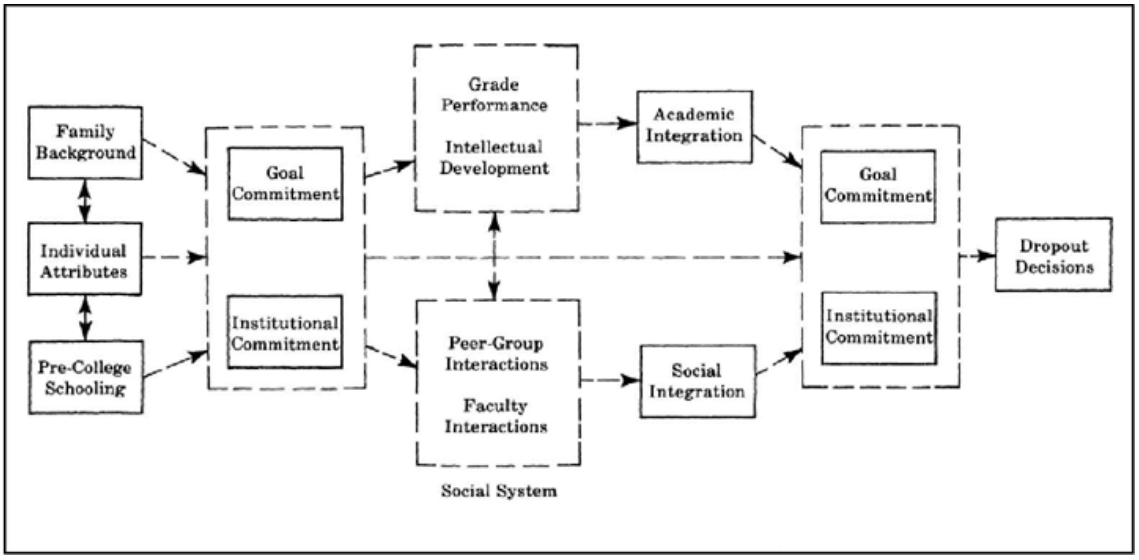


Figure 2.2: Dropout decision. Reference: (Tinto, 1975)

2.2 Machine Learning algorithms

According to (Samuel, 1967), machine learning is the ability of computers to learn without the need for programming them. In this research, predictive models of machine learning, which is a field of AI, are used to predict students' dropout and academic performance. A predictive model is defined by a function:

$$f(X, \beta) = Y$$

where X is the set of predictive features, (β) is the unknown factors that can influence the outcome (Baker *et al.*, 2011) and Y is the predicted variable, which in the context of this research can be the final status of a student or their academic performance.

The models of machine learning use mathematical functions to predict a target feature, known as y, from the other features of the dataset, known as x. Figure 2.3 represents the machine learning workflow.

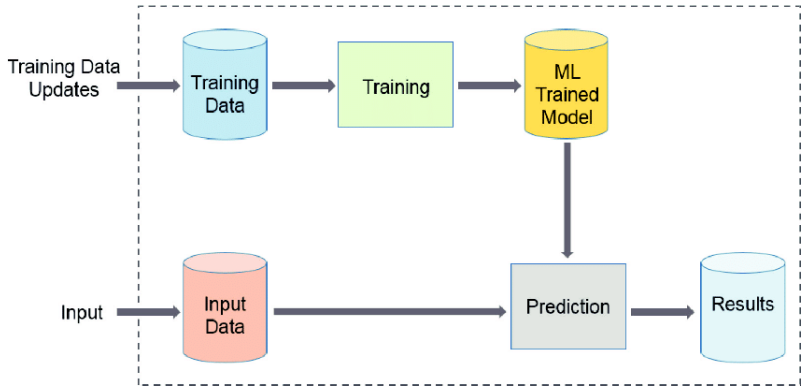


Figure 2.3: The workflow of machine learning. Reference: (Osman, 2019)

2.2.1 Algorithms to predict categorical features

These algorithms can be used to predict categorical features, such as if the student graduated or dropped out, if the student passed or failed a discipline the categorical grade the student get (for example: A, B, C, D, E or F) or the range of numeric grades (for example: 1 to 3.9, 4 to 6.9 or 7 to 10).

2.2.1.1 Decision Tree

Decision tree is an AI algorithm that classifies and predicts a target feature based on another feature following a path on a tree (Fürnkranz, 2010), which is composed of a root node, internal nodes, and leaves nodes. Each decision path on the tree represents a rule that the algorithm identified in the training dataset by Gini impurity or entropy. In this research, the Gini impurity criteria were used. Following the path-making decisions, the algorithm can predict the target variable. The Gini Impurity can be calculated by:

$$\text{Gini}(S) = 1 - \sum_{i=1}^{c} \left(\frac{|S_i|}{|S|} \right)^2$$

where:

- S is the dataset,
- c is the number of classes,
- $|S_i|$ is the number of occurrences of class i in S,
- $|S|$ is the total size of the dataset.

Recursively, the algorithm uses Gini to find the feature with minimum Gini impurity, which means the feature where there is clear distinction on data class. After that, the

algorithm builds a node, which represents a rule. Algorithm 1 describes the steps taken by the Decision Tree algorithm. The attribute in the algorithm that best classifies examples is calculated by Gini Impurity.

One problem with Decision Tree is overfitting, where the tree that was built is too fit to the training data, reaching a false high accuracy, since the tree would have difficulty to predict correctly new data, which is the objective. To avoid overfitting, one possible solution is to establish a minimum Gini Impurity at the trade-off of lowing the accuracy in training data.

Algorithm 1 Decision Tree Algorithm. Reference: an adaptation of (Michalski *et al.*, 2014)

Input: Examples, Target.attribute, Attributes
Output: A decision tree that classifies the examples

```
if all Examples are positive then
    return the single-node tree Root, with label = +
if all Examples are negative then
    return the single-node tree Root, with label = -
if Attributes is empty then
    return the single-node tree Root, with label = most common value of Target.attribute
      in Examples
Create a Root node for the tree
A ← the attribute from Attributes that best classifies Examples
The decision attribute for Root ← A
for each possible value, v_i, of A do
    Add a new tree branch below Root, corresponding to the test A = v_i  Let Examples_{v_i}
      be the subset of Examples that have value v_i for A  if Examples_{v_i} is empty then
        Then below this new branch add a leaf node with label = most common value of
          Target.attribute in Examples
    Else Below this new branch add the subtree: ID3(Examples_{v_i}, Target.attribute,
      Attributes - {A})
return Root
```

2.2.1.2 Random Forest

Random Forest is an AI algorithm introduced by (Breiman, 2001) that classifies and predicts a target feature based on the majority of votes of several decision trees, which are built with samples of features, contrary to the Decision Tree algorithm, that a single tree uses all the features available. Figure 2.4 shows an example of the random forest voting system. Algorithm 2 describes the steps taken by the Random Forest algorithm.

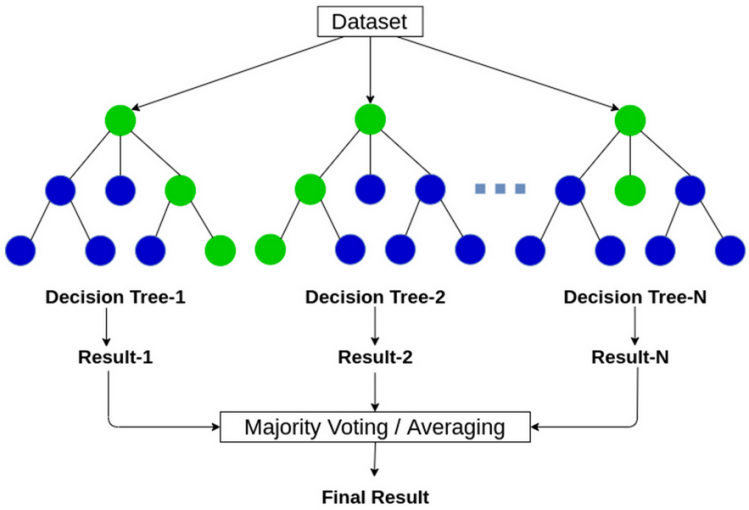


Figure 2.4: Random Forest. Reference: (Brital, 2021)

2.2.1.3 Gradient Boosting

Gradient Boosting is an AI algorithm introduced by (Friedman, 2000) that classifies and predicts a target feature based on boosting generations of decision trees. This algorithm starts with a decision tree, and then the residual, which is the difference between the predicted values and the true values is calculated by the loss function:

$$L(y, F(x)) = \sum_{i=1}^{n} L(y_i, F(x_i))$$

where:

- n is the number of training examples.
- y_i is the true value for the i-th example.
- $F(x_i)$ is the predicted value for the i-th example.
- $L(y_i, F(x_i))$ is the loss for the i-th example, typically chosen based on the problem at hand (e.g., mean squared error for regression, log loss for classification).

The learning rate of each new model added to the ensemble is given by:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{2.1}$$

where:

Algorithm 2 Random Forest algorithm. Reference: an adaptation of (Guo *et al.*, 2021)

Input: Training data D, number of classifiers c, subset percentage $x\%$

Output: A random forest of decision trees

```
for i = 1 to c do
    Randomly sample the training data D with replacement to produce D_i  Create a root node, N, containing D_i  Call BuildTree(N)

Function BuildTree(N):
    if N contains instances of only one class then
        return
    else
        Randomly select x% of the possible splitting features in N  Select the feature F with the highest information gain to split on  Create f child nodes of N, N_1, ..., N_f, where F has f possible values (F_1, ..., F_f)
        for i = 1 to f do
            Set the contents of N_i to D_i, where D_i is all instances in N that match F_i  Call BuildTree(N_i)
```

- $F_m(x)$: The model prediction after the m-th iteration.
- $F_{m-1}(x)$: The model prediction after the $(m-1)$-th iteration.
- η: The learning rate, a hyperparameter that controls the contribution of each weak learner.
- $h_m(x)$: The m-th weak learner, trained to correct the residual errors from the previous model.

The learning rate η determines the step size at each iteration while moving towards a minimum of the loss function.

After several interactions to boost the decision trees, the final model $F_M(x)$ in Gradient Boosting is given by:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \gamma_m h_m(x)$$

where:

- $F_0(x)$ is the initial model, often chosen as a constant value.
- M is the number of boosting iterations.
- γ_m is the multiplier (step size) for the m-th base learner.

- $h_m(x)$ is the m-th base learner (e.g., a decision tree).

Figure 2.5 represents the visualization of the workflow of the Gradient Boosting algorithm, described in Algorithm 3. At the end, the final model is the ensemble of each individual decision tree, known as base learners. Differently from Random Forest, the decision trees with less errors has more influence because of the γ_m, which increments in each step.

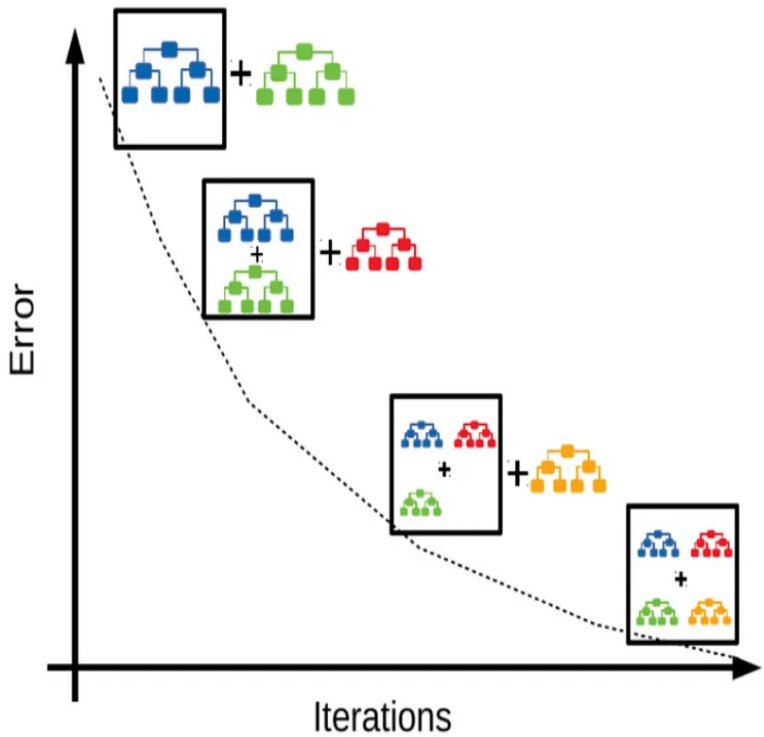


Figure 2.5: The workflow of Gradient Boosting algorithm. Reference: (Hemashreekilari, 2023)

Algorithm 3 Gradient Boosting. Reference: an adaptation of (Natekin; Knoll, 2013)

Input: $X, y, M, \texttt{LearningRate}$

$F_0(x) \leftarrow$ Initial model prediction (e.g., mean of y) **for** $m \leftarrow 1$ **to** M **do**

  Compute residuals: $r_i^{(m)} = y_i - F_{m-1}(x_i)$ for each i Fit a new model $h_m(x)$ to the residuals $r_i^{(m)}$ Update the model: $F_m(x) \leftarrow F_{m-1}(x) + \texttt{LearningRate} \cdot h_m(x)$

Output: $F_M(x)$ {Final model after M iterations}

2.2.2 Algorithms to predict numerical features

These algorithms can be used to predict numeric characteristics, such as grades or average grades (GPA or CR). In US, the GPA is a number between 0 and 4 because the letter grades (F, D, C, B and A) are converted to numbers. In Brazil, the CR is a number

between 0 and 10 because the grade system in Brazil is already numeric. Both GPA and CR are an average of grades, therefore, in the context of this research both are considered similar.

2.2.2.1 Linear Regression

Linear Regression is an algorithm to predict the numerical feature y based on the features x by a linear function $Y = X\beta + \epsilon$, where the slope of the hyperplane is X, and ϵ is the intercept (the value of y when $x = 0$). Figure 2.6 describes the Linear Regression algorithm formula.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

Figure 2.6: Linear regression formula. Reference: (PennState, 2018)

2.3 Main metrics of AI prediction algorithms

With the adoption of AI algorithms to serve in multiple sectors, the adoption of evaluation metrics is essential (Naidu *et al.*, 2023). There are metrics that are used in tests to evaluate the reliability of an AI model. The main metrics are accuracy, precision, recall, and F1-score for categorical features, such as graduated or dropped out, if the student passed or reproved, and the categorical discrete grade or group of grades the student achieved. The main metrics for numerical features, such as grades, are the r-value and RMSE.

2.3.1 Metrics for models that predict categorical features

These metrics are used in models that predict categorical features, usually Decision Tree, Random Forest, Neural Networks, and Naive Bayes. The following are the cases where the models predicted the categorical features correctly:

- True Positives (TP): Cases where the model predicted positive, and the actual outcome was positive. For instance: a student was predicted that would graduate and graduate factually.
- True Negatives (TN): Cases where the model predicted the negative outcome and the actual outcome was negative. For instance: a student who was predicted that would drop out and, factually, dropped out.

The following are the cases where the models predicted the categorical features wrongly:

- False Positives (FP): Cases where the model predicted positive but the actual outcome was negative. For instance: a student who was predicted that would graduate but factually dropped out.
- False Negatives (FN): Cases where the model predicted negative but the actual outcome was positive. For instance: a student who was predicted that would drop out but factually graduated.

Table 2.1 represents the theoretical confusion matrix. The following metrics are derived from these definitions and the table.

Table 2.1: Theoretical Confusion Matrix.

	Actual Positive	Actual Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

2.3.1.1 Accuracy

Accuracy is a measure of the model's success rate in relation to the total number of examples. It is calculated as the ratio between the number of correct predictions and the total number of examples. It is calculated by:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{TP} + \mathsf{TN} + \mathsf{FP} + \mathsf{FN}}$$

2.3.1.2 Precision

Precision measures the proportion of examples correctly classified in relation to the total number of classified examples in a certain class. It can be calculated for both TP and TN separately. It is calculated by:

$$\text{Precision (TP)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Precision (TN)} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

2.3.1.3 Recall

Recall measures the proportion of examples that were correctly classified in relation to the total number of predicted examples, including false positives or negatives. It can be calculated for both TP and TN separately. It is calculated by:

$$\text{Recall (TP)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Recall (TN)} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

2.3.1.4 F1 Score

The F1 score is a combined measure of precision and recall. It provides a balance between these two metrics by calculating the harmonic mean between them. It can be calculated for both TP and TN separately. It is calculated by:

$$\text{F1 score}(\text{TP}) = \frac{2 \times (\text{Precision(TP)} \times \text{Recall(TP)})}{\text{Precision(TP)} + \text{Recall(TP)}}$$

$$\text{F1 score}(\text{TN}) = \frac{2 \times (\text{Precision(TN)} \times \text{Recall(TN)})}{\text{Precision(TN)} + \text{Recall(TN)}}$$

2.3.2 Metrics for models that predict numerical features

These metrics are used to models that predict numerical features, usually Linear Regression and can be used to predict numeric grades and average grades.

2.3.2.1 R-Value

R-Value represents the correlation between the features the model use, known as x, to predict the target feature, known as y. It is calculated by:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where:

- n is the number of data points,
- $\sum xy$ is the sum of the product of each pair of corresponding x and y values,
- $\sum x$ and $\sum y$ are the sums of the x and y values, respectively,
- $\sum x^2$ and $\sum y^2$ are the sums of the squares of the x and y values, respectively.

The r-value can assume values between - 1 and 1. Values near 0 indicate no correlation between x and y, values near 1 indicates a positive correlation, and values near -1 indicate a negative correlation. Figure 2.7 represents an example of the possible values r can assume.

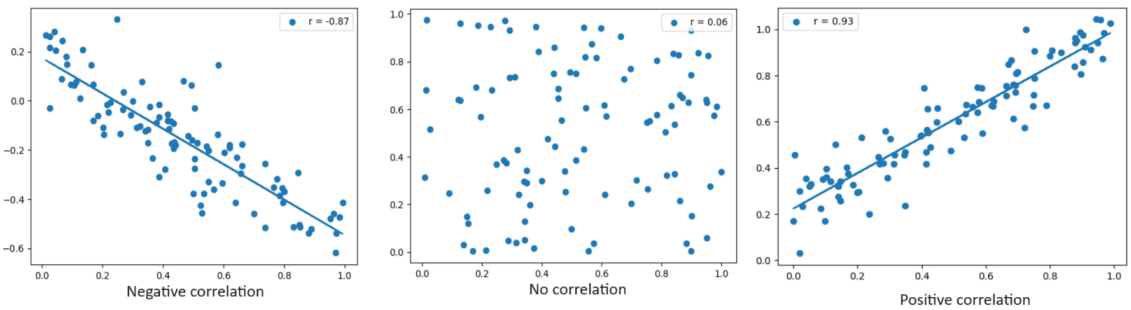


Figure 2.7: Possible correlations. Reference: the author

2.3.2.2 Root Mean Square Error (RMSE)

RMSE (Root Mean Square Error) is a measure of the difference between the values predicted by a model and the actual values. It is calculated by:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where:

- n is the number of observations,
- y_i is the actual value of the dependent variable for observation i,
- $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

2.4 Chi Square Test

The Chi Square test is a statistical test to verify if two categorical variables are independent or dependent on each other (Ugoni; Walker, 1995). To use the test, it is necessary to formulate two hypotheses:

- H_0 is the null hypothesis: there is not a significant statistical correlation between variables, therefore, the variables probably are independent.
- H_1 is the alternative hypothesis: there is a significant statistical correlation between variables, therefore, the variables probably are dependent on each other.

The Chi-square can be calculated by:

$$\chi^2 = \sum \frac{(\text{Observed values} - \text{Expected values})^2}{\text{Expected}}$$

The expected values in the formula are considering that the null hypothesis is true, leading to a more probable value of χ^2 if the values are random. If the values are not random, this indicates that there is a correlation between variables, leading to a more unlikely value of χ^2 considering the null hypothesis. This probability is called the p-value. If the p-value < 0.05, the null hypothesis is rejected.

In the context of this research, the Chi-Square Test is used to verify which factors have more statistical correlation with dropout.

3. Systematic Mapping Study

3.1 Systematic mapping study about dropout prediction

This chapter presents an SMS titled Artificial Intelligence Algorithms to Predict College Students' Dropout: a Systematic Mapping Study, which was published in Proceedings of the International Conference on Agents and Artificial Intelligence 2024 (ICAART) (Rodrigues *et al.*, 2024a).

3.1.1 Introduction

The objective of this study is to identify the most common algorithms used to predict student dropout, the features used by these algorithms, and the typical challenges in their implementation. To do so, we conducted a systematic mapping study (SMS) to identify and analyse the existing literature on experiments using AI algorithms to predict dropout in HEIs, contributing to an overview of this issue.

The remainder of this study is structured as follows: Subsection 3.1.2 details previous literature reviews on this topic; Subsection 3.1.3 presents the planning and conduction of this SMS; Subsection 3.1.4 details the results of this SMS; Subsection 3.1.7 discusses the findings of this SMS; Subsection 3.1.8 explores the threats to validity of the SMS; and Subsection 3.1.9 presents final remarks and future work.

3.1.2 Related Work

(Tete *et al.*, 2022) conducted a systematic literature review to analyse studies related to prediction models for student dropout from HEIs. The authors found that the most common algorithm is the Decision Tree. The most important features were grouped into five categories: socioeconomic (gender, age, professional position, income, ethnic group), academic (grades, Grade Point Average - GPA, frequency), psychological (learning difficulties, academic life satisfaction, sociability), health (well-being, diseases, health issues), and accessibility. This study did not identify any academic projects or actions to decrease student dropout.

(Silva; Roman, 2021) also conducted a systematic literature review. They found that the most analyzed features in the studies relate to sociodemographic and academic factors,

as well as psychological and motivational variables. They also concluded that the most frequently used algorithms are Naive Bayes, KNN, and Random Forest (a combination of several decision trees).

This study, as in previous reviews, also investigates the most common algorithms and features used to predict student dropout. Our contribution is to also investigate the accuracies reached by the algorithms and the most common limitations and difficulties faced in the implementation of such algorithms, besides validating the previous results found in the literature.

3.1.3 Research Method

We performed an SMS based on Kitchenham and Charters (Kitchenham, 2012) and Petersen et al. (Petersen *et al.*, 2015) guidelines, which prescribe the following phases: establish research scope, execute search, select studies, extract data, and perform analysis. The study was documented via Parsifal[1], an online tool to support SMS and it is detailed in the following subsections.

3.1.3.1 Search Strategy and Data Source

The research question that expresses the goal of this study was formulated following the criteria specified in the PIO (population, intervention, and outcome), as shown in Table 3.1. Therefore, the formulated research question (RQ) is *"How are the artificial intelligence algorithms used to predict dropout rates among higher education students?"*.

Table 3.1: PIO structure to formulate the research question.

PIO	
Population	Higher Education Institutions Dropout
Intervention	Artificial Intelligence Algorithms
Outcome	Algorithms, Difficulties, Accuracies, and Features

The desirable outcome of this research is to understand which AI algorithms are most commonly used, which variables are used by these algorithms, how well these algorithms can predict student dropout in terms of accuracy, and the most common difficulties and limitations on the implementation of such algorithms. Moreover, to

[1]https://parsif.al/

expand the comprehension of the research question, the following sub-questions (Sub-Q) were formulated:

(Sub-Q1): What are the difficulties in using AI to predict university dropout rates?

(Sub-Q2): How do the AI algorithms use features to predict university dropout rates?

The sources to search by the existing studies were: ACM Digital Library, IEEE Xplore, and Scopus.

3.1.3.2 Search String

A generic search string was created from the keywords and their synonyms. Keywords were connected using the AND logical operator, whereas variations and synonyms were connected using the OR operator. The terms of the search string were selected to conduct a broader search that included a wide range of studies. We tested different configurations of the search string in Scopus. After calibrating the search string, the final version was:

(“higher education" OR “college" OR “graduation" OR “university") AND (“predict*") AND (“artificial intelligence" OR “AI" OR “data science" OR “deep learning" OR “machine learning") AND (“drop off" OR “drop out" OR “dropout")

3.1.3.3 Selection Criteria

To properly address the research question and its subquestions, we established selection criteria to include studies relevant to the topic and exclude those that are not. In this study, publication year was not deemed a relevant criterion. The adopted selection criteria are shown in Table 3.2. No criteria were set regarding the publication date, and studies from any country were considered acceptable.

3.1.3.4 Study Selection Process

After retrieving studies from the sources, the following filters were used to select the studies: I) title, abstract, and keywords screening; II) introduction and conclusion screening; and III) full text screening.

3.1.3.5 Data Extraction

We extracted the following data for each of the accepted studies: Study ID, reference, algorithm(s) used, features used, algorithm accuracy, and study limitations. The extracted

Table 3.2: Selection Criteria.

Inclusion Criteria	
IC1	Study describes an AI technique for predicting dropouts in higher education.
Exclusion Criteria	
EC1	Study describes an AI technique for predicting dropouts in elementary, high school, or massive open online courses (MOOCs).
EC2	Duplicate study.
EC3	Study is not available for reading and data collection (files paid for or not made available by search engines).
EC4	Study is not peer-reviewed.
EC5	Secondary study.
EC6	Study is not written in English.
EC7	Study is not within the topic of AI techniques to predict higher education dropout.

data were saved in a spreadsheet form and later used to support the discussion of the SMS results.

3.1.4 Results

In this subsection, we present the survey's main findings.

3.1.4.1 Sources of Studies

The number of studies retrieved from each source is described in Table 3.3. From the search in the chosen sources, 223 studies were retrieved: 31 were retrieved by IEEE Xplore, 3 by ACM Digital Library, and 189 by Scopus.

After applying the inclusion and exclusion criteria and filtering, 23 studies were selected, as shown in Tables 3.4 and 3.5. Not all the features used in the studies are displayed in the tables. When multiple algorithms were used in a study, the algorithm with the highest accuracy was selected.

Table 3.3: Number of studies by source.

Quantity of studies by source	
IEEE Xplore	31
ACM Digital Library	3
Scopus	189

3.1.4.2 Filtering

The filtering process is described in Figure 3.1.

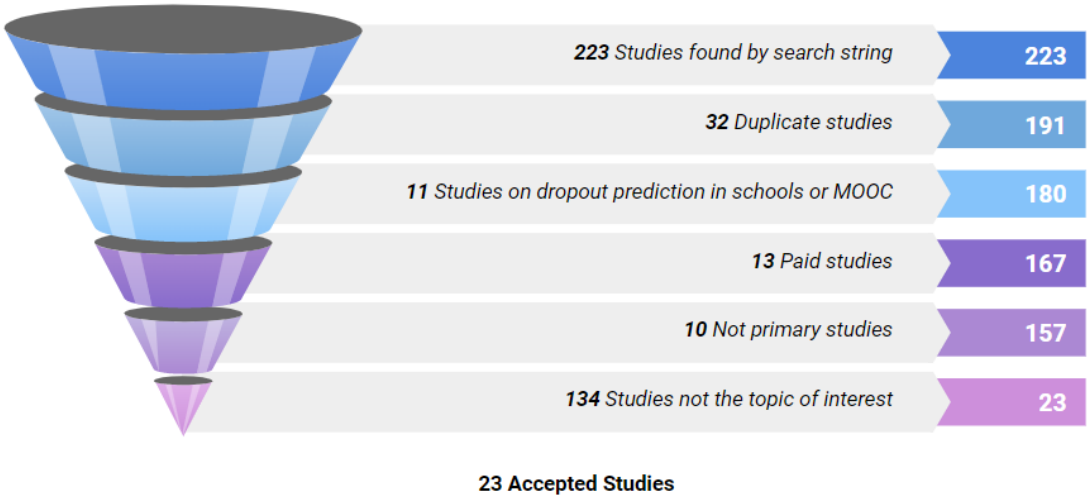


Figure 3.1: Studies' filtering process

From the set of studies retrieved, 11 were excluded because the dropout prediction focused on basic education or MOOCs. Thirteen studies were excluded due to paid access. Moreover, some studies were about AI algorithms to predict academic performance, not focusing on dropout risk.

3.1.4.3 Country of Origin of the Studies

The selected studies analysed dropout behaviour in HEIs of different countries. Figure 3.2 describes the number of studies conducted in each country.

We identified one study from Portugal, Hungary, Vietnam, the United Kingdom of Great Britain and Northern Ireland, Costa Rica, the United States of America, Brazil, Malaysia, Saudi Arabia, Italy, and India. We identified 2 studies each from Colombia, Chile, and Spain. Finally, we identified 4 studies from Peru, the country with the most studies identified in this SMS. From the set of 23 selected studies, 10 were from Latin America, 7 were from Europe, 5 were from Asia, and one study was from North America.

3.1.4.4 Algorithms in the Studies

The percentage of each algorithm found in each study is described in Figure 3.3.

In the case of studies that compared a set of algorithms, we only considered in Figure 3.3 the algorithm with the highest accuracy; therefore, it does not represent the

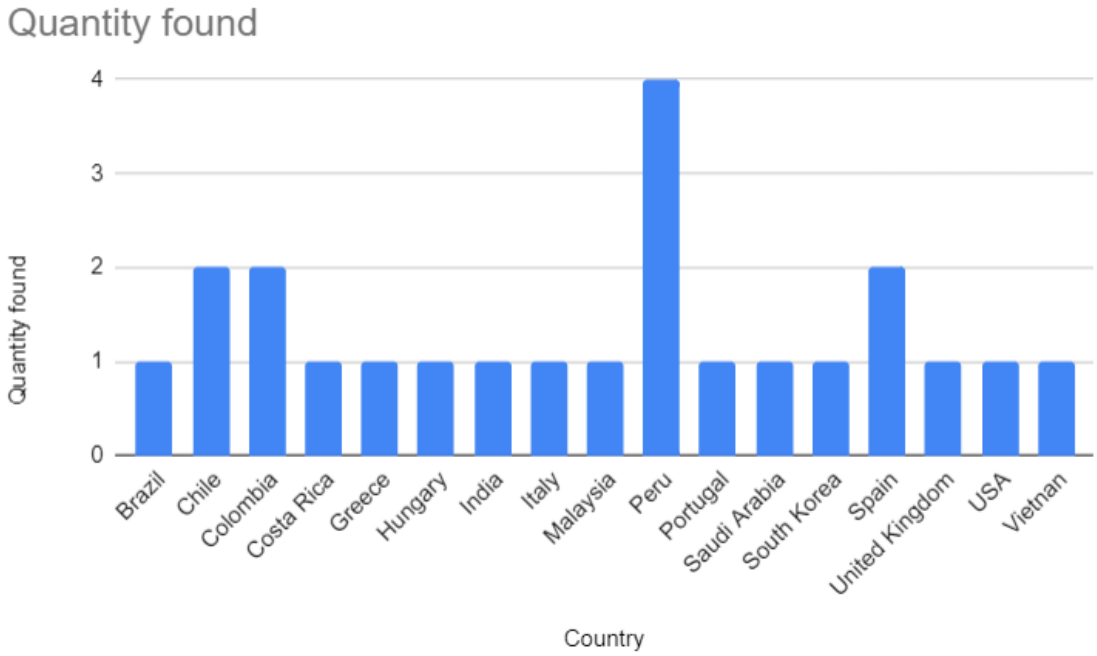


Figure 3.2: Studies' countries

total percentage for each algorithm used in the studies. In other words, it represents only the algorithm with the highest accuracy of each selected study. After performing data extraction from the selected studies, it was possible to answer the sub-questions, presented in the next subsection.

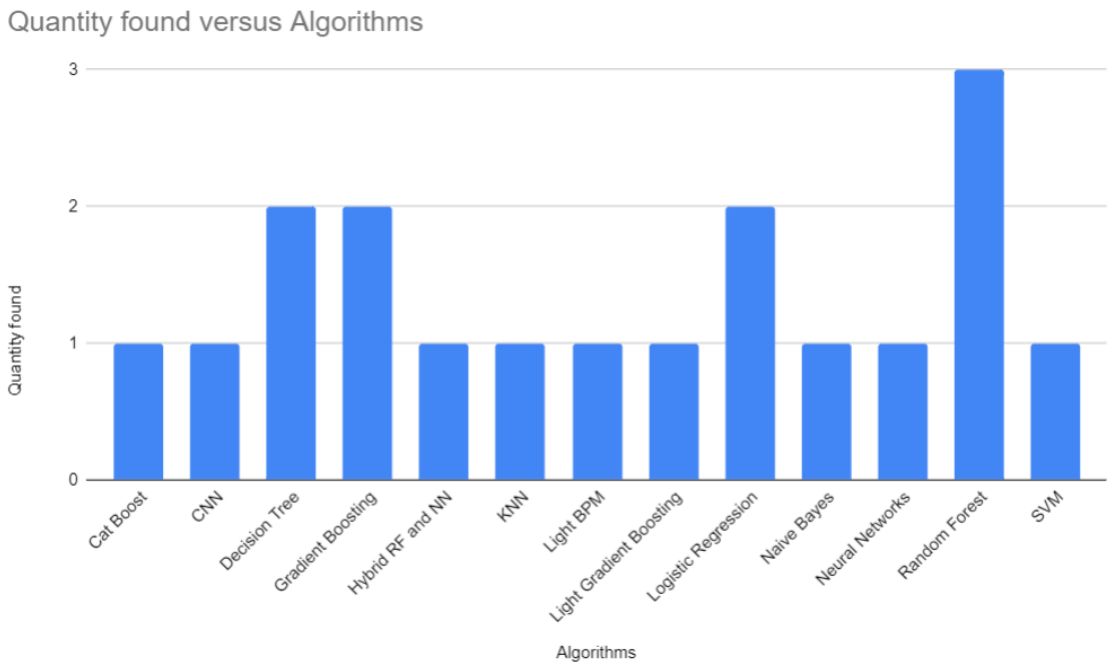


Figure 3.3: Algorithms explored in the selected studies

Table 3.4: Results extracted from the studies

ID	Reference	Algorithm	Main Variables	Best Accuracy	Limitations
S1	(Realinho *et al.*, 2022)	Random Forest	Marital status, parent's formation	N/A	Bias may occur
S2	(Nagy; Molontay, 2023)	Cat Boost	Hungarian entrance exam scores, course, gender, age	0.84	Limited to Budapest
S3	(Osorio; Santacoloma, 2023)	Logistic Regression	Depression, drug addictions	0.80	Not mentioned
S4	(Anh *et al.*, 2023)	Light Gradient Boosting	Grades in subjects, attendance in classes	0.95	Bias may occur
S5	(López-Angulo *et al.*, 2023)	Structural Equation Modelling	Satisfaction with HEI	N/A	Satisfaction with academic life may change in time
S6	(Jimenez-Macias *et al.*, 2022)	Random Forest	Grades, Employment, credits	0.99	Few data
S7	(Gutierrez-Pachas *et al.*, 2023)	CNN	Grades, GPA, HDI	0.98	Unequal behaviours
S8	(Zihan *et al.*, 2023)	Light BPM	Grades, GPA	0.93	Not mentioned
S9	(Kotsiantis *et al.*, 2003)	Naive Bayes	Occupation, grades, attendance on tutoring	0.83	Not mentioned
S10	(Moseley; Mead, 2008)	Decision Tree	Grades, age, gender	0.94	Few data
S11	(Solis *et al.*, 2018)	Random Forest	Average of Grades, academic records	0.91	Few data
S12	(Zhang; Rangwala, 2018)	Iterative Logistic Regression	Scores of SAT and ACT	0.98	New proposed algorithm

Table 3.5: Results extracted from the studies

ID	Reference	Algorithm	Main Variables	Best Accuracy	Limitations
S13	(Pachas *et al.*, 2021)	Random Forest	Quantity of fails	0.78	Lack of data diversity
S14	(Gismondi; Huiman, 2021)	Neural Networks	Grades, use of mobiles	0.87	Not mentioned
S15	(Fernández-García *et al.*, 2021)	Gradient Boosting	Not mentioned	0.72	Privacy issues
S16	(Santos *et al.*, 2020)	Decision Tree	GPA, Entrance exam scores	0.95	Unbalanced classes
S17	(Sani *et al.*, 2020)	Gradient Boosting	Academic year, high-school GPA, channels of admission	0.93	Not mentioned
S18	(Uliyan *et al.*, 2021)	Neural Networks	Grades, GPA	0.90	Not mentioned
S19	(Agrusti *et al.*, 2020)	CNN	Not mentioned	0.94	Data accuracy required.
S20	(Opazo *et al.*, 2021)	Gradient Boosting	Grades, GPA	0.69	Different HEIs may need different methods
S21	(Ramirez *et al.*, 2022)	Random Forests	Grades, age, gender, academic credits	0.99	Not mentioned
S22	(Daza *et al.*, 2022)	Hybrid Random Forest and Neural Networks	Gender, Age, Academic Credits	0.99	New proposed algorithm
S23	(Revathy *et al.*, 2022)	K-nearest neighbours	Not mentioned	0.97	Not mentioned

3.1.5 (Sub-Q1): What are the difficulties in using AI to predict university dropout rates?

The most common difficulties and limitations are related to data availability and its small volume, they are usually data from the authors' HEIs affiliation. This limitation causes biases in the analyses. Another limitation is that data may differ in time, courses, and different HEIs, such as the behaviour of dropout rates and student satisfaction with academic life.

By acknowledging and actively working to overcome these challenges, higher education institutions can harness the potential of AI algorithms to make significant strides in supporting student success and retention. Collaboration among institutions and researchers can facilitate the sharing of knowledge and resources, thus creating more robust, unbiased, and adaptable predictive models.

Two studies proposed two new algorithms to predict student dropout (S12 and S22). Both studies claimed very high accuracy for their algorithms, which should be replicated in other datasets to confirm such results.

3.1.6 (Sub-Q2): How do AI algorithms use features to predict higher institutions' dropout?

We found that the most commonly used variables to predict HEI student dropout can be grouped into socioeconomic (gender, age, professional position, income, ethnic group), academic (grades, GPA, frequency, scores at entrance exams, quantity of failed disciplines), and psychological (satisfaction with the academic life, sociability). The majority (11) of the analyzed studies used academic and socioeconomic variables, only a few used (2) psychological variables, and none used physical health and accessibility-related variables. Thus, we could not verify (Tete *et al.*, 2022) results.

The analysed studies on this SMS did not explore major differences between gender, ethnicity, and age group on the behaviour of dropout prediction. However, it does not refute the existence of differences between these social groups.

The most important factors related to college students' dropout are academic performance, such as grades, GPA, attendance in class, and credits taken. The most important external factors are the psychological state of the student, such as satisfaction with academic life and addiction to drugs. Finally, the most used variables in AI algorithms to predict student dropout are related to academic performance.

3.1.7 Discussion

Several researchers around the globe are investigating AI algorithms to predict student dropout, testing algorithms, such as Random Forest, Cat Boost, Logistic Regression, Neural Networks, Decision Tree, Naive Bayes, KNN, Gradient Boosting, CNN, Light Gradient Boosting, Light BPM, and SVM. Some selected studies in this SMS tested more than one algorithm. In such cases, this study reported the algorithm with higher accuracy. The Random Forest algorithm is the most frequent algorithm with better performance. Additionally, the difficulties reported are mostly related to the unavailability of large data sources because most of the analysed studies used data provided by the authors' affiliated HEIs.

To develop a more reliable AI algorithm to predict student dropout, it is necessary to retrieve anonymized data from several HEIs in a large data source. However, it is a hard task to execute since different HEIs have different data formats, such as grades that can be expressed on a scale of 0 to 10, on a scale from F to A, or another format and variables, by different legislations, such as the General Data Protection Regulation (GDPR) from the European Union (European Commission, 2016) or the General Law on Data Protection (LGPD) from Brazil (Brasil, 2018).

Collaborative efforts among HEIs, researchers, and regulatory bodies are essential to overcome these challenges. Establishing data-sharing agreements that adhere to legal requirements while facilitating the exchange of anonymized data for research purposes can help unlock the potential for more reliable AI algorithms. Furthermore, initiatives to create standardized data formats and encourage transparency in data collection practices can contribute to the development of a more cohesive and effective research ecosystem focused on predicting student dropout.

The majority of the analyzed algorithms used data related to academic performance, such as grades and GPA, to predict student dropout, or concluded that such categories of features are the most significant for making such predictions. However, it was not explored how grades are influenced by another variable. In future work, it will be possible to investigate how AI algorithms predict academic performance, such as based on grades.

Another aspect to be explored is the influence of non-academic features on academic performance. These could include socioeconomic factors, such as family background, financial stability, and access to support services. Additionally, personal factors such as motivation, study habits, and mental health can significantly impact a student's grades. Investigating how these variables interact with academic performance can help create a

more comprehensive understanding of the factors contributing to student dropout risk.

Moreover, a subject of interest could be the temporal aspect of academic performance prediction. Analysing how students' grades evolve and how early warning signs in academic performance can be identified can be crucial for proactive interventions to prevent dropout. Furthermore, the application of advanced AI techniques, such as machine learning interpretability methods, could help shed light on how certain features or variables contribute to academic performance predictions. This can provide valuable insights into the underlying mechanisms that drive the results of AI models.

3.1.8 Threats to validity

The main threats to this SMS are related to the strategies adopted to create the search string, retrieve primary studies, and extract data from these primary studies. The completeness of this SMS may have been affected by the missing relevant primary studies because some of them may not be retrieved by the search string, or because some of them were excluded by EC3 due to paid access. The authors are aware that considering only peer-reviewed studies on the topic of using AI algorithms for predicting HEI student dropout does not allow for the generalization of the results, as there may be relevant content on this topic in grey literature, such as technical reports.

In addition, the quality of this SMS may also be influenced by potential biases introduced during the selection and inclusion of primary studies. The criteria used to determine which studies to include and exclude could inadvertently introduce bias, affecting the overall comprehensiveness and representativeness of the findings.

3.1.9 Conclusion

We performed an SMS in which 23 studies were selected for analysis. The results reveal that several HEIs around the globe are testing algorithms to predict student dropout, trying to find the most significant features, sharing their limitations, and trying to maximize the algorithms' accuracy.

From the results, we conclude that there is no specific recommended algorithm to predict higher education students' dropouts. Many studies test different algorithms to perform this task, looking for the one with the highest accuracy. In our search, the Random Forest algorithm was the one that had a better performance in most of the studies. The most recommended features are related to academic performance, such as

grades, GPA, credits taken, and attendance at class. Psychological health features, such as satisfaction with academic life, drug addiction, and mental diseases are also present but are less used. The most common difficulties in implementing these AI algorithms are related to the unavailability of a large quantity of data to be used and the diversity of realities in which different HEIs and undergraduate courses are inserted.

Based on this study, we hope to contribute to the field by providing the current overview of the AI algorithms used in predicting HEI students' dropout.

4. Method

4.1 Ethical Assessment

The research was submitted for ethical review to the Brazilian Platform Ethics Committee, where it received a favourable assessment. The Certificate of Presentation of Ethical Review (CAAE) number for the ethical appraisal and approval process is 78896924.5.0000.5285. This indicates that the study has been evaluated and deemed to adhere to the ethical standards and guidelines required for research involving data in Brazil This was necessary to use the Brazilian Identification Number (CPF), which identifies a Brazilian person, in data crossings.

4.2 Cross Industry Standard Process for Data Mining

To address the RQ and the Sub-RQs presented in Chapter 1, the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology (Provost; Fawcett, 2016), EDM techniques, and the construction of AI models with Gradient Boosting (Friedman, 2000) were used.

The CRISP-DM was used in these phases:

- Data understanding: It was research on literature which data and models are most frequently used in the topic of AI prediction of dropout. It was observed that financial data was poorly used, therefore (Tete *et al.*, 2022; Rodrigues *et al.*, 2024a), this kind of data was chosen to be researched in this study.
- Data preparation: Three data sources were crossed to obtain students who became company owners, were employed during graduation, or received scholarships.
- Modeling: It was conducted an exploratory data analysis, and chi-square tests and it was trained Gradient Boosting models.
- Models Evaluation: the Gradient Boosting models were evaluated.
- Deployment: It was built a web system using these models

Figure 4.1 represents a general scheme of the CRISP-DM method, showing each step.

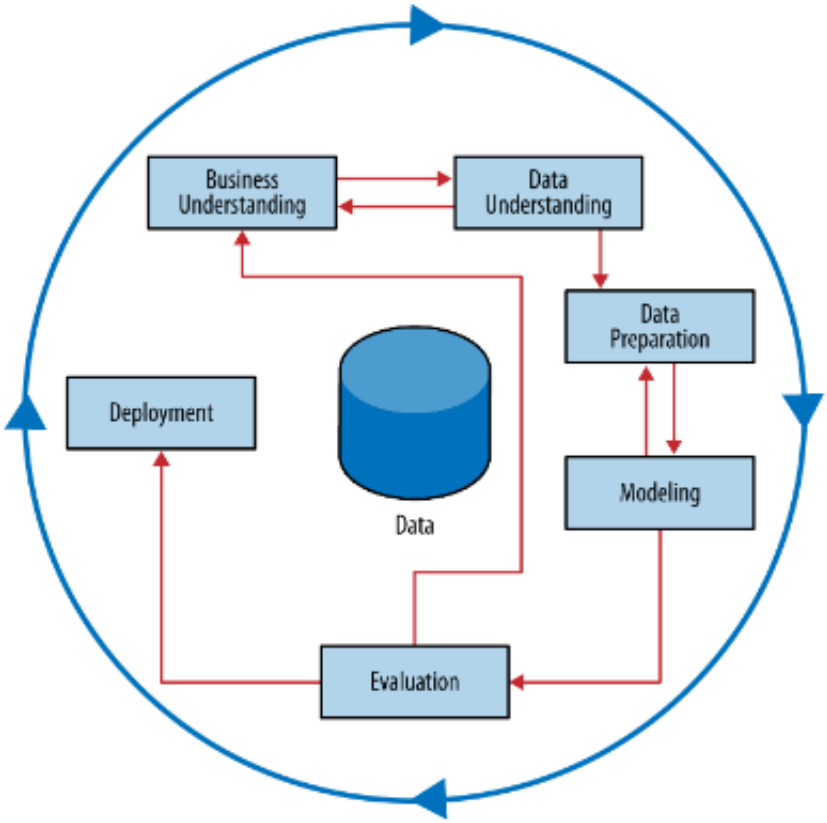


Figure 4.1: Cross Industry Standard Process for Data Mining. Reference: (Provost; Fawcett, 2016)

4.3 Method of the two following chapters

This section details the method adopted on the following chapters.

4.3.1 Chapter 5: exploratory analysis and preliminary study

In this chapter, it was made an exploratory analysis and Chi-Square Tests on the data, which are described in the Subsubsection 4.3.1.1. It also trained a preliminary decision tree model using only academic data from the Information System course 2000.1 to 2023.2, consisting of 853 unique students. This model was divided into 80% for training and 20% for tests.

4.3.1.1 Chi-Square Tests' Hypothesis

To determine if there is a statistically significant relationship between the variables, the chi-square test (Ugoni; Walker, 1995) was applied, with the following hypotheses:

- H_0 for **sub-RQ4**: there is no statistically significant correlation between dropout

and full-time employment.

- H_1 for **sub-RQ4**: there is a statistically significant correlation between dropout and full-time employment.

- H_0 for **sub-RQ5**: there is no statistically significant correlation between dropout and entrepreneurship.
- H_1 for **sub-RQ5**: there is a statistically significant correlation between dropout and entrepreneurship.

- H_0 for **sub-RQ6**: there is no statistically significant correlation between dropout and receiving scholarships.
- H_1 for **sub-RQ6**: there is a statistically significant correlation between dropout and receiving scholarships.

4.3.2 Chapter 6: follow-up and final study

It was trained the final Gradient Boosting models for all three STEM courses of UNIRIO, located at the CCET: Information Systems, Production Engineering, and Mathematics, and a general model for the whole CCET. The data was split 80% for training and 20% for tests.

5. Data Understanding and Exploratory Analysis

5.1 Database

For this research, we used Educational Data Mining (EDM) techniques (Baker et al. 2011), and a dataset that was built using data from only the Information System from 2000.1 to 2023.1 and data from the whole CCET at UNIRIO from 2013.1 to 2023.1; open data from the Brazilian Federal Revenue, which contains data of company owners from November 2021 to May 2024 [1] and data from Annual Relation of Social Information (RAIS) from 2013 to 2023, which was obtained from the Brazilian Ministry of Labour by the process 19964.212740/2024-11. Figure 5.1 summarizes the features collected from the SIE, CNPJ, and RAIS used in this study. After balancing the dataset based on feature engineering methods, such as removing rows with missing data and irrelevant outliers for the study (i.e., grades above the allowed average), the database contains 83,875 rows concerning the students' grades in each curricular activity, i.e., the same student appears several times in the database.

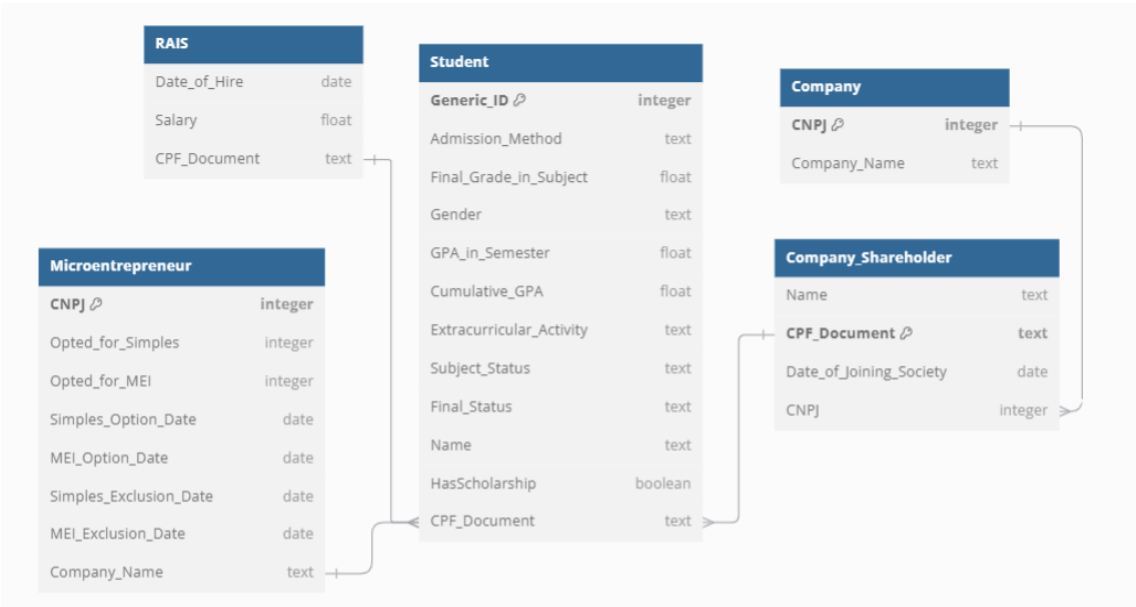


Figure 5.1: The database

In total, there are 974 distinct students in which 76% are males and 24% are females. Students who were still enrolled in the course, those who passed away during the course, and those who were transferred to other courses or institutions were not included in the

[1] Federal Revenue open data: https://basedosdados.org/dataset/e43f0d5b-43cf-4bfb-8d90-c38a4e0d7c4f?table=3dbb38d1-65af-44a3-b43a-7b088891ebc0

database. All rows in the dataset were filled, and there were no problems with missing data. Table 5.1 represents the dataset after the data crossings. Tables 5.2 represent the disciplines with more failures in Information Systems, 5.3 represents the disciplines with more failures in Production Engineering, and 5.4 represents the disciplines with more failures in Mathematics. The models were created using Python on Google Colab[2].

Table 5.1: Database features

Feature	**Description**
Generic Student ID	An ID to identify rows of the same student
Course	The student's undergraduate program
Year of enrolment in curricular activity	The year when the student enrolled in the curricular activity
Semester of enrolment in curricular activity	The first semester when the student enrolled in the course
Admission method to the course	How the student was admitted to the course (i.e., if the student was admitted by quota or broad competition)
Final grade in curricular activity	The final grade the student received in the curricular activity
Gender of student	The student's gender
GPA in the semester	The student's GPA for the semesters
Accumulated GPA of the former student	Student's accumulated GPA
Curricular activity name	The name given to each curricular activity of the course.
Status of curricular activity	Whether the student approved or failed the curricular activity
IsTheyBusinessperson	Whether the student is a company owner or not
Category of businessperson student	Whether the student is not a company owner or founded a company during graduation, after graduation or before enrolled HEI.
IsTheyEmployeeStudent	Whether the student is an employee or not
Has scholarship	Whether student had scholarship or not
Final Status of the student (i.e., course completed or not as this refers to what is predicted) **(target class)**	Whether the student completed the course or not

[2]https://colab.google/

Table 5.2: Database features: Information Systems Curricular Activities

Semester	Feature	Description
1st	grade_programming1	Grade of Programming 1 discipline
2nd	grade_programming2	Grade of Programming 2 discipline
1st	grade_basic_math	Grade of Basic Maths discipline
2nd	grade_calculus1	Grade of Calculus 1 discipline
2nd	grade_linear_algebra	Grade of Linear Algebra discipline
2nd	grade_logic	Grade of Logic discipline

Table 5.3: Database features: Production Engineering Curricular Activities

Semester	Feature	Description
1st	grade_programming1	Grade of Programming 1 discipline
1st	grade_calculus0	Grade of Calculus 0 discipline
2nd	grade_calculus1	Grade of Calculus 1 discipline
1st	grade_engineering_introduction	Grade of Engineering Introduction discipline
2nd	grade_linear_algebra	Grade of Linear Algebra discipline

Table 5.4: Database features: Mathematics Curricular Activities

Semester	Feature	Description
1st	grade_programming1	Grade of Programming 1 discipline
1st	grade_environment	Grade of Environment discipline
2nd	grade_geometry1	Grade of Geometry 1 discipline
2nd	grade_calculus1	Grade of Calculus 1 discipline
1st	grade_math_foundation	Grade of Maths Foundation discipline
1st	grade_analytic_geometry	Grade of Analytic Geometry discipline

5.2 Dropout rate formula

In the dataset, there are several possible values for the final status of the student. In this research, it was excluded the final status of currently enrolled students, transferred to another HEI students and students that unfortunately died during the graduation.

It was considered students who concluded successfully and students who dropped out, which includes abandonment of the course, withdrawal from the course, general cancellation of the course and dismissal.

Therefore, considering that Dropped Students + Concluded Students = Total Students, the mathematical formula used to calculate dropout rates is:

$$\text{Dropout Rate} = \frac{\text{Dropped Students}}{\text{Concluded Students} + \text{Dropped Students}}$$

5.3 Data analysis in Information System course using only academic data

This section presents the results of the paper titled "Predicting Student Dropout on the Information Systems Undergraduate Program of UNIRIO Using Decision Tree", which was published on the Workshop of Education in Computing at the Congress of Brazilian Society of Computing (WEI/CSBC) (Rodrigues *et al.*, 2024b).

From 853 distinct former students of Information Systems from 2000.0 to 2022.1, 432 (50.64%) graduated and 421 (49.36%) dropped out of the course. Of the total, 681 (79.84%) are male and 172 (20.16%) are female. Table 5.5 shows the specified number and percentage of each gender by who graduated and dropped out. The chi-square statistic was calculated, obtaining a p-value of 0.129. The null hypothesis was that the gender and the student's outcome (graduation or dropout) were independent. Since we did not reject the null hypothesis, we do not have sufficient evidence to state that there is an association between gender and outcome.

Table 5.5: Graduation or dropout by gender

Graduation or dropout by gender		
Total males	**Graduated males (%)**	**Dropped out males (%)**
681	49.34%	50.66%
Total females	**Graduated females (%)**	**Dropped out females (%)**
172	55.81%	44.19%

The Unified Selection System (SiSU) is a national university entrance exam and was adopted at UNIRIO in 2013 by a Brazilian resolution (Brasil 2012). The quotas were nationally adopted in federal universities with the creation of SiSU (Heringer, 2024). adopted Taking into account only the 234 students who entered after the SiSU was implemented, we performed a chi-square test, excluding students who were admitted before the adoption of SiSU, to test the statistical independence between the admission method and dropout, obtaining a p-value of 0.003. The null hypothesis, which was

rejected, was that the admission method and student outcome (graduation or dropping out) were independent, suggesting an association between them. Table 5.6 shows the specified number and percentage of students by admission method of who graduated and dropped out.

Table 5.6: Graduation or dropout by admission method

Graduation or dropout by admission method		
Number of students admitted before SiSU	**Graduated students admitted before SiSU**	**Dropped out students admitted before SiSU**
619	46.05%	53.95%
Number of students admitted by SiSU quotas	**Graduated students admitted by SiSU quotas**	**Dropped out students admitted by SiSU quotas**
95	30.52%	69.48%
Number of students admitted by SiSU (non-quotas)	**Graduated students admitted by SiSU (non-quotas)**	**Dropped out students admitted by SiSU (non-quotas)**
139	49.64%	50.36%

Figures 5.2 to 5.4 show the accumulated GPA, semester GPA, and curricular activity grade by who graduated and dropped out, respectively. Visual inspection of the box plot allows us to infer statistical significance between the accumulated and semester GPAs of those who graduated and dropped out. However, it cannot be extended to the grades of curricular activities. In this case, dropouts have a larger interquartile range than graduates, comprising the whole grade spectrum. A reason for this behaviour could be a difference in the difficulty of curricular activities, which means that some curricular activities may be responsible for "holding back" some students, leading to dropout. We will further investigate this phenomenon in Section 5.3.3.

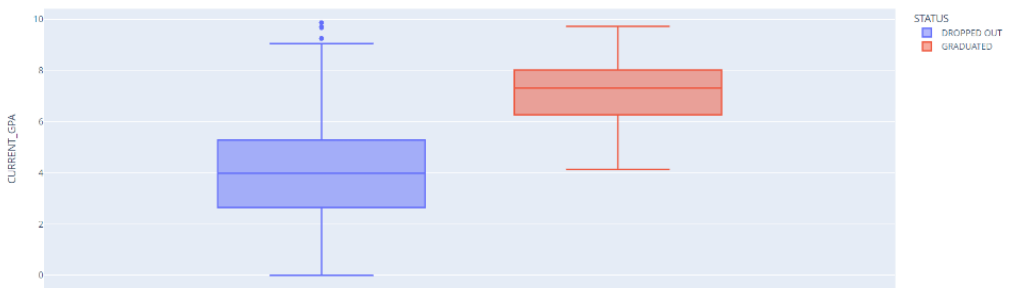


Figure 5.2: Accumulated GPA by outcome

Figure 5.5 presents the difference in the average GPA per semester between former students who graduated and those who dropped out. We can notice that after 16

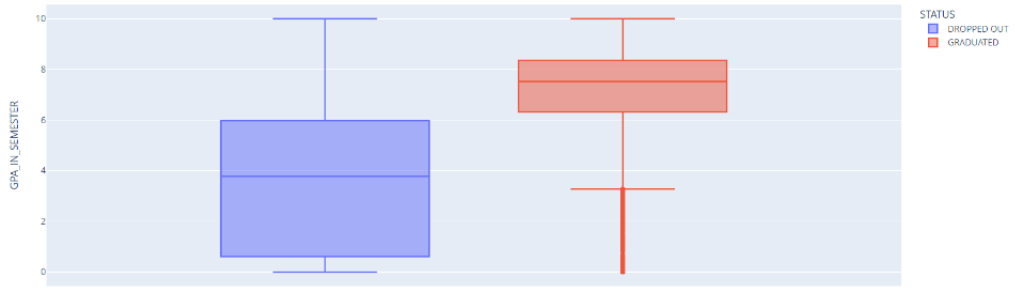


Figure 5.3: Semester GPA by outcome

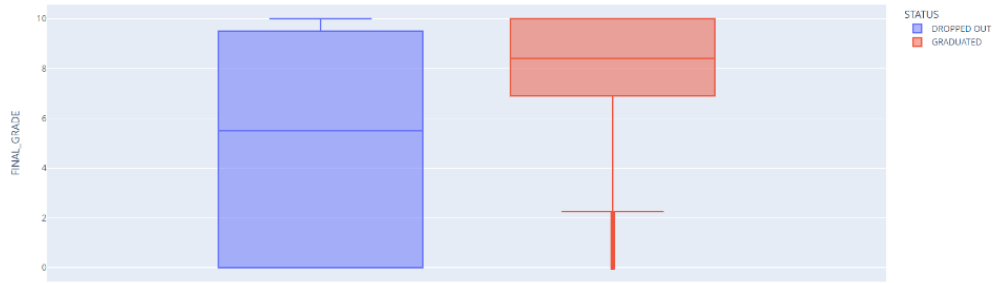


Figure 5.4: Curricular activities grades by outcome

semesters (eight years - the regular course is completed in four years), all dropped out students have already left the university. Those who continue after this period are likely to graduate, taking up to 21 semesters (11 years) to get the degree. In turn, Figure 5.6 shows the correlation between academic performance features and the outcome status (graduation or dropout). We conducted a Mann-Whitney U Test to verify if graduated and dropped-out students perform similar results in CURRENT_GPA, GPA_IN_SEMESTER, and FINAL_GRADE. All tests resulted in a p-value < 0.001, which means that the null hypothesis (two groups of students have the same academic performance) can be rejected.

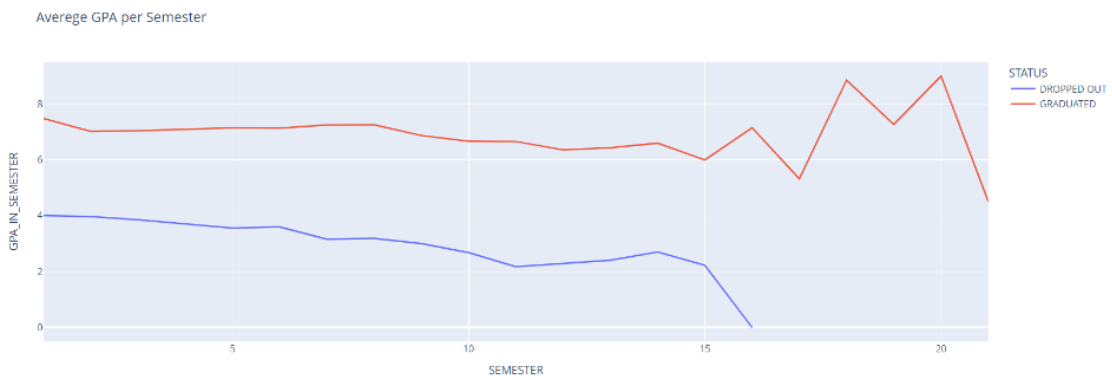


Figure 5.5: Average semester GPA by outcome

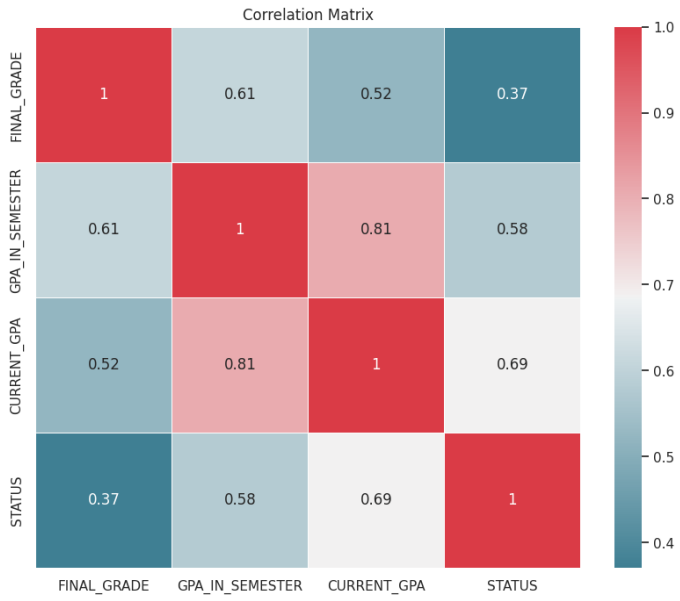


Figure 5.6: Correlation: academic performance X status (graduation/dropout)

5.3.1 Model: Decision Tree

We created a model focused on the student that left the undergraduate program, either successfully or not, using the unique 853 rows and the following features: admission method to the course, accumulated GPA, and gender. These features were used to develop the model because features related to academic performance and socioeconomic conditions were the most common to make this kind of model, according to the literature reviews on this topic (Silva and Roman 2021, Tete et al. 2022, Rodrigues et al. 2024). Figure 5.7 presents the result of the decision tree made by the model. The Gini criterion was used and the minimum impurity was defined as 0.005.

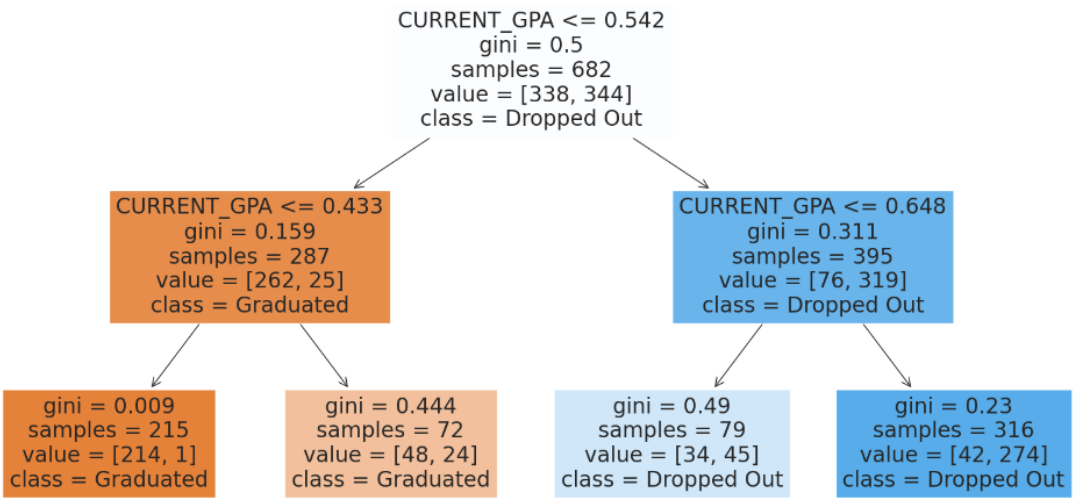


Figure 5.7: Decision tree of the model focused on the former student

As expected from the exploratory data analysis, the model considered the accumulated GPA (in the features named CURRENT_GPA) as the main factor to predict the status of graduation or dropout. The chi-square test suggested an association between the admission method and the predicted outcome. However, this feature and gender were not selected by the model as important features. The model had an accuracy of 83.04%. Table 5.7 presents the model's results metrics. The model predicted that 83 students graduated, 63 as true graduation, and 20 as false graduation. It also predicted that 88 students dropped out, 79 as true dropouts and 9 as false dropouts.

Table 5.7: Model accuracy and classification report

Model accuracy and classification report				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.88	0.76	0.81	83
Graduated students	0.80	0.90	0.84	88
		accuracy	0.83	

5.3.2 (Sub-RQ1): What are the most determining variables to predict dropout in BSI at UNIRIO?

After the data analysis and the prediction made by the model, it was demonstrated that the most determining variables to predict university dropout in BSI at UNIRIO is the Accumulated GPA.

5.3.3 (Sub-RQ2): In which years was there the highest dropout rate in BSI at UNIRIO?

Figure 5.8 shows the years until a former student dropped out, concretely. Most dropouts occur in the second year of the course, referred to as 1 in the graph since it was counted beginning at 0, which comprehends the third and fourth semesters. The fourth year is supposed to be the last year of the course if a student graduates by the established deadline by UNIRIO, but such a stage concentrates the second-highest dropout occurrences. The third-highest dropouts occur in the first year of the course. Therefore, the need to create early-warning dropout models, focusing on predicting student dropout during their first two years.

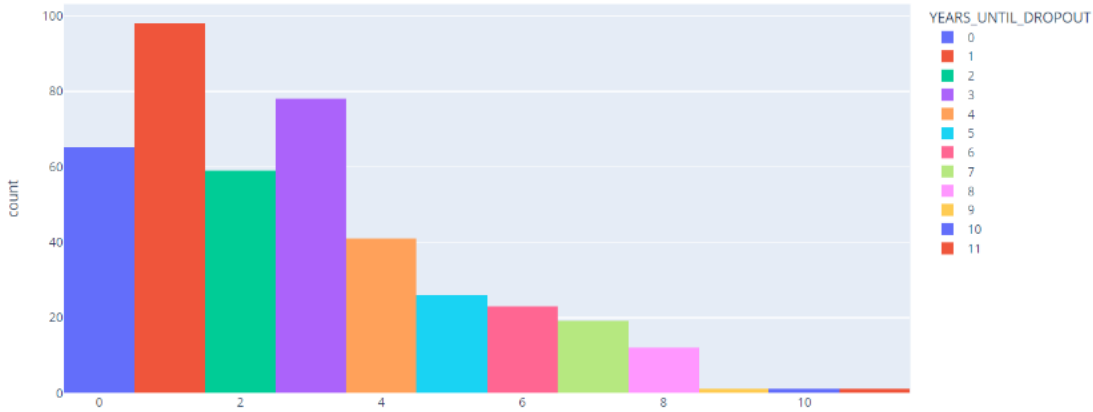


Figure 5.8: Dropout of students per years since the enrollment

5.3.4 (Sub-RQ3): Which curricular activities are the most decisive for dropout in BSI at UNIRIO?

Figure 5.9 shows the top 10 curricular activities that incur the most failures. Academic performance variables are key predictors of university dropout in BSI at UNIRIO. Notably, Accumulated GPA and semester GPA are significant. These GPAs are derived from final grades. Students who fail curricular activities receive grades below 5. Consequently, it can be inferred that the curricular activities with the highest failure rates are the most influential in determining dropout in BSI at UNIRIO.

The curricular activity incurring in the most failures is Graduation Project II. This curricular activity is the second step of the writing of the final-year project, which is done at the end of the course. A hypothesis to explain why this curricular activity has many failures is that students can delay the presentation of their final-year project to the following semester, resulting in a failure. A similar hypothesis can be formulated for Extension Curriculum Activities 1 in which students must present university documents of activities they are doing outside, such as internships, courses, or sports. The other eight curricular activities refer to the introduction to programming and mathematics subjects that are concentrated in the first half of the course. Therefore, these curricular activities are the most decisive for university dropouts in BSI at UNIRIO. They will be considered in out final model.

Most dropouts occurred in the first two years of the course. So, the model can predict early students at risk of dropout by the accumulated GPA of the first two semesters, especially with students who get failures in introduction to programming and mathematics curricular activities. Schoeffel et al. (2020) found out that students' motivation in introduction to programming can be a indicative of success or dropouts. As such, in a future work, it is possible to verify the students' motivation regarding the

BSI's introduction to programming.

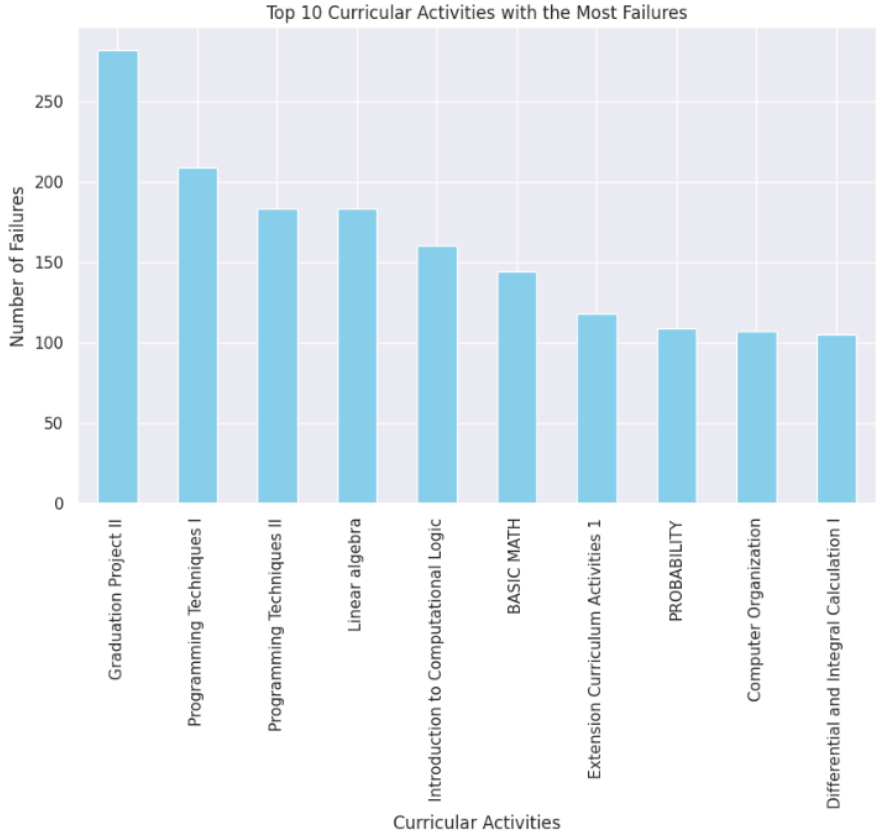


Figure 5.9: Top 10 curricular activities with the most student's failures

5.4 Data analysis in STEM courses using academic and financial data

Figure 5.10 represents the CCET dropout rates from 2013 to 2023, after the adoption of the Unified Selection System (SiSU). Information Systems got 73% of dropout rates, Mathematics got 86% and Production Engineering got 67%.

Figure 5.11 and Figure 5.12 represent the accumulated GPA from students who are employed in full-time jobs until the fourth semester and who became company owners. Students who became company owners got a median accumulated GPA of 5.49 while students who did not become company owners got a median of 4.21. Students who worked full-time during graduation got a median of accumulated GPA of 4.00 while students who did not work full-time got a median of 4.3. Our literature review(Rodrigues *et al.*, 2024a) observed that dropouts are influenced by academic performance. There is no significant difference in academic performance observed in these two groups of students.

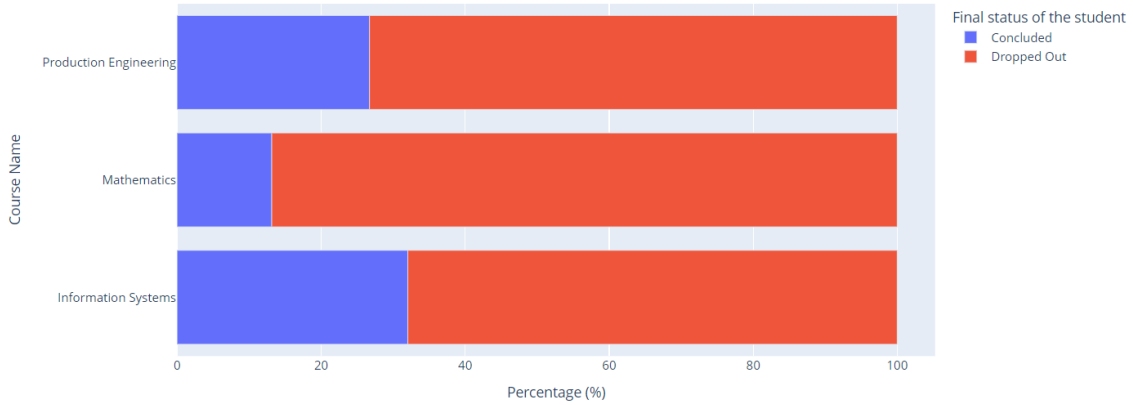


Figure 5.10: Dropout rates on CCET

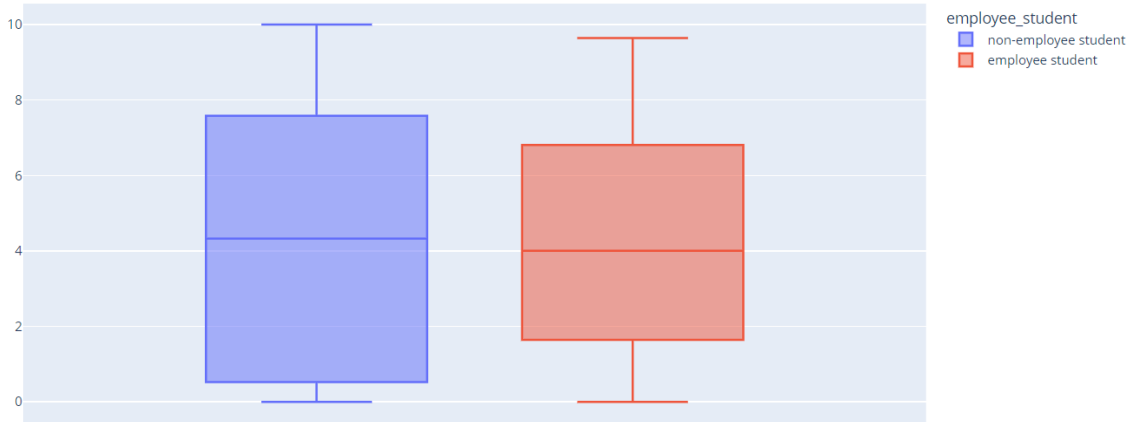


Figure 5.11: Accumulated GPA of employee-students at CCET

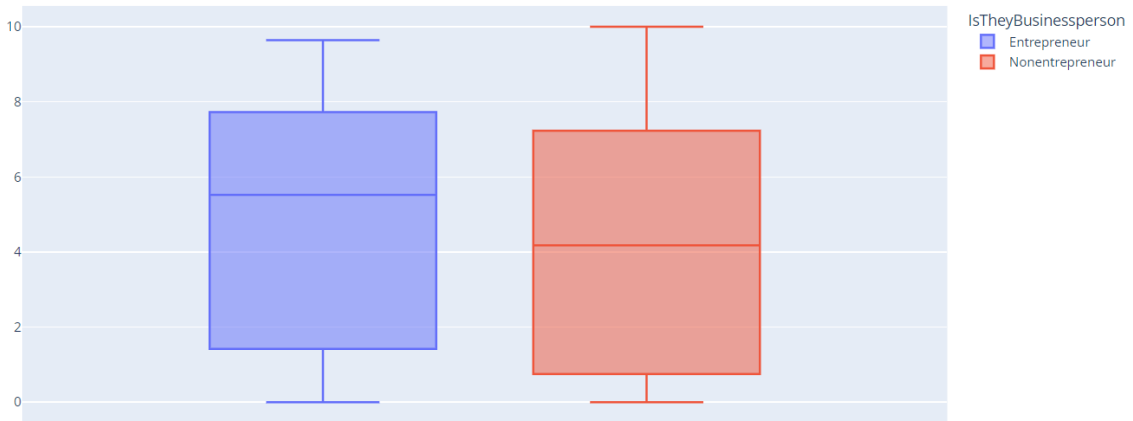


Figure 5.12: Accumulated GPA of students who become company owners at CCET

5.4.1 (sub-RQ4) Is there an association between full-time employment and dropout rates at CCET?

Table 5.8: Graduation or dropout by full-time work

Graduation or dropout by full-time work		
Quantity of employee-students	**Graduated employee-students**	**Dropped out employee-students**
213	22.06%	77.94%
Quantity of non-employee-students	**Graduated non-employee-students)**	**Dropped out non-employee-students**
761	24.31%	75.69%

Table 5.8 represents the dropout rates between students who worked full-time and students who did not work full-time. The chi-square statistic was calculated, obtaining a p-value of 0.55. The null hypothesis was that employment and the student's outcome (graduation or dropout) were independent. Since we did not reject the null hypothesis, we do not have sufficient evidence to state that there is an association between employment and outcome.

5.4.2 (Sub-RQ5): Is there an association between entrepreneurship and dropout rates at CCET?

Table 5.9: Graduation or dropout by company ownership

Graduation or dropout by company ownership		
Quantity of company owner students	**Graduated company owner students**	**Dropped out company owner students**
95	26.32%	73.68%
Quantity of non-company owner students	**Graduated non-company owner students)**	**Dropped out non-company owner students**
879	23.55%	76.45%

Table 5.9 represents the dropout rates between students who became company owners and students who did not become company owners. The chi-square statistic was calculated, obtaining a p-value of 0.63 The null hypothesis was that entrepreneurship and the student's outcome (graduation or dropout) were independent. Since we did not reject the null hypothesis, we do not have sufficient evidence to state that there is an association between entrepreneurship and outcome.

Figure 5.13 represents the category of company owners that studied in CCET, including those who graduated successfully and those who dropped out. Students who founded companies after graduating or dropping out were included because if they were

excluded, there would not be enough data. Therefore, we reframed the original scope of the sub-RQ5 to include them.

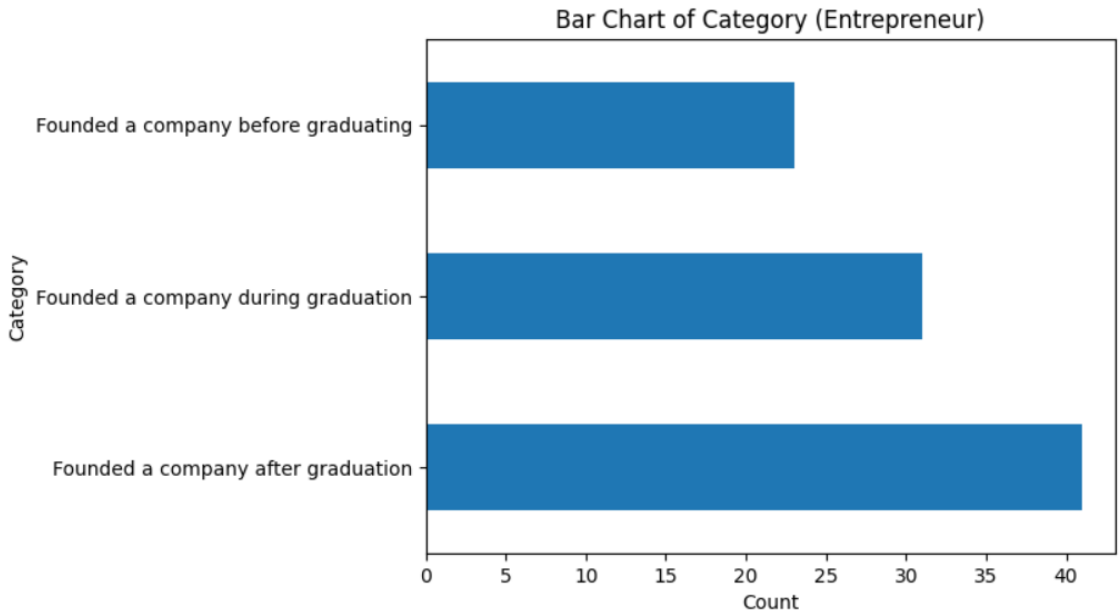


Figure 5.13: Category of the company owners of CCET

5.4.3 (Sub-RQ6): Do university scholarships help reduce dropout rates at CCET?

Table 5.10 represents the dropout rate between students who had or did not have scholarships. The chi-square statistic was calculated, obtaining a p-value <0.00001. The null hypothesis that the presence of scholarship and the student's outcome (graduation or dropout) were independent was rejected, indicating that students with scholarships graduate more than students without scholarships.

Table 5.10: Graduation or dropout by presence of scholarship

Graduation or dropout by presence of scholarship		
Quantity of students who had scholarships	**Graduated students who had scholarships**	**Dropped students who had scholarships**
176	65.90%	34.10%
Quantity of students who did not have scholarships	**Graduated non-employee-students)**	**Dropped students who did not have scholarships**
798	14.53%	85.47%

Figure 5.14 represents the dropout rates by admission methods and scholarships. Students admitted by quotas who did not receive scholarships got a dropout rate of

89.23%. Students admitted by quotas who received scholarships got a dropout rate of 54.21%. Students admitted by free concurrence who did not receive scholarships got a dropout rate of 82.87%. Students admitted by free concurrence who received scholarships got a dropout rate of 16.12%. There are two types of scholarships: academic and social. We planned to analyze each type separately, but we did not have enough data on social scholarships. Although it was observed that students who got scholarships graduated more, we can not explain, but there are hypotheses: students with better grades got academic scholarships, therefore they already have a higher probability of graduating, or students who got scholarships got more involved in academic life, which reduced the probability of dropout. Suggestions for future work include an in-depth analysis of different scholarship types.

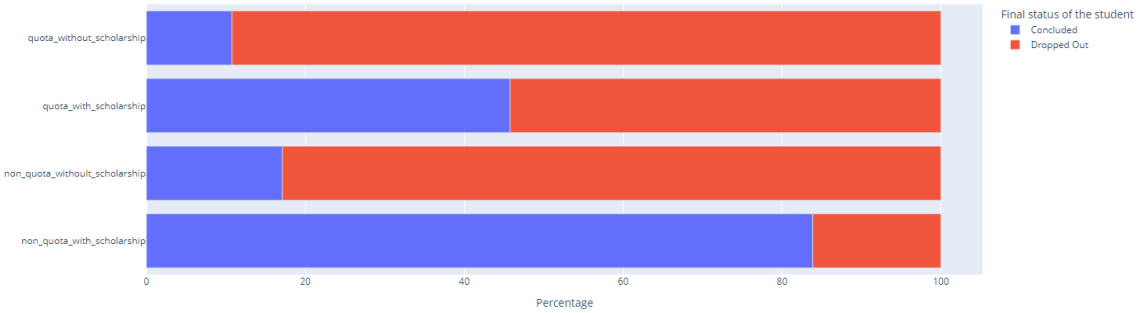


Figure 5.14: Dropout rates by admission methods and scholarships

Figure 5.15 represents the difference in academic performance between students who received scholarships and those who did not. Students who received scholarships during graduation got a median GPA of 8.4 and students who did not got a median GPA of 3.25.

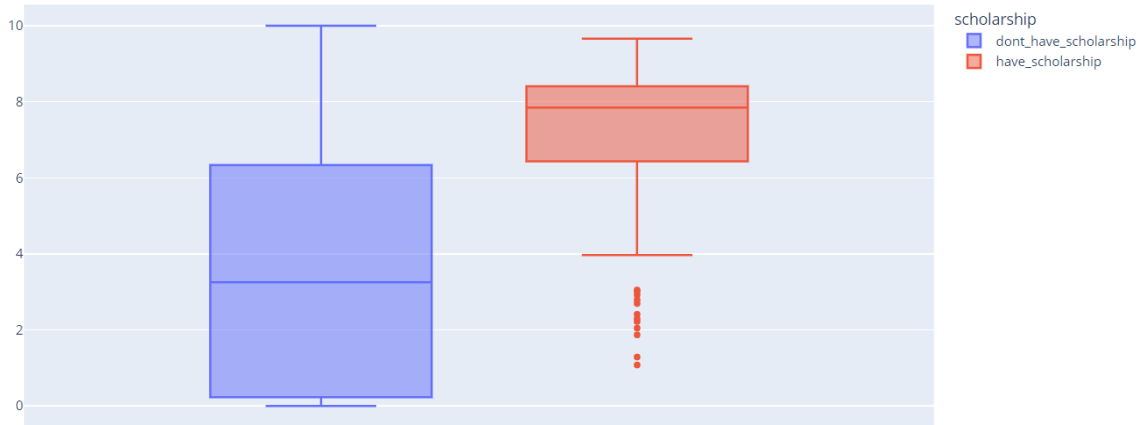


Figure 5.15: Accumulated GPA per students that have scholarship and students that don't have scholarship

6. Modeling

6.1 Gradient Boosting Models

Following the training of the preliminary Decision Tree model presented in Chapter 5, which only included academic data of Information Systems, Gradient Boosting models were trained that focus on the early prediction of dropout, which focuses on the first four semesters of the undergraduate programs due the fact that most dropouts occur on these semesters.

In this context, these models include not only academic data but also financial data, which include full-time work, company ownership, and scholarships. 80% of the data of each individual student were used to train the model, while 20% of the data were used to test the model. There was no crossing validation.

There are four models for each of the undergraduate programs: Information Systems, Production Engineering, and Mathematics.

All models have financial and socio-demographic data. The first model of each course has the first semester GPA and the grades of the disciplines of the first semester. The second model has the first, and second GPAs and the grades of disciplines of the first and second semesters. The third model has all the cited data plus the third GPA, and the fourth model has all the previously cited data plus the GPA of the fourth semester.

There are also general models for the whole CCET. There are also four CCET's models, following a similar structure, but without discipline grades.

To verify specifically the sub-RQ7, it was trained two models without using financial data to compare with the ones trained with financial data. The comparison was made on the general models regarding the whole CCET in the first and fourth semesters.

For each of the three courses, it was created 4 four models to predict the final status of the student. All models have the following features: Admission method, gender, IsTheyBusinessPerson, Category of businessperson, and IsTheyEmployeeStudent. Each of the four models per course represents progress in the course, therefore, the first model have the first semester GPA and the grades of the first semester disciplines that have more retention, the second model has the first semester GPA, second semester GPA, and the grades of the first-semester and second-semester disciplines that have more retention.

These disciplines were chosen following the results of the preliminary study, which had shown poor academic performance impacts dropout. The third and fourth models have in addition to the previous models the third semester GPA and the fourth semester GPA, respectively.

6.2 Process of training the models

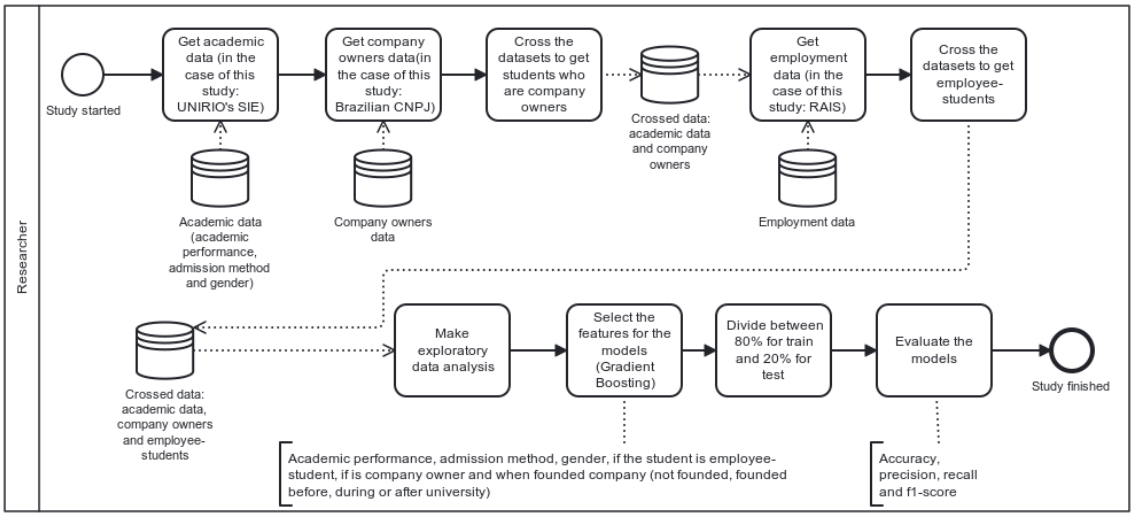


Figure 6.1: Study process

Figure 6.1 represents the process of the study, which uses the Business Process Modelling Notation (BPMN) (White, 2004). Firstly, it was retrieved academic data from UNIRIO's SIE. Secondly, the academic data was crossed with CNPJ data to get which students became company owners. Thirdly, this crossed data was crossed again with RAIS's data to get which students worked full-time during graduation. Finally, with the three original datasets crossed, the data analysis was performed and the model was trained with the final dataset that was built.

The data crossings were made using the CPF, which can be found on the three datasets. To use the CPF, it was necessary to seek approval from the Brazilian Platform Ethics Committee, which was discussed Subsection in 4.1 of the Chapter 4.

6.3 Results

We created models focused on the student that left the undergraduate program, either successfully or not, using the following features: admission method to the course, accumulated GPA, semester GPA from the first to fourth semester, gender, if the student

owns a company in a moment of their life, category of businessperson (not businessperson, founded a company before, during, or after the college), if the student works on a full-time job while studying, and the course.

These features were used to develop the model because features related to academic performance and socioeconomic conditions were the most common to make this kind of model and because according to (Tinto, 1975), these features can be classified as Family Background (Admission Method), Individual Attributes (gender and financial factors related to professional life), which influences the goal commitment to academic performance, which influences the students' decision to drop out or not. According to the literature reviews on this topic, features related to employment and entrepreneurship are poorly explored by related work (Silva; Roman, 2021; Tete *et al.*, 2022; Rodrigues *et al.*, 2024a)

Table 6.1 represent the model's evaluation metrics for the first four semesters of Information System, 6.2 represent the model's evaluation metrics for the first four semesters of Production Engineering, Table 6.3 represent the model's evaluation metrics for the first fours semesters of Mathematics and Table 6.4 represent the general model's evaluation metrics for the first fours semesters of the whole CCET. Figure 6.2 represents a graph of the F1-Score across the four models of the undergraduate programs and the CCET. It can be seen that this metric grows over time, indicating that the models have a higher predictive power with more academic performance data.

Table 6.5 represents the feature importance of the general model of the first semester of CCET. The most important feature considered by this model is related to academic performance, and the second most important whether students receive scholarships or not.

Table 6.1: Information System's model's accuracy and classification report

Information Systems model 1 (GPA until first semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.82	0.92	0.87	49
Graduated students	0.78	0.58	0.67	24
		accuracy	0.88	
Information Systems model 2 (GPA until second semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.90	0.88	0.89	49
Graduated students	0.76	0.79	0.78	24
		accuracy	0.84	
Information Systems model 3 (GPA until third semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.90	0.90	0.90	49
Graduated students	0.79	0.79	0.79	24
		accuracy	0.86	
Information Systems model 4 (GPA until fourth semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.92	0.90	0.91	49
Graduated students	0.80	0.83	0.82	24
		accuracy	0.87	

Table 6.2: Production Engineering's models accuracy and classification report

Production Engineering's model 1 (GPA until first semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.90	0.88	0.89	42
Graduated students	0.58	0.64	0.61	11
		accuracy	0.83	
Production Engineering model 2 (GPA until second semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.91	0.95	0.93	42
Graduated students	0.78	0.64	0.70	11
		accuracy	0.88	
Production Engineering model 3 (GPA until third semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.93	0.93	0.93	42
Graduated students	0.73	0.73	0.73	11
		accuracy	0.88	
Production Engineering model 4 (GPA until fourth semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.95	0.93	0.94	42
Graduated students	0.75	0.82	0.78	11
		accuracy	0.90	

Table 6.3: Mathematics's models accuracy and classification report

Mathematics model 1 (GPA until first semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.91	0.95	0.93	62
Graduated students	0.40	0.25	0.31	8
		accuracy	0.87	
Mathematics model 2 (GPA until second semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.95	0.97	0.96	62
Graduated students	0.71	0.62	0.67	8
		accuracy	0.92	
Mathematics model 3 (GPA until third semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.97	0.97	0.97	62
Graduated students	0.75	0.75	0.75	8
		accuracy	0.94	
Mathematics model 4 (GPA until fourth semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.98	0.97	0.98	62
Graduated students	0.78	0.88	0.82	8
		accuracy	0.95	

Table 6.4: CCET's general models accuracy and classification report

CCET model 1 (GPA until first semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.88	0.92	0.90	146
Graduated students	0.70	0.60	0.64	47
		accuracy	0.83	
CCET model 2 (GPA until second semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.94	0.92	0.93	146
Graduated students	0.78	0.81	0.79	47
		accuracy	0.89	
CCET model 3 (GPA until third semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.93	0.95	0.94	146
Graduated students	0.82	0.79	0.80	47
		accuracy	0.90	
CCET model 4 (GPA until fourth semester)				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.93	0.95	0.94	146
Graduated students	0.82	0.77	0.79	47
		accuracy	0.90	

Table 6.5: Analysis of the importance of variables for the CCET model 1

Feature	**Importance**
First Semester GPA	0,126141
Gender	0,028320
Has Scholarship	0,105705
Admission Method	0,047822
IsTheyBusinessperson	0,000000
Category of Businessperson	0,007573
IsTheyEmployeeStudent	0,001971

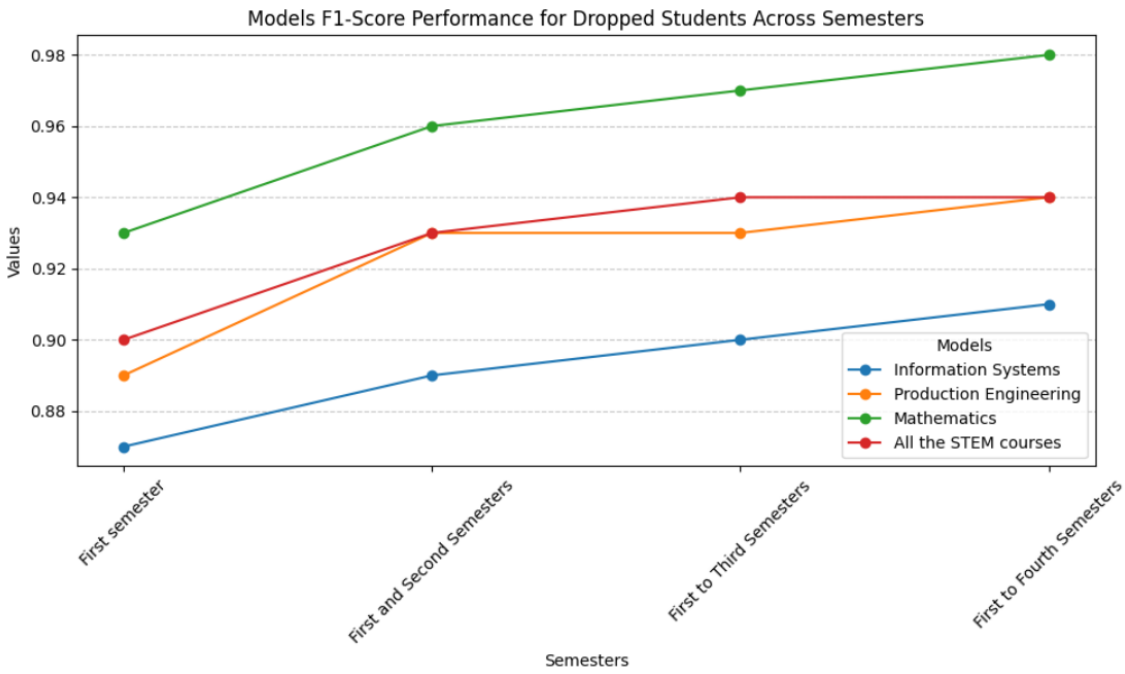


Figure 6.2: F1 Score across the models

6.3.1 (Sub-RQ7): Does a model that uses financial factors along with academic data have greater predictive power than a model that considers only academic data?

Table 6.6 represents the evaluation metrics on a CCET model trained solely using academic performance on the first semester.

Table 6.6: Overall accuracy and classification report without financial data: CCET model 1

CCET Model 1 without financial data				
	precision	**recall**	**f1-score**	**support**
Students who dropped out	0.82	0.90	0.86	148
Students who graduated	0.50	0.33	0.40	45
overall accuracy		0.76		

Table 6.7 represents the evaluation metrics on a CCET model trained solely using academic performance on the first semester to the fourth semester.

Table 6.7: Overall accuracy and classification report without financial data: CCET model 4

CCET Model 4 without financial data				
	precision	**recall**	**f1-score**	**support**
Students who dropped out	0.92	0.93	0.92	146
Students who graduated	0.75	0.73	0.74	47
overall accuracy		0.88		

There was an improvement in the metrics of the class of dropout students, mainly recall, when incorporating socioeconomic variables for the models of the first period, while for the models of the fourth period there was no significant improvement. However, the socioeconomic variable considered to have the greatest predictive power by the model was whether the student receives a scholarship or not, suggesting that some financial factor may be related to dropout.

6.3.2 Gender fairness in the model

Fairness in Machine Learning is defined as the avoidance of bias in a model regarding gender, race, disabilities, and other characteristics (Caton; Haas, 2024).

To verify if the models are fair with genders, it was tested the CCET model 1 with samples only containing male or female students, exclusively.

The Tables 6.8 and 6.9 represents the evaluation for the tests made exclusively by samples only containing males and female students respectively.

Although there are more male students in the CCET, this tested model had a higher accuracy in predicting female students' outcomes, especially females that would graduate, as seen in higher precision, recall, and f1-score of graduate females.

Table 6.8: CCET Model 1 accuracy and classification report for male students

CCET Model 1 accuracy and classification report for male students				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.88	0.88	0.88	117
Graduated students	0.53	0.53	0.53	30
		accuracy	0.80	

Table 6.9: CCET Model 1 accuracy and classification report for female students

CCET Model 1 accuracy and classification report for female students				
	precision	**recall**	**f1-score**	**support**
Dropped out students	0.80	0.97	0.88	29
Graduated students	0.91	0.59	0.71	17
		accuracy	0.82	

6.4 Discussion

According to the literature reviews conducted by (Silva; Roman, 2021), (Tete *et al.*, 2022) and (Rodrigues *et al.*, 2024a), this research was one of the first to verify if there is

an association between dropout and professional activities, such as full-time work and entrepreneurship, using EDM and AI prediction, but no such association was found, contrary to studies that used other methods (Hovdhaugen, 2013; Kocsis; Pusztai, 2020). Students that work and study simultaneously or that own a company have similar behaviour in dropout as students that do not work or do not own companies. Therefore, in the context of CCET/UNIRIO, these are not factors that influence dropout.

It was found that the insertion of financial data increased the predictive capacity of the model more with an academic performance from the first period than from the fourth, indicating that socioeconomic variables are more useful at the beginning of the course, before the grade for several periods is obtained.

The models can be considered fair regarding gender, as they performed similarly in tests containing only males or only females. However, it is worth noting that they were more successful in predicting female graduates We did not explore an explanation for that, but a hypothesis is that there are fewer females than males, therefore, the model could tend to overfit for female students. However, data from the preliminary study only using academic data of Information Systems had shown that dropout rates are higher among male students, so, it is possible that the model considered was fair enough to behave in a such way to perform better predicting graduate females.

7. Deployment

7.1 Architecture of the Dropout Predictor System

It was built a web system with the models presented in Chapter 6. The system was designed to be used by academic managers to identify students at risk of dropping out. Figure 7.1 shows the architecture of the web system, and Figure 7.2 shows the use cases (Seidl *et al.*, 2015) of the system. The use cases are described in Appendix A The system code can be found on GitHub at the link: <https://github.com/HenriqueSoaresRodrigues/Sistema-de-Predicao-de-Evasao>.

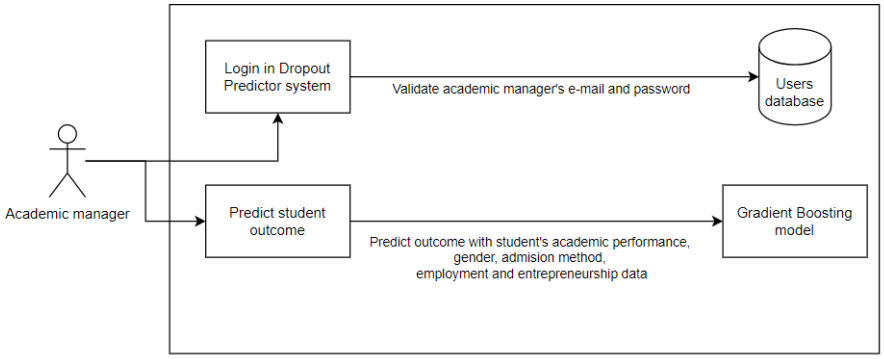


Figure 7.1: Dropout Predictor system: examples of prediction

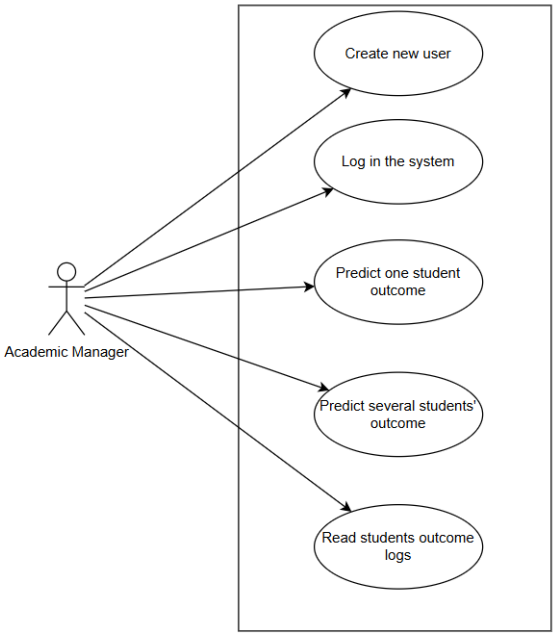


Figure 7.2: Dropout Predictor system: examples of prediction

7.2 Use Examples of the Dropout Predictor System

Figures 7.4 and 7.3 show two examples of the use of the system. Figure 7.5 shows the input page for Comma-Separate Values (CSV) files, where the user can generate predictions for several students with a CSV file generated by the process of crossing data in Google Colab. Figure 7.6 shows an example of logs generated after the predictions. The system is currently hosted at <https://sistemapreditorevasao.onrender.com/>

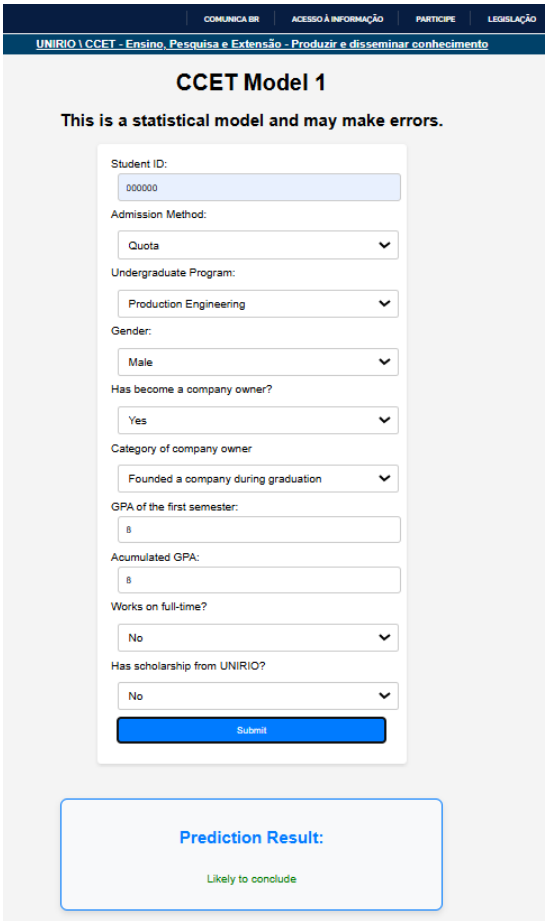


Figure 7.3: Dropout Predictor system: example of prediction of graduation

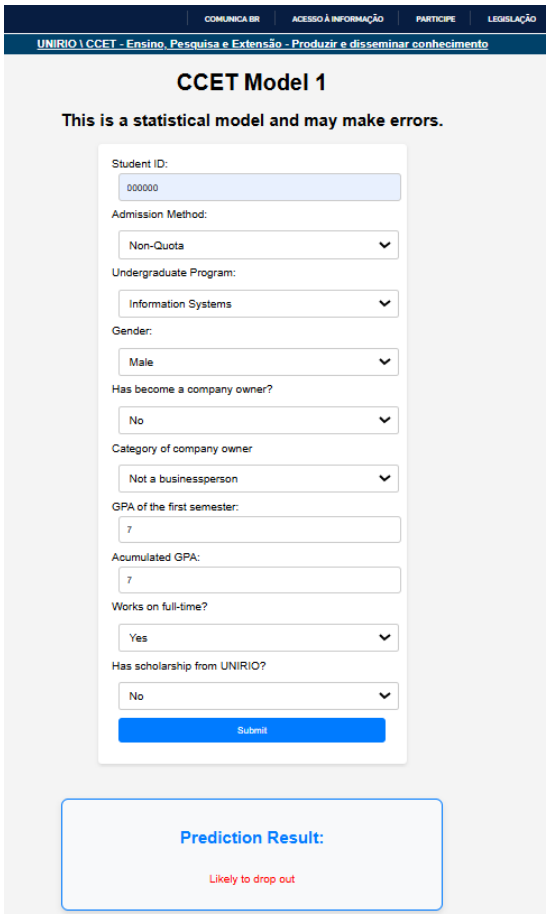


Figure 7.4: Dropout Predictor system: examples of prediction of dropout

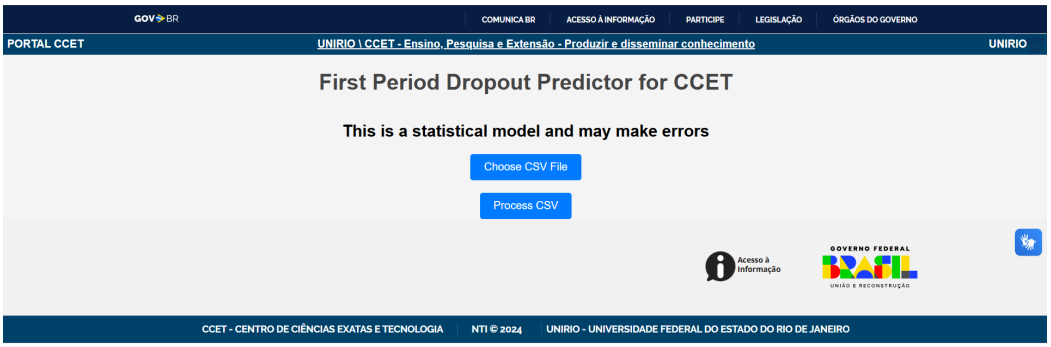
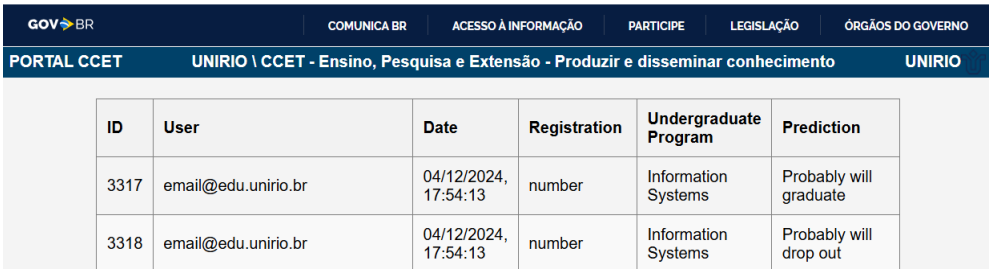


Figure 7.5: Dropout Predictor system: examples of CSV input page



ID	User	Date	Registration	Undergraduate Program	Prediction
3317	email@edu.unirio.br	04/12/2024, 17:54:13	number	Information Systems	Probably will graduate
3318	email@edu.unirio.br	04/12/2024, 17:54:13	number	Information Systems	Probably will drop out

Figure 7.6: Dropout Predictor system: examples of prediction logs

8. Final Remarks

8.1 Conclusion

The results of this analysis and the models on the data of the two primary studies show that students facing difficulties with curricular activities along the course, especially during the first half of the course, have a considerable probability of dropout.

This research did not find an association between dropout on UNIRIO'S STEM courses and financial factors, such as professional activities, such as full-time work, and company ownership.

It was found that the insertion of financial data increased the predictive capacity of the model more with the academic performance from the first period than from the fourth, indicating that socioeconomic variables are more useful at the beginning of the course before the grade for several semesters is obtained

A system called Dropout Prediction was built for UNIRIO's STEM courses to be used by academic managers to identify students at risk of dropout. This system is expected to help academic managers reduce the dropout rates in UNIRIO STEM courses.

This research reinforces the need for HEI, such as UNIRIO, to implement education policies to help students at risk of dropping out to continue their studies and eventually graduate with success.

8.2 Limitations

This research only covered STEM courses of UNIRIO, therefore the conclusion of the analysis may not be applicable to other contexts, such as other UNIRIO's courses or other universities, or other geographical places. It was considered to investigate the impact of internships on dropout, but internship data was not available.

The SMS conducted to research the status quo of the use of AI to predict academic performance showed that generally previous grades are used to predict future grades. This research also had difficulties in trying to predict the GPA using other available factors beyond previous academic performance, therefore, no algorithms to predict academic performance were presented, although it was originally planned at the start of the research.

The Gradient Boosting model used in the Dropout Prediction system was trained with data from BSI that precedes a curricular reform implemented in 2023.2, therefore, when in use by BSI academic managers, the system will not be in accordance with the current state of the BSI. In the future, a new model will be trained in accordance with the new BSI curricular components. This problem does not occur in Product Engineering and Mathematics.

Another limitation is that the analysis was limited by academic and socio-economic factors. Psychological and health factors were not considered.

8.3 Future Works

It is expected to include all undergraduate programs from UNIRIO in the Dropout Prediction system in the future, as well as include other factors in the system beyond academic performance and professional life.

Further investigation is needed to verify the reason why the models performed better in predicting graduate females. It will also be investigated in a data analysis of the dropout per credits already taken. Disciplines with more weekly classes hours have more credits.

It is considered to use cross-validation in the Gradient Boosting models and deploy the new models with cross-validation in the Dropout Predictor System as a future work.

It is considered in the future to investigate the impact of internships and compare the results of this research with undergraduate programs that have classes during only the morning and afternoon, as it is possible that dropouts on courses in daytime have differences between students who work and does not work.

It can also be cited as future work to conduct more research on influences on academic performance and build an AI model to predict it.

Bibliography

AGRUSTI, F.; MEZZINI, M.; BONAVOLONTÀ, G. Deep learning approach for predicting university dropout: A case study at roma tre university. **Journal of e-learning and knowledge society**, v. 16, n. 1, p. 44–54, 2020.

ALBAN, M.; MAURICIO, D. Predicting university dropout through data mining: A systematic literature. **Indian Journal of Science and Technology**, v. 12, n. 4, p. 1–12, 2019.

ALVIM, Í. V.; BITTENCOURT, R. A.; DURAN, R. S. Evasão nos cursos de graduação em computação no brasil. In: SBC. **Anais do IV Simpósio Brasileiro de Educação em Computação**. [S.l.], 2024. p. 1–11.

ANH, B. N. *et al.* An university student dropout detector based on academic data. In: **2023 IEEE Symposium on Industrial Electronics Applications (ISIEA)**. [S.l.: s.n.], 2023. p. 1–8.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. **Revista Brasileira de informática na educação**, v. 19, n. 02, p. 03, 2011.

BARDAGI, M.; HUTZ, C. S. Evasão universitária e serviços de apoio ao estudante: uma breve revisão da literatura brasileira. **Psicologia Revista**, v. 14, n. 2, p. 279–301, 2005.

BRASIL. Art.14 da portaria mec nº 18/2012. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2012. Available from: <https://www.unirio.br/caeg/sisu-sistema-de-selecao-unificada-1/legislacao/PORTARIANORMATIVAN18DE11DEOUTUBRODE2012AlteradapelaPortarian2.0272023.pdf>.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 2018. ISSN 1677-7042. Available from: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

BRASIL. **Censo da Educação Superior**. 2023. Https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior/resultados. Acessed in 10/30/2024.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, out. 2001. ISSN 1573-0565. Available from: <https://doi.org/10.1023/A:1010933404324>.

BRITAL, A. Random forest algorithm explained. 2021. Available from: <https://anasbrital98.github.io/blog/2021/Random-Forest/>.

CATON, S.; HAAS, C. Fairness in machine learning: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 56, n. 7, abr. 2024. ISSN 0360-0300. Available from: <https://doi.org/10.1145/3616865>.

CRUZ-CAMPOS, J.-C. de la; VICTORIA-MALDONADO, J.-J.; MARTÍNEZ-DOMINGO, J.-A.; CAMPOS-SOTO, M.-N. Causes of academic dropout in higher education in andalusia and proposals for its prevention at university: A systematic review. In: FRONTIERS MEDIA SA. **Frontiers in Education**. [S.l.], 2023. v. 8, p. 1130952.

DAZA, A.; GUERRA, C.; CERVERA, N.; BURGOS, E. A stacking based hybrid technique to predict student dropout at universities. **J Theor Appl Inf Technol**, v. 100, n. 13, p. 1–12, 2022.

European Commission. **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)**. European Commission, 2016. Available from: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

FERNáNDEZ-GARCíA, A. J. *et al.* A real-life machine learning experience for predicting university dropout at different stages using academic data. **IEEE Access**, v. 9, p. 133076–133090, 2021.

FOERSTER, H. V. Ethics and second-order cybernetics. **Understanding understanding: Essays on cybernetics and cognition**, Springer, p. 287–304, 2003.

FREIRE, J.; LANDIM, F.; MORAES, L.; DELGADO, C.; PEDREIRA, C. Modelo para previsão precoce de abandono de uma disciplina de introdução à programação. In: **Anais do XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024. p. 635–645. ISSN 2595-6175. Available from: <https://sol.sbc.org.br/index.php/wei/article/view/29663>.

FRIEDMAN, J. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, 11 2000.

FÜRNKRANZ, J. Decision tree. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 263–267. ISBN 978-0-387-30164-8. Available from: <https://doi.org/10.1007/978-0-387-30164-8_204>.

GISMONDI, H. E. C.; HUIMAN, L. V. U. Multilayer neural networks for predicting academic dropout at the national university of santa - peru. In: **2021 International Symposium on Accreditation of Engineering and Computing Education (ICACIT)**. [S.l.: s.n.], 2021. p. 1–4.

GUO, H.; NGUYEN, H.; VU, D.-A.; BUI, X.-N. Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. **Resources Policy**, v. 74, p. 101474, 2021. ISSN 0301-4207. Available from: <https://www.sciencedirect.com/science/article/pii/S0301420718306901>.

GUTIERREZ-PACHAS, D. A.; GARCIA-ZANABRIA, G.; CUADROS-VARGAS, E.; CAMARA-CHAVEZ, G.; GOMEZ-NIETO, E. Supporting Decision-Making Process on Higher Education Dropout by Analyzing Academic, Socioeconomic, and Equity Factors through Machine Learning and Survival Analysis Methods in the Latin American Context. **Education Sciences**, v. 13, n. 2, 2023. ISSN 2227-7102. Available from: <https://www.mdpi.com/2227-7102/13/2/154>.

HEMASHREEKILARI. Understanding gradient boosting. 2023. Available from: <https://medium.com/@hemashreekilari9/understanding-gradient-boosting-632939b98764>.

HERINGER, R. Affirmative action policies in higher education in brazil: Outcomes and future challenges. **Social Sciences**, v. 13, n. 3, 2024. ISSN 2076-0760. Available from: <https://www.mdpi.com/2076-0760/13/3/132>.

HOVDHAUGEN, E. Working while studying: the impact of term-time employment on dropout rates. **Journal of Education and Work**, Routledge, v. 28, n. 6, p. 631–651, 2013. Available from: <https://doi.org/10.1080/13639080.2013.869311>.

JIMENEZ-MACIAS, A.; MORENO-MARCOS, M.; MERINO, P.; ORTIZ, M.; DELGADO-KLOOS, C. Analyzing feature importance for a predictive undergraduate student dropout model. **Computer Science and Information Systems**, v. 20, p. 50–50, 01 2022.

KITCHENHAM, B. A. Systematic review in software engineering: Where we are and where we should be going. In: **Proceedings of the 2nd International Workshop on Evidential Assessment of Software Technologies**. New York, NY, USA: Association for Computing Machinery, 2012. (EAST '12), p. 1–2. ISBN 9781450315098. Available from: <https://doi.org/10.1145/2372233.2372235>.

KOCSIS, Z.; PUSZTAI, G. Student employment as a possible factor of dropout. **Acta Polytechnica Hungarica**, v. 17, n. 4, p. 183–199, 2020.

KOTSIANTIS, S. B.; PIERRAKEAS, C. J.; PINTELAS, P. E. Preventing student dropout in distance learning using machine learning techniques. In: PALADE, V.; HOWLETT, R. J.; JAIN, L. (Ed.). **Knowledge-Based Intelligent Information and Engineering Systems**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 267–274. ISBN 978-3-540-45226-3.

LÓPEZ-ANGULO, Y.; SÁEZ-DELGADO, F.; MELLA-NORAMBUENA, J.; BERNARDO, A. B.; DÍAZ-MUJICA, A. Predictive model of the dropout intention of chilean university students. **Frontiers in Psychology**, Frontiers, v. 13, p. 893894, 2023.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine Learning: An Artificial Intelligence Approach (Volume I)**. [S.l.]: Elsevier, 2014. v. 1.

MOSELEY, L.; MEAD, D. Predicting who will drop out of nursing courses: A machine learning exercise. **Nurse education today**, v. 28, p. 469–75, 06 2008.

NAGY, M.; MOLONTAY, R. Interpretable dropout prediction: Towards xai-based personalized intervention. **International Journal of Artificial Intelligence in Education**, 03 2023.

NAIDU, G.; ZUVA, T.; SIBANDA, E. M. A review of evaluation metrics in machine learning algorithms. In: SPRINGER. **Computer Science On-line Conference**. [S.l.], 2023. p. 15–25.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, Frontiers Media SA, v. 7, p. 21, 2013.

OLIVEIRA, R. dos S.; MEDEIROS, F. P. A. de. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. **Revista Brasileira de Informática na Educação**, v. 32, p. 1–21, 2024.

OPAZO, D.; MORENO, S.; MIRANDA, E. Álvarez; PEREIRA, J. Analysis of first-year university student dropout through machine learning models: A comparison between universities. **Mathematics**, v. 9, p. 2599, 10 2021.

OSMAN, A. F. Radiation oncology in the era of big data and machine learning for precision medicine. **Artificial Intelligence-Applications in Medicine and Biology. IntechOpen**, p. 41–70, 2019.

OSORIO, J.; SANTACOLOMA, G. Predictive model to identify college students with high dropout rates. **Revista Electrónica de Investigación Educativa**, v. 25, p. 1–10, 05 2023.

PACHAS, D. A. G. *et al.* A comparative study of who and when prediction approaches for early identification of university students at dropout risk. In: **2021 XLVII Latin American Computing Conference (CLEI)**. [S.l.: s.n.], 2021. p. 1–10.

PENNSTATE. A matrix formulation of the multiple regression model. 2018. Available from: <https://online.stat.psu.edu/stat462/node/132/,>.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and software technology**, Elsevier, v. 64, p. 1–18, 2015.

PRESTES, E. M. d. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da universidade federal da paraíba. **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 26, p. 869–889, 2018.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**. [S.l.: s.n.], 2016.

RAMIREZ, J. A. R.; GARCIA-BEDOYA, O.; GALPIN, I. Maximizing student retention using supervised models informed by student counseling data. CEUR-WS.org, v. 3282, p. 225–239, 2022. Available from: <http://ceur-ws.org/Vol-3282/icaiw_wdea_1.pdf>.

REALINHO, V.; MACHADO, J.; BAPTISTA, L.; MARTINS, M. V. Predicting student dropout and academic success. **Data**, v. 7, n. 11, 2022. ISSN 2306-5729. Available from: <https://www.mdpi.com/2306-5729/7/11/146>.

REVATHY, M.; KAMALAKKANNAN, S.; KAVITHA, P. Machine learning based prediction of dropout students from the education university using smote. In: **2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)**. [S.l.: s.n.], 2022. p. 1750–1758.

RODRIGUES, H. *et al.* Predicting student dropout on the information systems undergraduate program of unirio using decision trees. In: **Anais do XXXII Workshop sobre Educação em Computação**. Porto Alegre, RS, Brasil: SBC, 2024b. p. 588–598. ISSN 2595-6175. Available from: <https://sol.sbc.org.br/index.php/wei/article/view/29659>.

RODRIGUES, H. *et al.* Artificial intelligence algorithms to predict college students' dropout: A systematic mapping study. In: INSTICC. **Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART**. [S.l.]: SciTePress, 2024a. p. 344–351. ISBN 978-989-758-680-4. ISSN 2184-433X.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. **Estudos Econômicos (São Paulo)**, SciELO Brasil, v. 49, p. 337–373, 2019.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. ii—recent progress. **IBM Journal of research and development**, IBM, v. 11, n. 6, p. 601–617, 1967.

SANI, N. S.; FIKRI, A.; ALI, Z.; ZAKREE, M.; NADIYAH, K. Drop-out prediction in higher education among b40 students. **International Journal of Advanced Computer Science and Applications**, v. 11, p. 550–559, 11 2020.

SANTOS, G. *et al.* Evolvedtree: Analyzing student dropout in universities. In: **2020 International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.: s.n.], 2020. p. 173–178.

SANTOS, V. H. B. dos; SARAIVA, D. V.; OLIVEIRA, C. T. de. Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In: SBC. **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. [S.l.], 2021. p. 1196–1210.

SCHOEFFEL, P.; RAMOS, V. F. C.; WAZLAWICK, R. S. A method to predict at-risk students in introductory computing courses based on motivation. In: SOCIEDADE BRASILEIRA DE COMPUTAçãO. **Anais do 9º Concurso Alexandre Direne de Teses de Doutorado - Congresso Brasileiro de Informática na Educação (CBIE)**. Porto Alegre, 2020. p. 41–41.

SEIDL, M.; SCHOLZ, M.; HUEMER, C.; KAPPEL, G. The use case diagram. In: ____. **UML @ Classroom: An Introduction to Object-Oriented Modeling**. Cham: Springer International Publishing, 2015. p. 23–47. ISBN 978-3-319-12742-2. Available from: <https://doi.org/10.1007/978-3-319-12742-2_3>.

SILVA, J.; ROMAN, N. Predicting dropout in higher education: a systematic review. In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 1107–1117. ISSN 0000-0000. Available from: <https://sol.sbc.org.br/index.php/sbie/article/view/18134>.

SOLIS, M.; MOREIRA-MORA, T.; GONZALEZ, R.; FERNANDEZ, T.; HERNANDEZ, M. Perspectives to predict dropout in university students with machine learning. In: **2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)**. [S.l.: s.n.], 2018. p. 1–6.

TETE, M. F.; SOUSA, M. d. M.; SANTANA, T. S. de; SILVA, S. F. Predictive models for higher education dropout: A systematic literature review. **Education Policy Analysis Archives**, v. 30, p. (149), Oct. 2022. Available from: <https://epaa.asu.edu/index.php/epaa/article/view/6845>.

TINTO, V. Drop-outs from higher education: A theoretical synthesis of recent research. **Review of Educational Research**, v. 45, p. 89–125, 03 1975.

UGONI, A.; WALKER, B. F. The chi square test: an introduction. **COMSIG review**, BMC, v. 4, n. 3, p. 61, 1995.

ULIYAN, D. *et al.* Deep learning model to predict students retention using blstm and crf. **IEEE Access**, IEEE, v. 9, p. 135550–135558, 2021.

WHITE, S. A. Introduction to bpmn. **Ibm Cooperation**, v. 2, n. 0, p. 0, 2004.

ZHANG, L.; RANGWALA, H. Early identification of at-risk students using iterative logistic regression. In: **2018 International Conference on Artificial Intelligence in Education**. [S.l.: s.n.], 2018.

ZIHAN, S.; SUNG, S.-H.; PARK, D.-M.; PARK, B.-K. All-year dropout prediction modeling and analysis for university students. **Applied Sciences**, v. 13, p. 1143, 01 2023.

APPENDIX A – Dropout Prediction System Use Cases Description

Create new user

Description	The system shall be able to create a new user.
Precondition	The user must be logged as a root user.
Ordinary Sequence	1. Actor academic manager request the system to the create new user page. 2. The system loads the create new user pager. 3. Actor academic manager enters the e-mail, password and the confirm_password of the new user. 4. The system checkers if the password is equal to confirm_password and create new user.
Postcondition	The new user can log in.
Exceptions	1. If the password and the confirm_password are not equal, the system restarts the use case.

Log in the system

Description	The system shall be able to let authorized user log in.
Precondition	The user must not be logged.

Ordinary Sequence	1. Actor academic manager request the log in page. 2. The system loads the log in page. 3. Actor academic manager enters the e-mail and password. 4. The system validates the e-mail and password to let the actor academic manager log in.
Postcondition	The user can now use the system.
Exceptions	1. If the system does not validate e-mail and password, the system restarts the use case.

Predict one student outcome

Predict one student outcome	The system shall be able to predict one student outcome.
Precondition	The user must be logged as a root or common user.
Ordinary Sequence	1. Actor academic manager request one of the models available. 2. The system loads the model page. 3. Actor academic manager enters data requested by the model. 4. The system predict an outcome based on the data.
Postcondition	The prediction done is now available to be read in the logs
Exceptions	1. If one of the numeric data is not in the correct format or if the data related to company ownership is incoherent, the system restarts the use case

Predict several students' outcome

Description	The system shall be able to predict several students' outcome.
Precondition	The user must be logged as a root or common user.
Ordinary Sequence	1. Actor academic manager request one of the models available. 2. The system loads the model page. 3. Actor academic manager enters a CSV generated by data crossed on Google Colab. 4. The system predict outcomes based on the data and generated a CSV file.
Postcondition	The predictions done are now available to be read in the logs
Exceptions	1. If the submitted CSV file is not suitable, the system restarts the use case.

Read students' outcomes logs

Description	The system shall be able to let previous predictions to be available to be read
Precondition	The user must be logged as a root or common user.
Ordinary Sequence	1. Actor academic manager request the logs page. 2. The system loads the logs pager. 3. Actor academic manager requests to download the CSV file. 4. The system let the browser download the CSV file.