Early detection of university dropout with cluster analysis and machine learning classification techniques

Nathalia R. González Duarte[1], Juan V. Bogado Machuca[2]

[1]*Faculty of Sciences and Technology, National University of Caaguazú*
Coronel Oviedo, 3300, Caaguazú, Paraguay
nrgonzalezd@fctunca.edu.py
[2]*National University of Caaguazú*
Coronel Oviedo, 3300, Caaguazú, Paraguay
jvbogado@fctunca.edu.py

Abstract. This study develops an early warning model to address student dropout at the Faculty of Sciences and Technologies of the National University of Caaguazú, employing advanced data science techniques. Initial cluster analysis identified four distinct groups of students: graduates, early dropouts, late dropouts, and thesis-stage students. Subsequently, a detailed characterization of these groups was carried out, followed by the training of various machine learning models across multiple data configurations. The models were evaluated using metrics such as precision, accuracy, recall, and F1 score. Of the configurations tested, the third was the most effective, showing how model performance varies among different academic programs. In the Computer Science program, the K-Nearest Neighbors model was the most effective. On the other hand, the Decision Tree model had the lowest performance in this field. In the Electrical and Civil Engineering programs, the Decision Tree (DT) model was more effective. In contrast, the KNN model was the least effective in these fields. In Electronics, the Logistic Regression and K-Nearest Neighbors models showed better performance. The results highlight the effectiveness of tailored models in early identification of at-risk students and recommend integrating socioeconomic and psychological factors for future research in this area.

Keywords: Data science, Student dropout, Predictive models, Machine learning, Cluster analysis.

1 Introduction

University dropout represents a significant challenge for the educational system, as it negatively impacts various aspects of national development[1]. This project was developed to address and better understand the patterns underlying this phenomenon, using Machine Learning technologies. The issue of high university dropout rates in Paraguay is a matter of great concern that reflects the difficulties students face in completing their studies in the country. Currently, only about 10% of young Paraguayans manage to finish their university studies, leaving an overwhelming majority of 90% who at some point make the difficult decision to abandon their studies [2]. In the Faculty of Sciences and Technologies, student dropout is a commonly observed problem and represents a weakness for the educational institution. The development plan for one of the programs states that it is necessary to consolidate the mechanism for analyzing retention, dropout, transfer, and promotion to evaluate the internal efficiency of the program [3]. Predicting student dropout using Machine Learning emerged as an alternative to address this issue, seeking to analyze historical data in order to identify early indicators that a student is at risk of abandoning their studies. This early identification can help intervene in a timely manner and provide personalized support, aiming to prevent dropout. The justification for this work was based on the belief that the predictive model developed with Machine Learning would provide detailed information about each student, which would facilitate the creation of intervention strategies tailored to individual needs. Also an innovative and similar approach was explored with socioeconomic information.[4]. The aim of this work is to identify students at high risk of dropping out based on academic factors using machine learning techniques.

2 Methodology

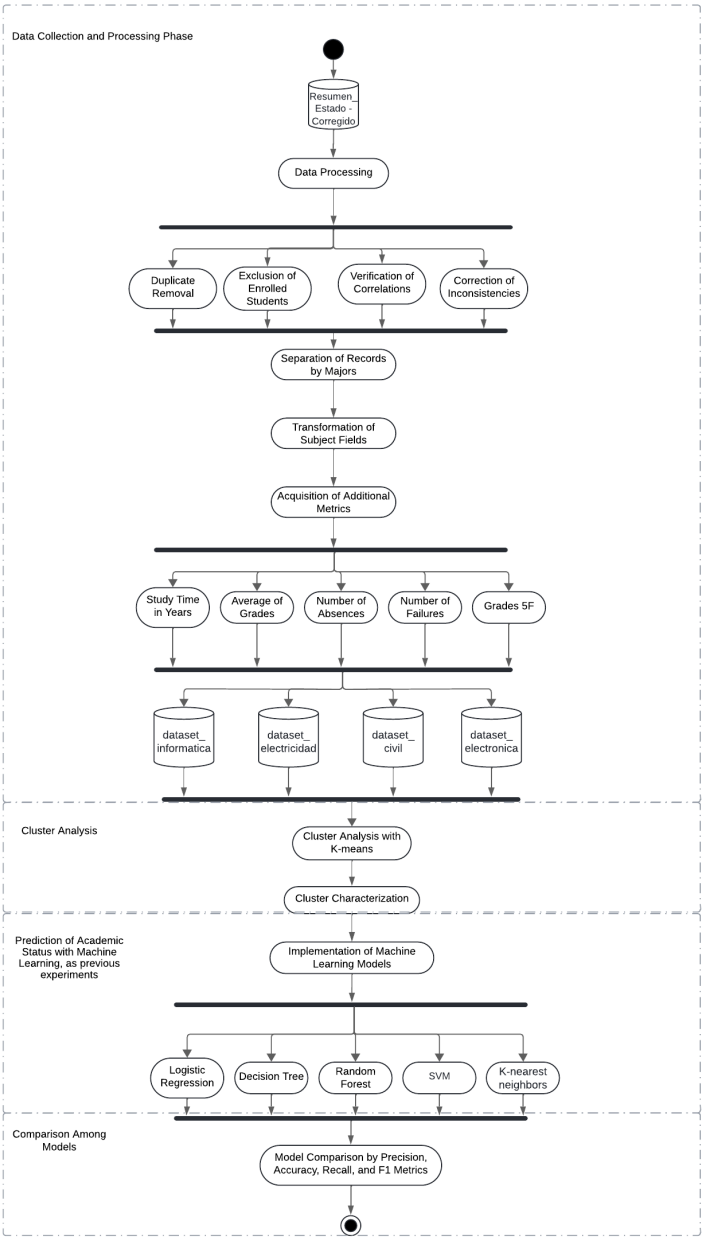

Figure 1. Summary workflow indicating the methodology used. Data collection, cluster analysis, model comparison.

This workflow diagram outlines the comprehensive process of data collection, processing, and analysis aimed at understanding academic performance across different majors. Initially, data is gathered and corrected for inaccuracies, which involves removing duplicate entries, excluding currently enrolled students, and correcting any inconsistencies. The data is then segmented by major, with transformation and additional metrics like study duration, average grades, and number of absences being calculated. Subsequent phases involve a detailed cluster analysis using the K-means algorithm to categorize the data, followed by the implementation of various machine learning models, including logistic regression, decision trees, Random Forest, SVM, and K-nearest neighbors. These models aim to predict academic outcomes and are evaluated based on precision, accuracy, recall, and F1 metrics to determine their effectiveness

2.1 Data Collection and Preprocessing

The data for this study were obtained from the Faculty of Sciences and Technologies at the National University of Caaguazú, with a dataset covering student records from 2010 to 2023, including 75,937 records of 1,422 students distributed across four engineering programs: Computer Engineering, Civil Engineering, Electrical Engineering, and Electronic Engineering. The dataset includes detailed fields such as date of entry, student ID, gender, program, subject ID, subject name, grades, evaluation period, academic year, and date of the final exam, among others. Data cleaning was performed, which involved removing duplicate records and correcting inconsistencies such as entry dates and first exam dates. Additionally, currently enrolled students were excluded to improve the accuracy of dropout predictions.

Additional measures such as study time, grade averages, absences, failures, and the "grade 5 congratulated" were employed. Each subject was treated as an individual field within the dataset, assigning the final grade of the student as the value.

Correlations of subjects for students who changed programs were verified, ensuring data consistency, and records were separated by program given the specificity of the subjects in each one.

2.2 Cluster Analysis

A cluster analysis was conducted to determine the existence of well-differentiated groups within the database. This analysis helped identify various statuses among the students, based on their academic activity and current status in their programs. The K-means algorithm was used, applying the elbow method to determine the number of groups formed, and then each cluster was analyzed using statistical techniques to characterize them.

2.3 Academic Status Prediction with Machine Learning

Once the number of existing clusters in the database and their characterization were determined, machine learning techniques were implemented to predict patterns of dropout and academic success. Different training and validation configurations were experimented with, which allowed for improved accuracy of the predictive models. The initial training phase used all relevant student data, while subsequent phases fine-tuned the model to achieve greater accuracy based on the students' initial years. Various models were used, including logistic regression, decision trees, random forest, SVM, and k-nearest neighbors. The results were compared using accuracy, recall, precision, and F1 score as metrics.

3 Results

In this section we describe the results obtained with the different combinations of machine learning techniques and evaluating them with accuracy, recall, f1 and precision.

3.1 Academic Data Analysis by Program

Table 1. Summary of Academic Data Analysis by Program

Program	Study Time (years)	Absences in Finals	Failures	Average Grade
Computer Engineering	3.46	6.90	4.46	3.24
Civil Engineering	3.70	8.00	5.13	3.20
Electrical Engineering	3.73	8.61	5.74	3.13
Electronic Engineering	3.62	8.42	6.46	3.15

With the data collected, an analysis was conducted, presenting additional variables for each major as shown in Table 1. This initial analysis focused on key academic characteristics of the students. After analyzing the presented data, different academic characteristics among students from the various majors analyzed were observed. This preliminary analysis has revealed not only variations in indicators such as the average of these variables. These differences underscore the need for a better understanding of the grouping and similarities among students, leading to the next crucial step in the research.

3.2 Cluster Analysis

To explore the presence of groupings within the dataset and determine similar academic characteristics among students, a clustering analysis using the K-Means algorithm was implemented. The elbow method was used to identify the optimal number of clusters, finding that four groupings (k=4) were ideal for capturing variability in the data. This analysis revealed that in the different programs of Computer Science, Electrical, Electronic, and Civil Engineering, the relationship between the number of clusters and distortion, at number K=4. Additionally, a visualization of the clusters was performed using principal component analysis (PCA) to reduce dimensionality to two main components. This simplified representation allowed for the observation of how the clusters dispersed in a two-dimensional space, revealing various groupings with some overlap among them. This pattern not only indicated significant variations in student profiles but also suggested the existence of common characteristics among students that could extend across several clusters. As the cluster analysis revealed the presence of four distinct groups within the dataset, it allowed for the definition of four specific academic states of interest for the research. These states are derived from unique combinations of academic characteristics, such as the completion of the curriculum and recent academic activity. Below is a description of each state, illustrated in table 2:

Table 2. Description of Defined Academic States

Academic State	**Description**
2	**Graduated Students:** Completion of the entire curriculum and final project.
3	**Early Dropouts:** No academic activity since 2018 and less than 5 years of study.
4	**Late Dropouts:** No academic activity since 2018 and more than 5 years of study.
5	**Pending Final Project:** Curriculum completed but the final project not approved.

3.3 Predictive Models

The experimental design utilized several machine learning models, including SVM, Random Forest, Logistic Regression, KNN, and Decision Tree. Initially, the models were trained and tested using a complete dataset with all performance variables, revealing high-performance metrics, but when tested with early-stage data, the models were not accurate. Subsequent experiments adjusted the scope of the data, focusing on the first three years and integrating different academic statuses to refine prediction accuracy, addressing the distribution mismatches observed in the initial tests.

Other experiments involved training models with only the first three years of data to address real-world prediction scenarios, revealing notable performance differences. However, at this point, the models confused graduates with students who only had their final project pending. Therefore, combining the categories of graduates and students with pending final projects improved model clarity and reduced confusion.

The experiments show that for the Computer Science major, the KNN model with data up to the third year and the combination of academic status categories 2 and 5 provided the best performance metrics. In Electrical Engineering, the Decision Tree stood out, showing high precision and accuracy, those results can be seen in Table 3. For Electronics, Logistic Regression was the most effective. In the Civil Engineering major, the Decision Tree consistently showed superiority.

The final models and data configurations reflect an approach aimed at optimizing accuracy and applicability in realistic scenarios where not all student data is immediately available. The decision to combine states 2 and 5 proved critical in improving prediction clarity and reducing confusion rates, suggesting that grouping similar categories can be an effective strategy in contexts where differences between states are minimal but critically important.

With this final configuration, the models significantly improved their ability to accurately predict academic outcomes in the four studied majors. This improvement was particularly notable in the context of more realistic and practical predictions, suitable for decision-making in educational environments where early interventions may be necessary. This approach highlights the importance of adapting machine learning models to the peculiarities of the data and the specific needs of the application environment to achieve optimal results.

Table 3. Summary of Predictive Models by Major, training, and testing with data up to the third year, merging states 2 and 5

Major	Metric	LR	DT	RF	SVM	KNN
Computer Science	ACC	0.827	0.793	0.862	0.862	**0.896**
	PREC	0.863	0.853	0.876	0.876	**0.896**
	RECALL	0.827	0.793	0.862	0.862	**0.896**
	F1	0.839	0.814	0.865	0.865	**0.895**
Electrical Engineering	ACC	0.901	**0.980**	0.921	0.862	0.823
	PREC	0.877	**0.981**	0.911	0.809	0.805
	RECALL	0.901	**0.980**	0.921	0.862	0.823
	F1	0.882	**0.979**	0.912	0.835	0.814
Electronics	ACC	**0.888**	0.777	0.833	0.833	**0.888**
	PREC	**0.898**	0.809	0.740	0.740	**0.898**
	RECALL	**0.888**	0.777	0.833	0.833	**0.888**
	F1	**0.882**	0.790	0.783	0.783	**0.882**
Civil Engineering	ACC	0.875	**0.968**	0.906	0.906	0.843
	PREC	0.889	**0.976**	0.920	0.909	0.850
	RECALL	0.875	**0.968**	0.906	0.906	0.843
	F1	0.878	**0.970**	0.911	0.904	0.843

4 Discussion

This analysis focuses solely on academic data as the university database does not contain records related to socioeconomic factors, social variables, and other potential influences on dropout rates. Additionally, the database lacks records of students who have dropped out; therefore, we had to perform calculations and assign statuses to those students who no longer had academic activity. This approach underscores the constraints of our dataset and highlights the need for a cautious interpretation of the results. Moving forward, integrating a broader spectrum of data, including dropout rates and external factors, could provide a more comprehensive understanding of the dynamics affecting academic outcomes and facilitate the development of more nuanced intervention strategies.

During the analysis, it was identified that there is an imbalance in the distribution of academic statuses, with some statuses being overrepresented compared to others. Ideally, a balanced dataset would allow the models to learn with equal representation of each status, thereby enabling a more equitable evaluation of the model's predictive capacity. However, given the nature of the available data and the importance of retaining as much information as possible for a thorough evaluation, the dataset was not balanced.

The decision to keep the dataset in its original form is based on the limited number of records available, which assigns significant value to each individual instance. Removing data could result in the loss of critical and potentially valuable information, while generating synthetic data for minority classes could introduce biases and variations that do not reflect the reality of the academic environment studied.

During these experiments, a total of five predictive models were implemented: SVM, Random Forest, Logistic Regression, KNN, and Decision Tree. Each of these models was evaluated on various data configurations, which included full datasets, datasets only up to the third year, datasets only up to the third year combining 2 states, and another dataset only up to the fourth year of studies. At the end of the experiments, the selected approach was the use of data up to the third year with the combination of academic status categories 2 and 5 into one. This setup was chosen because it demonstrated a better balance in prediction performance across different majors and models.

The experiments show that for the Computer Science major, the KNN model using data up to the third year combined with academic status categories 2 and 5 into one provided the best performance metrics. In Electrical, the Decision Tree particularly stood out, showing high precision and accuracy. For Electronics, Logistic Regression was the most effective. In the Civil Engineering major, the Decision Tree demonstrated consistent superiority. As shown in Table 4 , the final models and data configurations reflect an approach aimed at optimizing accuracy and

Table 4. Summary of the Best Model by Major, with the selected approach.

Major	Winning Model	Key Metrics	Reason for Choice
Computer Science	KNN	Accuracy: 0.896, Precision: 0.896	High performance and simplicity.
Electrical	DT	Accuracy: 0.980, Precision: 0.981	Excellent precision, easy interpretation of results.
Electronics	RL	Accuracy: 0.888, Precision: 0.898	High precision, provides clear insights into variables.
Civil	DT	Accuracy: 0.968, Precision: 0.970	Consistently high performance and good comprehensibility.

applicability in realistic scenarios where not all student data is immediately available. The decision to combine states 2 and 5 proved to be critical in enhancing prediction clarity and reducing confusion rates, suggesting that grouping similar categories can be an effective strategy in contexts where differences between states are minimal but critically important. With this final configuration, the models significantly improved their ability to accurately predict academic outcomes across the four majors studied. This improvement was particularly notable in the context of more realistic and practical predictions, suitable for decision-making in educational environments where early interventions may be necessary. This approach underscores the importance of tailoring machine learning models to the peculiarities of the data and the specific needs of the application environment to achieve optimal results.

5 Conclusions

A model for early warning of student dropout was built for the Faculty of Science and Technology, using estimates based on relevant academic factors extracted from the faculty's academic database.

In the conducted study, an effective characterization of the academic database was achieved using advanced data science techniques. Initially, the elbow method was used to determine the optimal number of clusters, identifying four distinct groups. The student population was segmented into graduates, early dropouts (students who left before five years of study), late dropouts (those who left after five years), and students who completed the curriculum but still need to present their final project. This detailed analysis allowed for a better understanding of the academic distribution.

Machine learning predictive models were adjusted and evaluated through different data configurations in a series of training and prediction experiments. The most effective method was the third experiment, which combined data from students in states 2 and 5 up to the third year. This combination created a more homogeneous and representative dataset of academic success, allowing the models to more accurately identify key patterns and factors that predict successful academic outcomes.

With the selection of the third experiment, for the different majors, the optimal models varied: for Computer Science, the best model turned out to be K-Nearest Neighbors (KNN); for Electricity and Civil, the Decision Tree (DT) model was the most effective, however the one that had the lowest performance in the two races was KNN, and for Electronics, the Logistic Regression (RL) and K-Nearest Neighbors (KNN)) demonstrated better performance, unlike the Decision Tree Model (DT) demonstrated lower accuracy, compared to the RF and SVM models that demonstrated lower precision.

Acknowledgements.

The authors thanks to the Faculty of Sciences and Technologies and the Rectorate of National University of Caaguazú for providing us with the resources and the conducive environment to carry out this project.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] Pablo Díaz, Alexis Tejedor De Leó. El cadesun. un nuevo instrumento para analizar la deserción estudiantil universitaria. [Online]. Available: https://www.redalyc.org/journal/340/34056722012/html/, 2017.

[2] Latitud25. Alta deserción universitaria: hablemos de lo complejo que es estudiar en paraguay. https://enlatitud25.com/news/alta-desercion-universitaria-hablemos-de-lo-complejo-que-es-estudiar-en-%paraguay/. [Online; accessed 2-February-2023], 2023.

[3] C. D. FCyT. Plan de desarrollo ingeniería en electricidad. http://www.fctunca.edu.py/application/files/6814/9092/2147/Plan_de_Desarrollo_Ingenieria_en_Electricidad_1.pdf. [Online; accessed 26-October-2016], 2016.

[4] S. V. Orea, A. S. Vargas, and M. G. Alonso. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Ene*, vol. 779, n. 73, pp. 33, 2005.