A Predictive Model for Academic Performance in Engineering Students

Ms. Cristian Saavedra-Acuna, Universidad Andres Bello, Concepcion, Chile

Cristian Saavedra is an assistant professor at the School of Engineering at the University Andres Bello in Concepcion, Chile. He holds a bachelor's degree in Electronics Engineering and a master's degree in Technological Innovation and Entrepreneurship. Cristian is certified in Industrial Engineering, University Teaching, Online Hybrid and Blended Education, and Entrepreneurship Educators. He teaches industrial engineering students and carries out academic management activities. His main research interest areas are Innovation, entrepreneurship, engineering education, gender perspective studies in STEM education, and data analysis and visualization.

Dr. Monica Quezada-Espinoza, Universidad Andres Bello, Santiago, Chile

Monica Quezada-Espinoza is a professor and researcher at the School of Engineering at Universidad Andrés Bello in Santiago, Chile, where she collaborates with the Educational and Academic Innovation Unit (UNIDA) as an instructor in active learning methodologies and mentors engineering faculty in educational research. She is the Secretary of the Women in Engineering Division (WIED) of the American Society for Engineering Education (ASEE) and an associate researcher in the STEM Latin America Network, specifically in the STEM + Gender group. Her research interests are diverse and focus on university education in STEM fields, faculty and professional development, research-based methodologies, and the use of evaluation tools and technology for education. She is also passionate about investigating conceptual learning in abstract physics topics, developing strategies to improve the retention of first-year engineering students, and enhancing skills and competencies in higher education. Additionally, Monica is dedicated to exploring gender issues in STEM education, with a particular emphasis on studying and proposing improvements for the inclusion of women in highly male-dominated fields. For more information on her work, visit her ORCID profile.

Ms. Danilo Alberto Gomez, Universidad Andres Bello, Concepcion, Chile

Danilo Gómez is an assistant professor at the School of Engineering at the Andrés Bello University in Concepción, Chile. He has a Master's degree in applied statistics and Industrial engineering. In addition, Danilo has certifications in data science, machine learning, and big data. In his role as a teacher, Danilo specializes in teaching industrial engineering students and carries out academic management activities. His main research areas can be reviewed at: https://orcid.org/0000-0002-8735-7832

A Predictive Model for Academic Performance in Engineering Students

Abstract

This research article proposes developing a predictive model to identify, at an early stage, students at risk of low academic performance. Academic performance is a critical indicator in higher education, essential for student and professional success, and has been extensively studied due to its impact on retention and timely graduation. Socio-demographic factors such as socioeconomic status, family environment, work responsibilities, study habits, financial support, and psychological factors have been shown to influence student performance and attrition rates significantly. While progress has been made using linear regression models, recent years have seen the incorporation of advanced artificial intelligence techniques, offering new opportunities to enhance academic management. The objective of this article is to design a predictive model based on the entry profile of engineering students to assess their risk of low academic performance. The study employs a non-experimental quantitative methodology and machine learning techniques within a Knowledge Discovery in Databases (KDD) framework. The data used in the model includes Weighted Average Grades and socio-demographic factors from the characterization survey that students complete upon entering the university. The sample comprises 1,266 students from the Faculty of Engineering at a private university in Chile who enrolled in the first semester of 2022. Their academic performance is analyzed from that semester until the first semester of 2024, covering five semesters and reaching 50% of their curricular progression. Additionally, socio-demographic data such as family, economic, and work backgrounds, collected in the characterization survey conducted at the time of their entry in 2022, are utilized. The results of this research are expected to identify key factors affecting academic performance, such as the number of working hours, study methodologies, and the source of financing for their studies. The developed model is anticipated to classify academic outcomes into four performance levels: no mastery, insufficient mastery, satisfactory mastery, and outstanding mastery, with 98% accuracy and a 3% margin of error. This predictive model aims to contribute to academic management by facilitating the early implementation of support measures or programs for students at academic risk. Moreover, institutions can adjust curricular design and teaching methods by analyzing the factors influencing academic performance. The actions derived from this model are expected to improve students' academic performance, potentially reducing dropout rates and increasing timely graduation rates, thereby inspiring and motivating educators and policymakers in engineering education.

Keywords: Data Science, Academic Performance, Predictive Model, Machine Learning, Student Retention

Introduction

The transition to university life marks a critical point in students' academic trajectories, with the first two years being particularly decisive, as this period sees the highest dropout rates [1-2]. This phenomenon has significant implications at multiple levels: it impacts institutional accreditation processes, educational management, and public policies, while also posing economic and emotional challenges for the families involved [3-4].

Several factors contribute to dropout during this stage, including difficulties adapting to the university environment and the high academic demands of higher education [1-3, 5]. These

challenges can lead to frustration and demotivation, thereby increasing the likelihood of student withdrawal [4]. The effects of dropout are not limited to individuals; educational institutions experience declines in quality, reputation, and performance on key indicators, while families and governments face far-reaching economic and social consequences [3-4].

In response, higher education institutions have implemented measures such as adaptation programs, leveling courses, and personalized tutoring to strengthen students' academic competencies and reduce risk factors associated with dropout [1-3]. A crucial aspect of these strategies is early monitoring of academic performance, as low performance is directly associated with elevated dropout rates [1-2]. By identifying at-risk students early, institutions can implement targeted interventions that promote integration and retention within the educational system [3-4].

Background

Academic dropout and predictive models have garnered significant interest in educational research due to their impact on students, institutions, and society. This section reviews both topics, integrating key concepts and evidence reported in the literature.
Academic dropout is a complex issue with economic, social, and educational repercussions. It not only affects individuals' quality of life by limiting access to better jobs and opportunities but also has implications for social cohesion and economic growth [6-7]. Commonly defined as the premature abandonment of studies, dropout hinders progression to higher levels of education. This process can manifest in various ways, including low academic performance, academic delay, failure, and permanent withdrawal [8-9].

Several factors contribute to academic dropout, including personal, institutional, and socioeconomic aspects. Personal factors encompass psychological characteristics, motivation, academic and family background, age at enrollment, marital status, and financial situation [10]. From an institutional perspective, educational quality, infrastructure, and student support play important roles, while access to scholarships, loans, and the type of prior schooling are key determinants on the socioeconomic front. Dropout is often more pronounced in science, technology, engineering, and mathematics (STEM) programs, which have significantly higher rates of attrition compared to other disciplines [7, 11]. Other studies emphasize the importance of tutoring strategies based on data mining to mitigate dropout in these contexts [12].

Dropout is a multifaceted phenomenon recognized as a priority in educational policies at both institutional and governmental levels [7, 13]. To address this challenge, research has focused on identifying predictive factors to design effective interventions to prevent student attrition. Other researchers have explored data mining techniques to predict student dropout in higher education institutions, highlighting their ability to generate accurate and adaptable predictive models [14].

Predictive models have emerged as valuable tools for detecting at-risk students and facilitating timely interventions. These tools, based on educational data mining (EDM) techniques, enable the analysis of large datasets to uncover patterns associated with academic performance and dropout [11, 15-16]. Other researchers propose deep learning-based models, such as the IC-BTCN framework, specifically designed to predict dropout in massive open online courses (MOOCs), demonstrating their effectiveness in the early identification of at-risk students [17].

Predictive models operate through various methodologies, such as classification, regression, and clustering. For instance, algorithms like decision trees, neural networks, and support vector machines are widely used to classify students into academic risk categories or predict continuous variables like GPA [6-7]. These models consider a wide range of variables, such as academic data (e.g., prior grades, accumulated credits), socio-demographic information (e.g., age, gender, type of prior schooling), and participation data from virtual platforms [10].

For example, Vives et al. [18] employed long short-term memory (LSTM) neural networks to predict academic performance in programming courses, showcasing this approach's ability to capture complex sequential patterns in student data. Additionally, the stages involved in building these models include data extraction and selection, algorithm application, precision evaluation, and practical implementation within educational institutions [7, 9]. Using specialized tools such as WEKA, RapidMiner, and KEEL has facilitated the application of EDM techniques, allowing educational institutions to identify at-risk students and plan targeted interventions [7, 13]. Model evaluation is conducted using methods like cross-validation and analysis of the area under the ROC curve, ensuring precision and generalizability.

In summary, academic dropout is a multifactorial challenge that requires comprehensive, evidence-based approaches for effective intervention. Predictive models, supported by technological and methodological advancements, represent key tools for anticipating risks, improving academic performance, and reducing dropout rates, thereby enhancing the quality and equity of higher education.

Research Questions

RQ1: Can machine learning algorithms effectively predict academic performance based on socio-demographic data?
RQ2: What methods and algorithms are applicable for predicting academic performance?
RQ3: What socio-demographic factors most determine a student's academic performance?

This research aims to establish the foundation for designing and developing predictive models that enable the early identification of socio-demographic and academic factors with the greatest impact on student performance upon entering the Faculty of Engineering. Implementing these models aims to detect students at higher risk of dropout and understand their specific needs. This will allow the implementation of personalized support strategies, which may include financial aid, flexible work schedules, study methodology reinforcement activities, or academic and career guidance programs. By anticipating potential causes of dropout, institutions can strengthen student retention, enhancing both the educational experience and institutional indicators of quality and academic success.

Methodology

This study employs Machine Learning tools and applies a Knowledge Discovery in Databases (KDD) methodology tailored to the higher education context. The process is structured into four stages, as outlined in Table 1.

Table 1. Stages of the methodology for data analysis, model creation, and interpretation.

Stage 1. Data Preprocessing	A dataset containing 861 records and 36 columns corresponding to students who enrolled in engineering programs in 2022 was used. The data was extracted from the institutional characterization survey administered to all prospective students. Data cleaning processes were applied to address outliers and missing values, resulting in a final dataset of 823 records.
Stage 2. Data Transformation	The variables were transformed, specifically standardizing the independent variables. The column names were also revised to facilitate analysis and modeling.
Stage 3. Data Selection	The variables were selected based on the significance of each independent variable concerning the dependent variables. This process reduced the dataset to 18 variables from the initial 36.
Stage 4. Data Modeling	Predictive models were developed using various algorithms, including linear regression, K-Neighbors Regressor, AdaBoost Regressor, Random Forest Regressor, and Gradient Boosting Regressor. Model evaluation was conducted through cross-validation and performance metrics.
Stage 5. Interpretation	To assess the quality of each model, performance metrics such as R^2, MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error) were employed.

This research focuses on the student population from the Faculty of Engineering at a private university in Chile. Data was drawn from the characterization survey completed by 1,266 new students enrolled during the first semester of 2022. These students' academic performance was monitored from the first semester of 2022 through the first semester of 2024. By the end of this period, 862 students remained active. The survey mentioned in Stage 1 of Table 1 included four dimensions:

i) family and social factors,
ii) economic factors,
iii) prior educational experience, and
iv) personal skills and study habits.

Additionally, the study considered the Weighted Academic Average (PPA) for the first, second, and third semesters, which, in this context, ranges between 1 and 7. Student academic performance tended to decline, particularly during the second year. Specifically, the PPA decreased by 7.7% and 7.0% compared to the first semester of 2022.

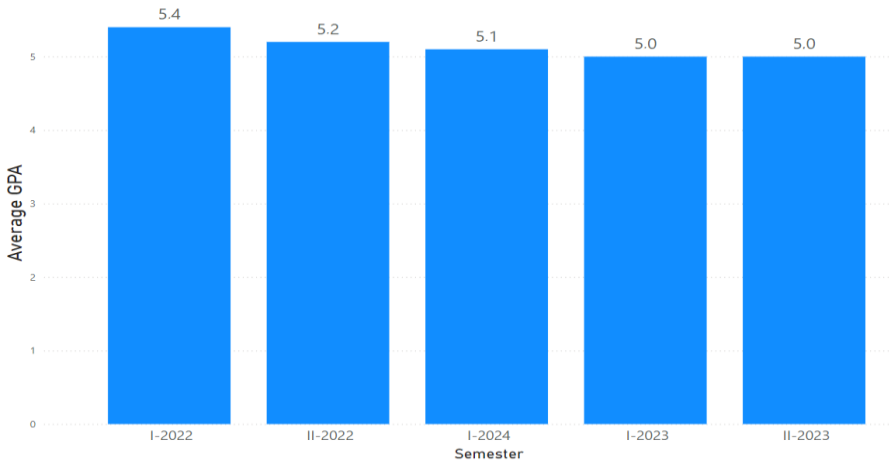

Figure 1. Average GPA Trends by Semester (2022–2023).

Similarly, it can be observed in Figure 2 that the segment of students with lower GPA averages (GPA range of 1 to 4 and 4 to 5) increases as students' progress in their curriculum, while the excellence segment (GPA range of 6 to 7) systematically decreases, reaching 63.9% by the second semester of 2023.

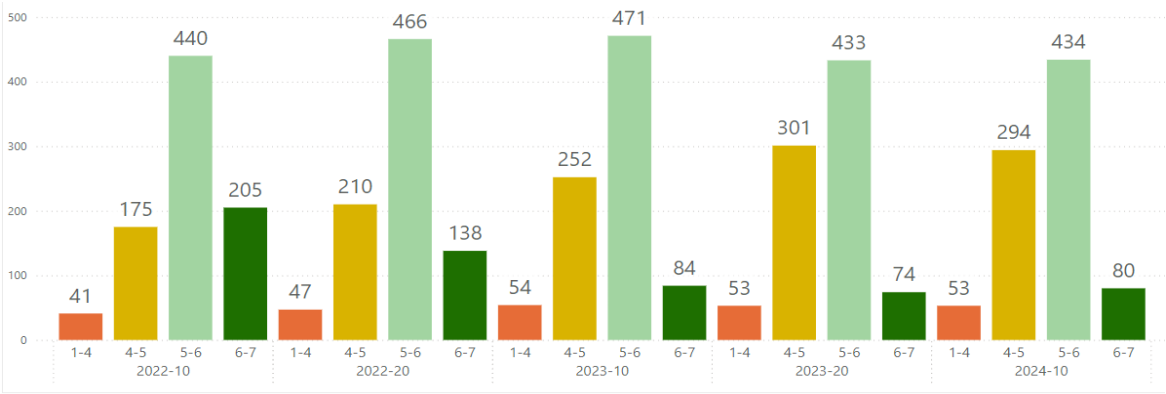

Figure 2. Distribution of Students by GPA Range and Semester (2022–2024).

The period of lowest academic performance for students corresponds to the first and second semesters of 2023. Therefore, the student's GPA from 2023 will be used as a reference for the predictive model.

This study utilized the PyCaret library as the primary environment to implement and compare various regression models, assessing their ability to predict the academic performance of university students. The dataset, comprising 862 active students as of the first semester of 2024, underwent rigorous preprocessing. This included the normalization and transformation of 36 predictive variables (detailed in Appendix A) to ensure data quality and homogeneity before integrating them into the predictive models.

The models selected for evaluation, Gradient Boosting Regressor (GBR), Random Forest Regressor (RF), AdaBoost Regressor (ADA), K-Neighbors Regressor (KNN), and Linear Regression (LR), were chosen for their flexibility in capturing non-linear relationships and their adaptability to various data patterns. The methodology involved an initial split of the data into training (80%) and testing (20%) sets, along with a 10-fold cross-validation scheme to ensure stability and representativeness of the results. Subsequently, hyperparameter optimization was performed using Grid Search, refining the performance of the most promising model, the Gradient Boosting Regressor.

Performance analysis was conducted using robust metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R^2 (Coefficient of Determination), and MAPE (Mean Absolute Percentage Error). The results achieved by the models are presented in Table 2.

Table 2. Performance metrics of regression models for predicting academic performance.

Modelo	**MAE**	**MSE**	**RMSE**	**R^2**	**MAPE**
Gradient Boosting Regressor	0.2830	0.1350	0.3656	0.6109	5.59%
Random Forest Regressor	0.2925	0.1437	0.3772	0.5850	5.78%
AdaBoost Regressor	0.2983	0.1507	0.3868	0.5674	5.91%
K Neighbors Regressor	0.3214	0.1697	0.4105	0.5126	6.37%
Linear Regression	0.3185	0.1755	0.4177	0.4962	6.27%

Based on the results presented in Table 2, the Gradient Boosting Regressor (GBR) model excels in accuracy and provides a robust analytical framework for identifying the factors that most influence academic performance. In this regard, it is essential to delve deeper into the analysis of the predictive variables that contribute to the model's performance, as these offer key insights into the areas with the greatest impact on the 2023 GPA average.

In Figure 3, the ranking of the importance of predictive variables based on the GBR model is presented. This analysis provides a clear visualization of the variables that carry the most significant weight in the predictions, offering a starting point for intervention and improvement strategies. The results demonstrate that the Gradient Boosting Regressor (GBR) model consistently outperforms the others across all key metrics, particularly in terms of MAE, MSE, and R^2. This highlights its effectiveness as a predictive tool for evaluating student academic performance, enabling the identification of relevant patterns and key variables that impact the 2023 GPA average.

The importance of predictive variables in the GBR model, as shown in Figure 3, highlights the dominant role of specific variables, such as x1 and x7, in predicting academic performance. This ranking of importance provides a solid foundation for designing personalized strategies to enhance academic success by focusing on the most influential factors.

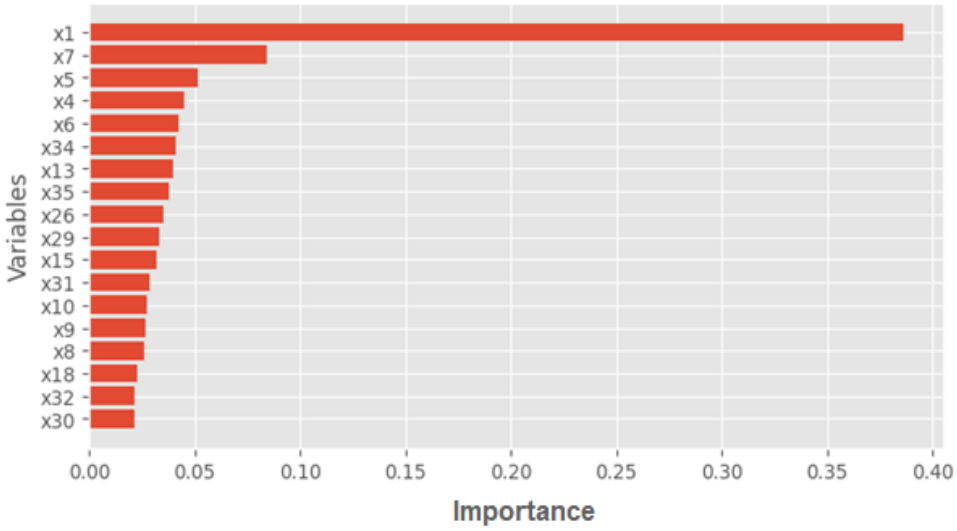

Figure 3. Ranking of variable importance in the Gradient Boosting Regressor Model.

In conclusion, the findings validate the utility of the Gradient Boosting Regressor for predictive analysis and emphasize the importance of integrating advanced machine learning techniques into educational management. These tools enable institutions to identify at-risk students early, optimize decision-making, and implement more effective interventions to improve academic outcomes and overall educational experience.

Discussion

This study aims to establish the foundation for designing and developing predictive models capable of identifying, at an early stage, the socio-demographic and academic factors that significantly impact student performance upon entering the School of Engineering. To achieve this, a methodology based on the Knowledge Discovery in Databases (KDD) process

was implemented and adapted to the context of Higher Education. This methodology included stages of data preprocessing, transformation, selection, and modeling, utilizing advanced Machine Learning tools to ensure the quality and relevance of the developed models. The integration of these models seeks to detect groups of students at higher risk of attrition and also to generate personalized support strategies that address their specific needs and promote their retention and academic success.

In this research, five artificial intelligence models were evaluated to predict student academic performance, among which the Gradient Boosting Regressor and Random Forest Regressor demonstrated superior performance. These results align with previous findings reported in the literature. For instance, Urbina et al. [7] identified Random Forest as one of the most accurate algorithms for predicting attrition, achieving a Matthews Correlation Coefficient of 87.43% and an accuracy of 94.34%, demonstrating its robustness in educational scenarios [7]. Similarly, the Gradient Boosting Regressor has been recognized for its ability to handle complex non-linear relationships and its effectiveness in similar contexts, as highlighted in the work of Thomas [13], which employed advanced Machine Learning techniques to analyze predictive factors of academic performance (Thomas & Celis) [13, 19].

These two tree-based methods were selected for their ability to manage moderately noisy datasets while maintaining interpretability and efficiency. In contrast, deep neural networks were excluded due to their limited dataset size and lower interpretability, which can hinder decision-making in educational settings. Simpler models, such as K-Neighbors Regressor and Linear Regression, were also tested, but they yielded lower accuracy and predictive performance based on MAE and R^2 metrics.

On the other hand, models based on K-Neighbors Regressor and Linear Regression showed inferior performance in this study, a finding also supported by the literature. Fernández-García [6] and Cruz Castro [11] noted that these models are generally less effective in capturing non-linear relationships and interactions among multiple predictive variables, limiting their applicability in educational contexts where data are inherently complex and multivariate [6, 13]. Furthermore, the KDD-based methodology used in this study emphasizes the importance of proper selection and transformation of predictive variables, a point also highlighted in prior studies, such as Alruwais [20], who underscored the need for data preprocessing to ensure more accurate and generalizable predictive models.

These findings reinforce the utility of decision-tree-based models, such as Random Forest and Gradient Boosting, and highlight the limitations of simpler methods. They justify the need for advanced approaches to improve accuracy and interpretability in predicting student academic performance. Notably, the Gradient Boosting Regressor (GBR) model achieved the best results in terms of MAE, MSE, and R^2, proving to be the most suitable model for this dataset.

The main variables significantly impacting the academic prediction model include the number of approved subjects (X1), schedule (X7), program (X5), campus (X4), major (X6), source of funding (X34), first-time enrollment in higher education (X13), and study financing source (X35). Below, we discuss the impact of these variables:

- X1 (Number of Approved Subjects): This factor has the highest impact on the model, demonstrating a direct relationship between the number of approved subjects and strong academic performance.

- Schedule, Program, and Major: Table 3 summarizes the GPA results and the percentage of students with a GPA below 4.0.

This analysis underscores the critical role of advanced predictive models and key variables in identifying areas for intervention to enhance student outcomes and academic success.

Table 3. Average GPA and percentage of students with GPA below 4.0 by major and academic schedule.

Program	N° Students	Shift	Average GPA 2023	% Students with GPA < 4.0
Automation and Robotics Engineering	6	Evening shift	5.72	0.0 %
Automation and Robotics Engineering	59	Daytime shift	5.04	8.5 %
Civil Engineering	25	Daytime shift	5.18	0.0 %
Computer Civil Engineering	155	Daytime shift	5.08	6.5 %
Computer Engineering	17	Evening shift	5.57	0.0 %
Computer Engineering	134	Daytime shift	5.11	6.0 %
Construction Engineering	25	Daytime shift	4.76	12.0 %
Geology	98	Daytime shift	4.88	7.1 %
Industrial Civil Engineering	20	Evening shift	5.01	15.0 %
Industrial Civil Engineering	139	Daytime shift	4.99	6.5 %
Industrial Engineering	7	Daytime shift	4.72	14.3 %
Industrial Engineering	13	Evening shift	5.14	7.7 %
Logistics and Transportation Engineering	13	Evening shift	5.60	0.0 %
Merchant Marine Engineering	93	Daytime shift	4.69	18.3 %
Mines Civil Engineering	57	Daytime shift	5.04	3.5 %

The majors with the lowest average GPA and the highest percentage of students with a GPA below 4.0 are Construction Engineering, Industrial Engineering, and Merchant Marine Engineering, all of which correspond to the daytime schedule. In contrast, students enrolled in evening programs demonstrate higher average GPAs, with most maintaining a GPA above 4.0, except for those in the Industrial Civil Engineering program.

When conducting a global analysis by academic schedule, evening program students exhibit better academic performance, with an average GPA of 5.41 and a percentage of students with a GPA below 4.0 of 4.54%. In contrast, daytime program students have an average GPA of 4.95 and a percentage of students with a GPA below 4.0 of 8.27%. Regarding the campus factor (X4), it is observed that students from Campus 1 and Campus 2 achieve higher average GPAs and have a lower percentage of students with a GPA below 4.0. Conversely, students from Campus 3 exhibit the lowest academic performance.

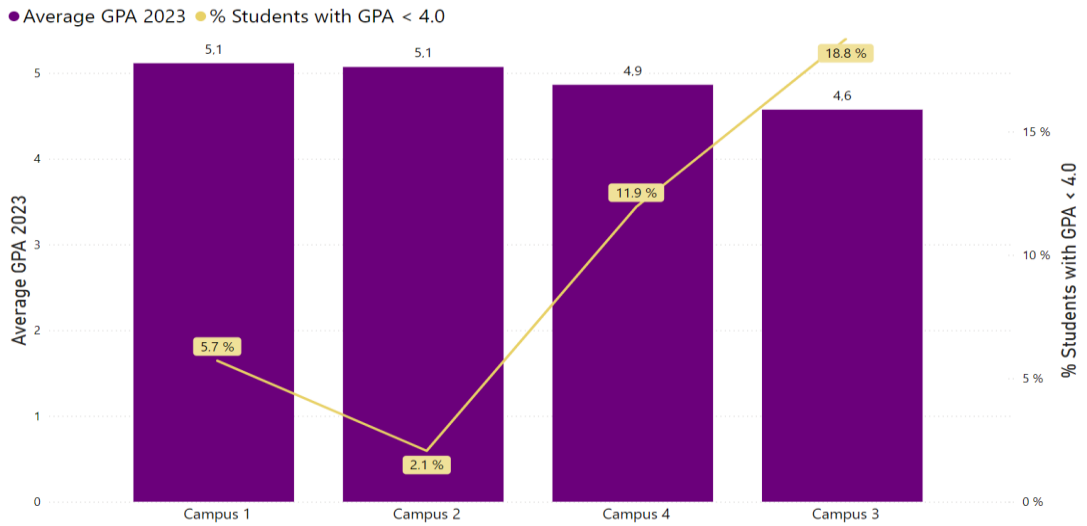

Figure 4. Average GPA by Campus and Percentage of Students with GPA < 4.0.

Regarding economic factors, specifically the Source of Funding (X34) and the Person Responsible for Financing Studies (X35), Table 4 identifies that the segment with the lowest academic performance corresponds to students who finance their studies with personal resources and share the responsibility of financing with other members of their household. Conversely, the segment with the best academic performance includes students solely responsible for financing their studies through State-Guaranteed Loans/Scholarships or their own resources.

Table 4. Average GPA in relation to sources of funding and who finances the studies.

	Source of financing / Who will finance the studies?	State Credit/Scholarships/Others	Mixed	Own resources
GPA	My Parents/	4.89	5.02	5.11
	Mixed (me/others)	4.91	4.93	4.08
	Me	5.22	4.85	5.29
% Students with GPA < 4.0	My Parents/	8.0%	8.2%	5.5%
	Mixed (me/others)	9.8%	13.3%	28.6%
	Me	0.0%	17.4%	5.8%

Finally, regarding the factor of whether it is the student's first time entering higher education (X13), it is identified that students who have graduated from a previous program, either from the same institution or another, represent the segment with the highest academic performance. Conversely, the segment with the lowest academic performance consists of students who have dropped out of more than one program previously.

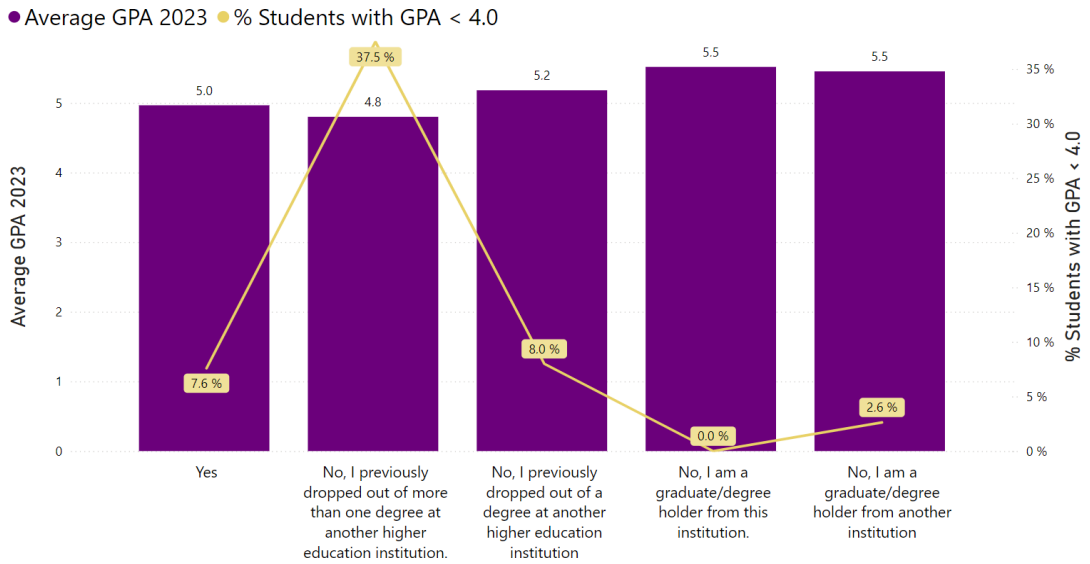

Figure 5. Average GPA by previous higher education experience.

In previous research, various factors influencing student academic performance have been identified. For instance, motivation has been recognized as a key component in predicting

performance, as highlighted by Cruz [11] in her analysis of university students. Similarly, Naghi et al. [21] emphasized the role of proactive personality and critical thinking skills in improving academic performance, while Verma et al. [22] underscored the influence of socio-demographic factors, such as gender and parental education level. Complementing these findings, Sabanal et al. [23] developed a predictive model for higher mathematics performance based on psychological factors, including learning engagement, motivation, and self-efficacy, and instructional factors such as teacher-related elements and the learning environment. Their model explained over 82% of the variance in performance, highlighting the relevance of these variables in designing targeted academic interventions in higher education. Additionally, Karim-Abdallah et al [24], through a systematic literature review of 60 studies, identified decision trees, random forests, and artificial neural networks as the most applied algorithms in predicting academic performance. His findings also emphasized the importance of incorporating diverse data sources, including academic records, demographic characteristics, and behavioral indicators, to build effective and inclusive predictive models that support early interventions and personalized learning in higher education.

In contrast, this study identified additional factors within the framework of the student characterization profile that significantly impact academic performance. Among these, evening program students exhibited better academic performance compared to daytime students. However, specific at-risk segments were also detected in both modalities. In the daytime program, students in Construction Engineering, Industrial Engineering, and Merchant Marine Engineering demonstrated lower academic performance, whereas in the evening program, the at-risk segment was concentrated among Industrial Civil Engineering students. These findings not only expand the understanding of factors affecting academic performance but also underscore the importance of considering student profile particularities when designing academic support strategies and intervention programs.

When comparing the results of this study with prior research focused on dropout prediction models [16], cross-cutting factors emerge in both models, such as economic factors highlighted by Thomas et al. [13]. In the present study, the highest-risk segment includes students who finance their studies with personal resources, sharing the financial responsibility with one or more additional individuals. On the other hand, the best-performing segment comprises students who also finance their studies with personal resources but assume 100% of the responsibility for their funding.

The experience of having pursued a previous degree shows a significant influence on academic performance. Students who completed a prior degree or dropped out only once tend to perform better academically, possibly due to acquired skills or greater clarity in their educational goals. This finding aligns with studies like Urbina-Nájera [7], who identified that prior academic experiences can positively influence performance, provided they are not associated with repeated patterns of dropout [7].

Conversely, students who have abandoned multiple programs exhibit low academic performance, which could be explained by factors such as demotivation or persistent difficulties. This aligns with findings by Fernández García [6] and Alruwais [20], who highlighted the negative influence of multiple dropouts on academic performance due to their impact on students' self-perception and confidence [6, 11].

For first-time students entering their initial degree program, the cumulative weighted average (GPA) is 5.0, with 7.6% of these students having a GPA below 4.0, placing them in an academic risk category. This aligns with the analysis by Cruz Castro [11], who emphasized the importance of monitoring first-year students to identify and mitigate early risk factors associated with low academic performance [12]. These findings highlight the need to design personalized strategies considering students' educational history, including those with prior dropout experiences, to provide more effective support and improve their academic performance.

Conclusions

This study evaluated five regression and decision tree models to predict the academic performance of students who enrolled in the first semester of 2022 and remained active through the end of their fifth semester. The dataset was derived from a characterization survey administered at the beginning of their studies, along with information related to the selected academic program. The findings confirm the feasibility of predicting academic performance during the early stages of university education. Among the models assessed, the Gradient Boosting Regressor (GBR) demonstrated the best performance, achieving strong metrics—MAE (0.283), MSE (0.135), and R^2 (0.6109). This model proved particularly effective in capturing complex patterns and non-linear relationships within the data. Using the characterization survey, this predictive model enables the generation of academic performance profiles and the segmentation of students based on risk levels. This provides valuable insights into identifying vulnerable student groups, such as those associated with specific program types or class schedules, and supports evidence-based decision-making.

A key takeaway from these findings is the importance of using predictive models not solely for risk classification, but also as a foundation for designing tailored interventions to improve academic outcomes. Early identification of at-risk students during the first two years of an engineering program represents a powerful strategy for institutional academic management. Currently, the institution assigns risk status to first-year students based exclusively on their university entrance scores. By incorporating a multivariable predictive model, institutions can more accurately identify students at academic risk and develop a system that enhances dropout prediction. This, in turn, would allow for a more effective allocation of resources toward leveling courses, personalized tutoring, and academic support services, ultimately improving graduation timelines and student retention. From the educators' perspective, early access to student profiles based on expected academic performance can enable faculty to understand their student cohorts' composition better. This information allows instructors to adapt their teaching strategies and methodologies to better suit the needs and potential of their students.

Limitations

The limitations of this study are primarily methodological. First, the data analyzed were collected exclusively from one university and a specific subset of engineering programs, which may limit the generalizability of the findings to other institutions or academic disciplines. Additionally, the predictive models were based on data collected at a single point in time, at the start of the students' studies, which restricts their ability to capture changes in students' characteristics and circumstances over time. Finally, external factors such as dynamic socioeconomic conditions, family contexts, and institutional policies were not

included in the analysis, despite their potential significant impact on students' academic performance.

Future work

The findings of this study provide a foundation for enhancing predictive modeling in higher education. Building on these results, future efforts can explore strategies to develop more accurate and efficient algorithms in the short term. For instance, combining ensemble learning techniques with robust feature selection processes may help improve model performance without unnecessarily increasing complexity. Additionally, incorporating dynamic information, such as the evolution of student performance across semesters, could lead to models more reflective of academic realities and allow for more timely and targeted interventions for at-risk students. This would enable informed decision-making and foster more effective, personalized academic support.

This research has demonstrated the feasibility of designing a predictive model based on the characterization survey administered to students upon university entry. As a next step, it is critical to strengthen and expand the impact of this work by replicating the methodology across different academic disciplines and institutions with diverse demographic profiles. Doing so would allow for evaluating the generalizability and robustness of the models in varied contexts. Moreover, incorporating data collected at multiple points throughout the academic cycle will support the development of more adaptive models that can respond to changes over time. Including external variables, such as family economic conditions, access to technological resources, and involvement in extracurricular activities, may significantly enhance the models' explanatory power.

Lastly, designing and evaluating pilot programs grounded in predictive insights is essential. These initiatives should enable the implementation of measurable, real-time, and adaptable academic support strategies. Exploring advanced techniques, such as deep neural networks or hybrid approaches, can further improve model precision and enrich the understanding of complex patterns in educational data.

Acknowledgments

The authors gratefully acknowledge the leadership and financial support of the School of Engineering at the Universidad Andres Bello, Chile. We also thank the Educational Research and Academic Development Unit (UNIDA) for its mentorship and guidance in developing research skills for higher education faculty.

References

[1] M. O. González-Morales, D. López-Aguilar, P. R. Álvarez-Pérez, and P. A. Toledo-Delgado, "Dropping out of higher education: Analysis of variables that characterise students who interrupt their studies," Acta Psychologica, vol. 252, p. 104669, Feb. 2025, doi: 10.1016/j.actpsy.2024.104669.

[2] Cruz L., Li T., Ciner L., Douglas K., Greg C., (2022) Predicting learning outcome in a first-year engineering course: a human-centered learning analytics approach. Recuperado de: https://peer.asee.org/predicting-learning-outcome-in-a-first-year-engineering-course-a-human-centered-learning-analytics-approach.pdf

[3] G. Bilquise, S. Abdallah, and T. Kobbaey, "Predicting Student Retention Among a Homogeneous Population Using Data Mining," in Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019 (AISI 2019), A. Hassanien, K. Shaalan, and M. Tolba, Eds., Advances in Intelligent Systems and Computing, vol. 1058. Cham, Switzerland: Springer, 2020, pp. 31–41. doi: 10.1007/978-3-030-31129-2_4.

[4] O. Aquines Gutiérrez, D. M. Hernández Taylor, A. Santos-Guevara, W. X. Chavarría-Garza, H. Martínez-Huerta, and R. K. Galloway, "How the Entry Profiles and Early Study Habits Are Related to First-Year Academic Performance in Engineering Programs," *Sustainability*, vol. 14, no. 15400, pp. 1-19, Nov. 2022. DOI: 10.3390/su142215400.

[5] I. Alcauter, L. Martinez-Villaseñor, y H. Ponce, "Explaining Factors of Student Attrition at Higher Education", *CyS*, vol. 27, n.º 4, dic. 2023, doi: 10.13053/cys-27-4-4776. Available in: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/4776

[6] A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sánchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEE Access*, vol. 9, pp. 133076–133094, 2021. DOI: 10.1109/ACCESS.2021.3115851.

[7] A. B. Urbina-Nájera and L. A. Méndez-Ortega, "Predictive Model for Taking Decision to Prevent University Dropout," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 205-216, Jan. 2022. DOI: 10.9781/ijimai.2022.01.006.

[8] N. M. Alruwais, "Deep FM-Based Predictive Model for Student Dropout in Online Classes", *IEEE Access*, vol. 11, pp. 96954-96970, 2023, doi: 10.1109/ACCESS.2023.3312150. Disponible en: https://ieeexplore.ieee.org/document/10239344/.

[9] A. Jiménez-Macías, P. M. Moreno-Marcos, P. J. Muñoz-Merino, M. Ortiz-Rojas, and C. Delgado Kloos, "Analyzing Feature Importance for a Predictive Undergraduate Student Dropout Model," *Lecture Notes in Computer Science*, vol. 20, no. 1, pp. 175–194, 2023. DOI: 10.2298/CSIS211110050J.

[10] G. Priya, S. Eliyas, and S. Kumar, "Detecting and Predicting Learner's Dropout Using KNN Algorithm," presented at the 2024 IEEE International Technology Conference (OTCON), 2024. DOI: 10.1109/OTCON60325.2024.10688123.

[11] L. M. Cruz Castro, T. Li, L. Ciner, K. A. Douglas, and C. G. Brinton, "Predicting Learning Outcome in a First-Year Engineering Course: A Human-Centered Learning Analytics Approach," presented at the ASEE Annual Conference & Exposition, 2022.

[12] C. Burgos, M. L. Campanario, D. D. L. Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Computers & Electrical Engineering*, vol. 66, pp. 541–556, Feb. 2018, doi: 10.1016/j.compeleceng.2017.03.005.

[13] P. B. Thomas, C. R. Bego, and A. D. Piemonte, "Predicting Student Retention via Expectancy Value Theory Using Data Gathered before the Semester Begins," presented at the ASEE Annual Conference & Exposition, 2023.

[14] W. F. Wan Yaacob, N. Mohd Sobri, S. A. M. Nasir, W. F. Wan Yaacob, N. D. Norshahidi, and W. Z. Wan Husin, "Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques," *Journal of Physics: Conference Series*, vol. 1496, p. 012005, Mar. 2020, doi: 10.1088/1742-6596/1496/1/012005.

[15] I. Uysal, P. E. Spector, C. Ferekides, M. B. Ayanoglu, and R. Elashmawy, "Predicting Academic Performance for Pre/Post-Intervention on Action-State Orientation Surveys," presented at the ASEE Annual Conference & Exposition, 2023.

[16] C. Saavedra-Acuna, M. Quezada-Espinoza, y D. Gomez, “Analyzing Attrition: Predictive Model of Dropout Causes among Engineering Students”, in *2024 ASEE Annual Conference & Exposition*, Portland, Oregon, jun. 2024. doi: 10.18260/1-2--46574. Available in: https://peer.asee.org/46574
[17] X. Zhang, X. Wang, J. Zhao, B. Zhang, and F. Zhang, "IC-BTCN: A Deep Learning Model for Dropout Prediction of MOOCs Students," *IEEE Transactions on Education*, vol. 67, no. 6, pp. 974–982, Dec. 2024, doi: 10.1109/TE.2024.3398771.
[18] L. Vives *et al*., "Prediction of Students’ Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks," *IEEE Access*, vol. 12, pp. 5882–5898, 2024, doi: 10.1109/ACCESS.2024.3350169.
[19] S. Celis, L. Moreno, P. Poblete, J. Villanueva, and R. Weber, "Un modelo analítico para la predicción del rendimiento académico de estudiantes de ingeniería," *Revista Ingeniería de Sistemas*, vol. XXIX, pp. 1-10, Sept. 2015.
[20] N. M. Alruwais, "Deep FM-Based Predictive Model for Student Dropout in Online Classes," *IEEE Access*, vol. 11, pp. 96954–96965, Sept. 2023. DOI: 10.1109/ACCESS.2023.3312150.
[21] M. Nagahi, R. Jaradat, S. Davarzani, M. Nagahisarchoghaei, y S. R. Goerger, “Academic Performance of Engineering Students”, in *ASEE Virtual Conference*, Virtual, jun. 2020. doi: 10.18260/1-2--34084. Disponible en: https://peer.asee.org/34084
[22] S. Verma y R. K. Yadav, “Effect of Different Attributes on the Academic Performance of Engineering Students”, in *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)*, Buldhana, India: IEEE, dic. 2020, pp. 1-4. doi: 10.1109/ICATMRI51801.2020.9398442. Disponible en: https://ieeexplore.ieee.org/document/9398442/.
[23] D. Sabanal, M. Gako, H. Dela Torre, J. Sabanal, R. Boi So, J. B. Bacal, . Dim Corgio, J. F. Laroga, C. Camallere (Jr.), M. J. Pagador, R. J. Barino, K. Mameng, M. Go, N. Goles “Predictive model for college students’ performance in higher mathematics”, *Social Sciences & Humanities Open*, vol. 10, p. 101134, jan. 2024, doi: 10.1016/j.ssaho.2024.101134. Available in: https://www.sciencedirect.com/science/article/pii/S2590291124003310.
[24] B. Karim-Abdallah, M. A. Junior, P. Appiahene, E. Harris, y D. K. Binful, “Application of Machine Learning Algorithms in Predicting Academic Performance of Students in Higher Education Institutes (HEIs): A Systematic Review and Bibliographic Analysis”, *African Journal of Applied Research*, vol. 11, n.o 1, pp. 536-559, ene. 2025, doi: 10.26437/ajar.v11i1.869. Available in: https://www.ajaronline.com/index.php/AJAR/article/view/869.

Appendix A. Variables associated with the dropout prediction model.

No.	Name	Description	Data Type
X1	Courses Approved	N°. of courses approved	Integer
X2	Credits Approved	N°. of credits approved	Integer
X3	Gender	Indicates the student's gender	Binary
X4	Campus	Specifies the campus location where the student is attending.	Nominal
X5	Program Code	A unique identifier for the academic program.	Nominal
X6	Program	Refers to the academic program in which the student is enrolled.	Nominal
X7	Shift	Evening shift or Daytime shift	Binary

X8	Living Arrangement	Specifies with whom the student will live upon starting their studies	Nominal
X9	Currently Working	Indicates whether the student is currently working before starting studies	Binary
X10	Will Work Upon Starting Studies	Indicates whether the student intends to work after starting studies	Binary
X11	Father's Educational Level	Indicates the highest level of education completed by the student's father	Nominal
X12	Mother's Educational Level	Indicates the highest level of education completed by the student's mother	Nominal
X13	First Entry to Higher Education	Indicates whether the student is enrolling in higher education for the first time	Binary
X14	Uses Study Techniques	Indicates whether the student uses any study techniques	Nominal
X15	Reason for Choosing Career	Indicates the main reason for choosing the career	Nominal
X16	Representation of the Career Path	Indicate what best represents you in the field of study	Nominal
X17	Skills	It indicates the skills that the student deems relevant	Nominal
X18	Teamwork	Indicates the level of importance the student assigns to teamwork	Nominal
X19	Leadership	Indicates the level of importance the student assigns to leadership	Nominal
X20	Effective Communication	Indicates the level of importance the student assigns to effective communication	Nominal
X21	Negotiation	Indicates the level of importance the student assigns to negotiation	Nominal
X22	Civic Education	Indicates the level of importance the student assigns to civic education	Nominal
X23	Innovation and Entrepreneurship	Indicates the level of importance the student assigns to innovation and entrepreneurship	Nominal
X24	Contact Level	Assess the level of contact with various close groups	Nominal
X25	Career	Name of the student's career	Nominal
X26	Extended Family	Indicates the level of contact the student maintains with aunts, uncles, cousins	Nominal
X27	Friends	Indicates the level of contact the student maintains with friends	Nominal
X28	Teachers	Indicates the level of contact the student maintains with teachers	Nominal
X29	Partner	Indicates the level of contact the student maintains with their partner	Nominal
X30	Marital Status	Indicates the student's marital status	Nominal
X31	Family Head	Indicates who assumes the role of head of family	Nominal
X32	Parental Status	Indicates whether the student has children	Nominal
X33	University Selection	This variable indicates a relevant factor in the student's decision to choose the Institution	Nominal
X34	Funding Source	Indicates the source of funding for the studies	Nominal
X35	Financial Responsibility	Indicates who is responsible for payments associated with the studies	Nominal
X36	Intent to Apply for State Aid	Indicates whether the student will apply for state aid	Binary

Appendix B. Careers and programs associated with the study.

N	Program	Shift	code	N° Students
1	Computer Engineering	Evening shift	UNAB21500	17
2	Industrial Civil Engineering	Daytime shift	UNAB12100	139
3	Logistics and Transportation Engineering	Evening shift	UNAB29203	13
4	Geology	Daytime shift	UNAB11350	98
5	Computer Civil Engineering	Daytime shift	UNAB12210	155
6	Industrial Civil Engineering	Evening shift	UNAB22100	20
7	Engineering in Automation and Robotics	Evening shift	UNAB29202	6
8	Industrial Engineering	Evening shift	UNAB22510	13
9	Merchant Marine Engineering	Daytime shift	UNAB14001	93
10	Civil Engineering	Daytime shift	UNAB11300	25
11	Engineering in Automation and Robotics	Daytime shift	UNAB19202	59
12	Civil Engineering in Mines	Daytime shift	UNAB11303	57
13	Computer Engineering	Daytime shift	UNAB11500	134
14	Industrial Engineering	Daytime shift	UNAB12510	7
15	Construction Engineering	Daytime shift	UNAB11400	25