Analyzing Attrition: Predictive Model of Dropout Causes among 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 the Universidad Andres Bello in Santiago, Chile, where currently collaborates with the Educational and Academic Innovation Unit, UNIDA (for its acronym in Spanish), as an instructor in active learning methodologies. Her research interest topics involve university education in STEM areas, faculty and continuing professional development, research-based methodologies, community engagement projects, evaluation tools and technology, and gender issues in STEM education. https://orcid.org/0000- 0002-0383-0179

Ms. Danilo Alberto Gomez Correa, 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

Analyzing Attrition: Predictive Model of Dropout Causes among Engineering Students

Abstract
This Complete Research develops a predictive model to elucidate factors affecting dropout rates in the first two years of tertiary education, using data from 1266 students at a School of Engineering in Chile. Focusing on socio-demographic variables from an institutional survey, such as family background, economic status, and employment, the study employs a quantitative, non-experimental methodology alongside Machine Learning techniques within a Knowledge Discovery in Databases (KDD) framework. Of the methods tested, including Neural Networks (NN), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR), the NN model proved most effective, demonstrating high Accuracy, Sensitivity, and Specificity (all above 0.7). A Weight-Based Feature Importance analysis identified economic factors, family composition, and social relationships as the top variables impacting dropout. This research enhances our understanding of factors influencing student attrition in the School of Engineering. The model acts as an early alert system, identifying potential dropouts and at-risk student groups before they commence their studies. Consequently, it allows the implementation of early interventions, such as financial support, improved study methods, and professional counseling, thereby significantly reducing dropout rates and improving student success.
Keywords: *AI, data mining, dropout, engineering, first-year students, higher education*

Introduction

Over the years, many studies have been conducted to understand why students leave their studies in Science, Technology, Engineering, and Mathematics (STEM) disciplines prematurely. Research has delved into sociocognitive factors that play a critical role in student persistence in university. For instance, sense of belonging [1, 2], self-efficacy [3, 4], identity [5, 6], and intrinsic motivation [7], which are vital to student persistence in university. For instance, Andrews et al. [8] researched how the incorporation of makerspaces impacts students' self-efficacy and sense of belonging concerning design, engineering, and technology. The authors administered a survey and used paired t-tests to analyze changes in nine factors among students over a semester. They observed significant improvements in innovation orientation, design, and technology self-efficacy, and sense of belonging to the makerspace and the engineering community. Additional analyses showed that the effects were based on the student's academic year, gender, race, and interactions between academic year, race, and time. The results support the use of makerspace projects in STEM curricula. However, the study also highlighted disparities in STEM education, with gender gaps in self-efficacy and racial differences in the sense of belonging. Therefore, addressing these disparities and providing equal opportunities for all students in STEM education is necessary.

Concurrently, recent studies have investigated how the structures of educational programs and institutional support impact student retention in STEM fields [9, 10]. These findings suggest that the educational environment, curricular experiences, personal, sociocultural, and demographic factors significantly influence students' decisions to continue or discontinue their university studies. For instance, Olewnik et al. [11] researched co-curricular activities in STEM education, particularly engineering, noting that many students do not regularly participate. The study, which took place at a large public university, examined student motivations for such activities. It used the Expectancy-Value Theory and combined quantitative methodology with interviews to gather their data. The results reveal that

participating students perceive lower costs associated with co-curricular activities. These perceived costs encompass a range of disadvantages, sacrifices, or inconveniences. Such costs include the required time commitment, mental or physical effort, potential interference with other academic or personal responsibilities, and other factors deemed as a 'price' or 'cost' for involvement. Conversely, achievement values, including perceived benefits like acquired skills and enriching experiences, only sometimes motivate participation. The findings suggest that institutions should align co-curricular experiences with learning values to promote student participation.

It is pertinent to note that several factors influence academic achievement, including sociodemographic factors such as family background, economic and employment conditions, and first-generation attendance at a university. Recent studies have demonstrated that these factors substantially impact students' academic performance [12-16]. Boone and Kirn [17] studied engineering students, specifically first-generation students (FGS), who face unique challenges such as limited study skills and familial and economic responsibilities. The research revealed that FGS exhibit a robust sense of belonging to their specialization and possess an engineering identity, motivations, and experiences comparable to those of continuing-generation counterparts (CGC) students. However, FGS value their classroom experiences more positively, attributing this to their professors' support and teaching methods despite having fewer familial resources. The study focused on upper-division students at a Western university, using a 106-item survey to evaluate their social capital, experience, identity, and belongingness. These findings emphasize the importance of considering FGS's unique experiences and perceptions when designing engineering educational programs. It is critical to develop educational programs that cater to the needs of FGS and provide support to help them overcome academic challenges. In summary, the research on student dropout in STEM education reveals a multifaceted problem that requires a multidimensional solution. A comprehensive understanding of the phenomenon can be achieved by combining various perspectives related to sociocognitive and sociodemographic factors, curricular structure, and institutional support. This understanding can lead to developing more effective strategies to improve retention and success in STEM education.

In the context presented, this research paper seeks to introduce an advanced predictive model to analyze and foresee the causes of dropout among engineering students. This model combines sociocognitive and sociodemographic factors to provide a comprehensive analytical approach. The application of advanced artificial intelligence (AI) techniques facilitates the identification of key patterns and trends, which is crucial for formulating and implementing effective student retention strategies. In AI applied to education, educational data mining stands out as crucial for institutions for its ability to anticipate and improve academic performance. These models identify patterns that allow educators and administrators to make better-informed decisions [18, 19].

This work continues a previous study that analyzed factors influencing university persistence [hidden citation 1, 20]. The present research aims to delve deeper into this issue, presenting a predictive model based on sociodemographic factors from the institutional entrance survey, such as family, economic, and employment backgrounds, and contrasting them with dropout rates in the first years of study. It also includes student work preferences (individual or collaborative) and self-perception of skills such as leadership (among others).

Research Questions

RQ1: Can machine learning algorithms effectively predict retention or dropout using academic performance and sociodemographic data?
RQ2: What methods and algorithms are applicable for predicting student dropout?
RQ3: What are the determinative sociodemographic factors for predicting a student's dropout?

The research's target group is students at the School of Engineering at a large private university in Chile. The data from the characterization survey answered by 1266 new students who entered in the first semester of 2022, their respective academic performances from the first semester of 2022 to the first semester of 2023, and the total number of students who have dropped out of the program by the end of the third semester were considered. This research will provide the basis for developing models that facilitate identifying factors that may have a high impact on student dropout upon entering the School of Engineering. This allows for early detection of student groups that may be prone to dropout, enabling intervention to support students according to their specific needs, whether financial, employment, study methodology activities, or career guidance.

The methodology implemented for developing the predictive model is detailed in the subsequent sections. Section II comprehensively describes the procedures, data analysis techniques, and criteria for constructing and validating the model. Next, the obtained results are presented, where the data are analyzed and interpreted in the context of the research objective and questions. Subsequently, the discussion section delves into the meaning and implications of these results, examining how they align with or differ from previous studies and what this might mean for the field of study. This part also addresses possible explanations for the findings and explores the practical implications. Finally, the conclusions synthesize the study's main findings, highlighting its relevance and contribution to existing knowledge. Here, the inherent limitations of the research are also discussed, offering a critical and reflective perspective. Recommendations are proposed based on the study's results, oriented towards educational practices and future research. Furthermore, future work is incorporated, suggesting areas that require further exploration and how the current findings can be a springboard for additional research.

Methodology

The research used a quantitative, non-experimental approach to understand and model the causes of student dropout without manipulating input variables. Machine Learning tools were used to develop the model, using a Knowledge Discovery in Databases (KDD) methodology associated with the Higher Education environment. This methodology involves four stages, as shown in Table 1.

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

Stage 1. Data Preprocessing	Stage 2. Data Transformation	Stage 3. Data Modeling	Stage 4. Interpretation
A database of 1266 students who enrolled in an engineering degree in 2022 was extracted from the institutional characterization survey that all interested students must complete.	A dichotomous transformation of the dependent variable CWA* 2023-10 was performed. The variables in the database were renamed to facilitate analysis and modeling.	Prediction models were developed using machine learning tools such as Multiple Linear Regression, Decision Trees, Neural Networks, and K-Nearest Neighbors (KNN) to classify student dropout. The models were evaluated through cross-validation and performance metrics.	Performance metrics were used to evaluate each model's accuracy and classification errors. These metrics include accuracy, sensitivity, specificity. and F1.

* Current Weighted Average (CWA). The CWA (Chile) and the GPA (USA) are similar metrics used to evaluate academic performance by considering course grades weighted by credits. However, the main difference lies in the grading scales used in the two countries. Chile uses a 1-7 scale, with 4 being the passing score.

The survey in Stage 1 (Tab. 1) covers four dimensions: family, social, and economic factors, previous study experience, and personal skills and study habits. It is worth noting that the Current Weighted Average (CWA) of the first, second, and third semesters of study, which can range from 1 to 7, is also considered. The CWA was collected from the institutional database.

Results

The predictive model's results are presented in the following manner: first, a detailed account of the data processing is given; second, an exhaustive description of the model development process is provided; finally, the predictive model's final output is presented, highlighting its key features and the implications of its results.

Data Preprocessing

After removing non-essential attributes from the original 50-field database, our research dataset was reduced to 34 significant fields. Subsequently, an additional field was added, representing a dependent variable that indicates the student's academic status (Dropout/Non-Dropout), with the value *1(one)* assigned to students who drop out and *0 (zero)* to those who do not. Following this, these variables were normalized for proper integration into the predictive models. The specific fields selected for the design of the predictive model are detailed in Table 2.

Table 2. Fields associated with the dropout prediction model.

No.|	Name	Description	Data Type
X1	Living Arrangement	Specifies with whom the student will live upon starting their studies	Nominal
X2	Currently Working	Indicates whether the student is currently working before starting studies	Binary
X3	Will Work Upon Starting Studies	Indicates whether the student intends to work after starting studies	Binary
X4	Father's Educational Level	Indicates the highest level of education completed by the student's father	Nominal
X5	Mother's Educational Level	Indicates the highest level of education completed by the student's mother	Nominal
X6	First Entry to Higher Education	Indicates whether the student is enrolling in higher education for the first time	Binary
X7	Uses Study Techniques	Indicates whether the student uses any study techniques	Nominal
X8	Reason for Choosing Career	Indicates the main reason for choosing the career	Nominal
X9	Representation of the Career Path	Indicate what best represents you in the field of study	Nominal
X10	Skills	It indicates the skills that the student deems relevant	Nominal
X11	Teamwork	Indicates the level of importance the student assigns to teamwork	Nominal
X12	Leadership	Indicates the level of importance the student assigns to leadership	Nominal
X13	Effective Communication	Indicates the level of importance the student assigns to effective communication	Nominal
X14	Negotiation	Indicates the level of importance the student assigns to negotiation	Nominal
X15	Civic Education	Indicates the level of importance the student assigns to civic education	Nominal
X16	Innovation and Entrepreneurship	Indicates the level of importance the student assigns to innovation and entrepreneurship	Nominal
X17	Contact Level	Assess the level of contact with various close groups	Nominal
X18	Career	Name of the student's career	Nominal
X19	Extended Family	Indicates the level of contact the student maintains with aunts, uncles, cousins, etc.	Nominal
X20	Friends	Indicates the level of contact the student maintains with friends	Nominal

X21	Teachers	Indicates the level of contact the student maintains with teachers	Nominal
X22	Partner	Indicates the level of contact the student maintains with their partner	Nominal
X23	Marital Status	Indicates the student's marital status	Nominal
X24	Family Head	Indicates who assumes the role of head of family	Nominal
X25	Parental Status	Indicates whether the student has children	Nominal
X26	University Selection	This variable indicates a relevant factor in the student's decision to choose the Institution	Nominal
X27	Funding Source	Indicates the source of funding for the studies	Nominal
X28	Financial Responsibility	Indicates who is responsible for payments associated with the studies	Nominal
X29	Intent to Apply for State Aid	Indicates whether the student will apply for state aid	Binary
X30	Program	Refers to the academic program in which the student is enrolled.	Nominal
X31	Shift	Indicates the shift of the student's classes (e.g., morning, afternoon, evening).	Nominal
X32	Program Code	A unique identifier for the academic program modality.	Nominal
X33	Campus	Specify the campus location where the student is attending.	Nominal
X34	Gender	Indicates the student's gender	Binary

Model development

The purpose of the model is to predict the dropout rate of first-year students. To achieve this, a training dataset was used with the help of R-Studio to implement predictive models based on various methods such as Neural Networks (NN), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR). In developing the models, their global performance was evaluated primarily through the *Accuracy* indicator. This metric considers both Type 1 and Type 2 errors. Specifically, Type 1 error refers to instances where a student is incorrectly classified as a dropout. In contrast, Type 2 error pertains to instances where a student is mistakenly identified as not dropping out. To provide a more comprehensive quantification of 'accuracy,' it includes the aggregate of Type 1 and Type 2 errors across all instances. The models' performance was further based on the Kappa statistic. The model incorporated cross-validation, repeated 50 times for the five algorithms; the graph below (Fig. 1) illustrates each algorithm's performance results.

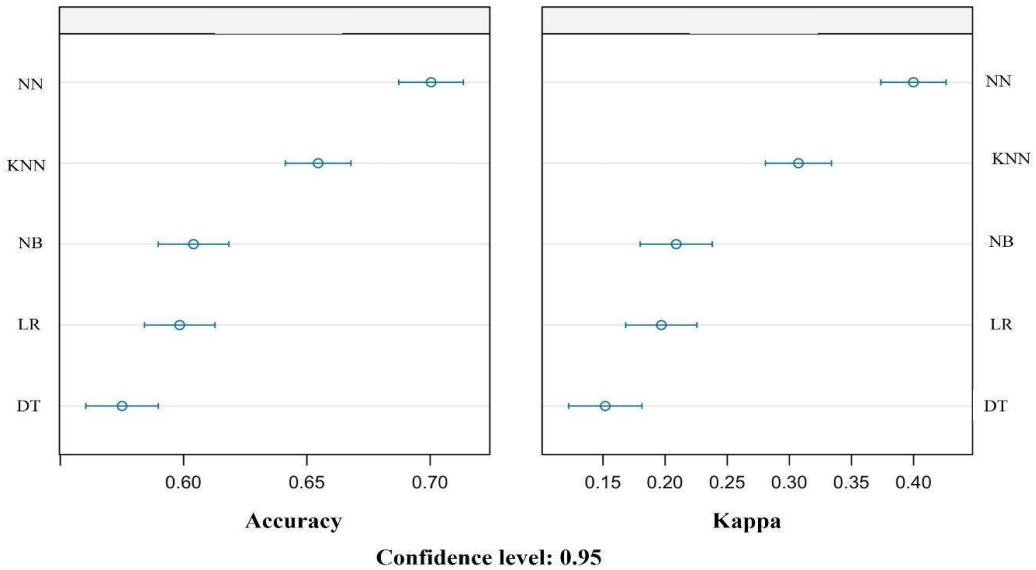

Figure 1. Evaluation of Neural Networks, K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR) models in student dropout prediction.

In this initial training phase, it is evident that the Neural Networks model yields the best results, with an average Accuracy of 0.71 and a Kappa of 0.42. To further assess the accuracy of the developed models, testing and scoring tools were applied based on the criteria of Sensitivity, Specificity, Precision, Recall, and F1 Score. The latter serves as an estimator of an algorithm's classification capability [21]. The results obtained in each model are described in Table 3.

Table 3. Results of the predictive models' evaluation.

Measures	**LR**	**DT**	**NN**	**NB**	**KNN**
Accuracy	0.6000	0.5455	0.7065	0.5584	0.6701
Sensitivity	0.5737	0.8053	0.7105	0.6211	0.5263
Specificity	0.6256	0.2923	0.7026	0.4974	0.8103
Precision	0.5989	0.5258	0.6995	0.5463	0.7299
Recall	0.5737	0.8053	0.7105	0.6211	0.5263
F1	0.5860	0.6362	0.7050	0.5813	0.6116

Among the analyzed models, the best-performing one, based on the main variables associated with the accuracy of the student dropout prediction model, is the Neural Networks model. It scored the following in the most significant variables: 0.7065 (Accuracy), 0.7105 (Sensitivity), 0.7026 (Specificity). The general averages of the performance variables are shown in Table 4.

Table 4. Neural Network model performance metrics.

Model	Average score	Difference
NN	0.7058	
KNN	0.6458	-8.5%
DT	0.6017	-14.7%
LR	0.5930	-16%
NB	0.5709	-19.1%

Final Model

This study employed an artificial neural network with a specific structure to analyze and model a dataset (see Fig. 2). The network featured a hidden layer comprising two neurons, a choice-balancing model complexity, and efficiency. The network's target variable was "Dropout," and all other available dataset variables were used as inputs to predict this target. This configuration allowed for an in-depth exploration of the relationships between" Dropout" and other variables. A key feature of the network was its focus on classification, reflected in its nonlinear output. The logistic function was chosen as the activation function, a common selection for classification problems due to its effectiveness in modeling probabilities.

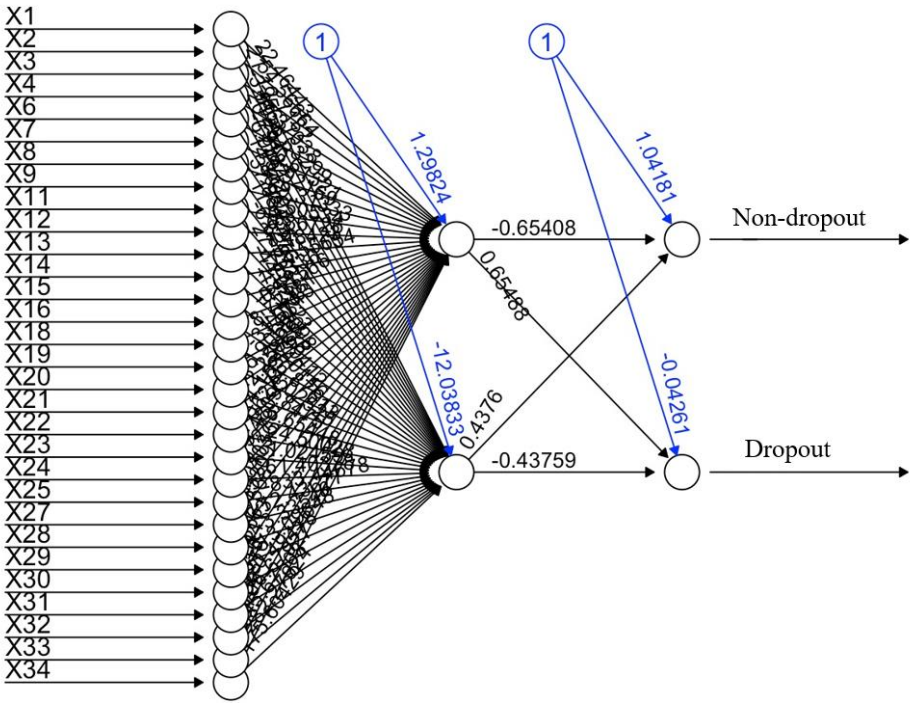

Figure 2. Neural network architecture for predicting student dropout and retention.

Network visualization was a crucial tool for understanding its structure and operation, aiding in interpreting and presenting results. Overall, this neural network configuration offered a balanced and effective approach to data analysis.

Discussion

During the research, five artificial intelligence models were evaluated to determine their effectiveness in predicting student dropout. The models were tested, and it was found that the Neural Networks and K-Nearest Neighbor algorithms performed the best. Meanwhile, the Naive Bayes and Logistic Regression models showed the poorest results. The Neural Networks model was the most reliable tool for prediction, as evidenced by its high values for Accuracy, Sensitivity, and Specificity, all of which exceeded 0.7. This indicates that the Neural Networks model can offer accurate projections for student attrition. To identify the variables that have the most significant impact on classifying whether students drop out or not, the *Weight-Based Feature Importance* method is applied. The results show a range of weights from 7 to 0.03. These results are distributed as shown in Table 5 below.

Table 5. Feature importance in student dropout classification.

Variable	Weight*	Variable	Weight*	Variable	Weight*
X28	*7.00*	X13	3.34	X22	2.82
X25	*5.77*	X16	3.30	X3	2.77
X24	*5.20*	X6	3.20	X23	2.75
X34	*4.36*	X2	3.13	X19	2.70
X20	*4.26*	X8	3.06	X11	2.56
X14	3.61	X21	3.01	X15	2.39
X30	3.49	X4	3.00	X12	2.36
X29	3.42	X31	2.91	X32	2.15
X7	3.40	X27	2.90	X9	2.14
X18	3.36	X1	2.89	X33	0.03

*Variable weights determined by the permutation importance algorithm [22].

The top 5 variables with the most significant impact on classifying the student's status (Dropout/Non-Dropout) are related to economic factors, family composition, and social relationships. These are reflected in the questions of the characterization instrument, which include: Who is the financially responsible person? (X28), Does the student have children? (X25), Who assumes the head role in the family? (X24), Gender (X34), and Level of contact with friends (X20). The following discussion examines the impact of each of these five variables.

Financial support

Of 315 students who dropped out, 211—representing 67% — belong to the group whose education is financed by their parents or relatives. The second-highest percentage was among self-financed students, at 34.21%, and the third-highest percentage was observed in the mixed financing segment, at 30.23%. Of 315 dropouts, 211 were financed by parents/guardians, 65

by themselves, and 39 had mixed financing. This represents a dropout rate of 22.3% among students financed by their parents/guardians, 34.2% among self-financed students, and 30.2% among those with mixed financing. To summarize, most students who dropped out were financed by their parents/guardians, while the highest risk of dropping out was observed in the self-financed and mixed financing segments.

Table 6. Dropout and non-dropout rates by study financing source.

X28.- Who will finance your studies?	Dropout	Non-dropout	% Dropout
My parents/Tutor	211	735	22.3%
Mixed (myself/others)	39	90	30.2%
Myself	65	125	34.2%

When considering the analysis presented on the significance of who finances a student's education, it becomes clear that the source of funding is a vital factor in a student's educational path and perseverance. This relationship is particularly significant in understanding the dynamics that affect student dropout rates. Financial support, or the lack thereof, enables access to education and significantly impacts students' ability to continue and excel in their studies. The differences in dropout rates among various funding segments illustrate the diverse challenges and pressures students face based on their financial support (see Fig. 3). For example, self-finance students may experience more significant difficulties balancing work and study, leading to higher dropout rates.

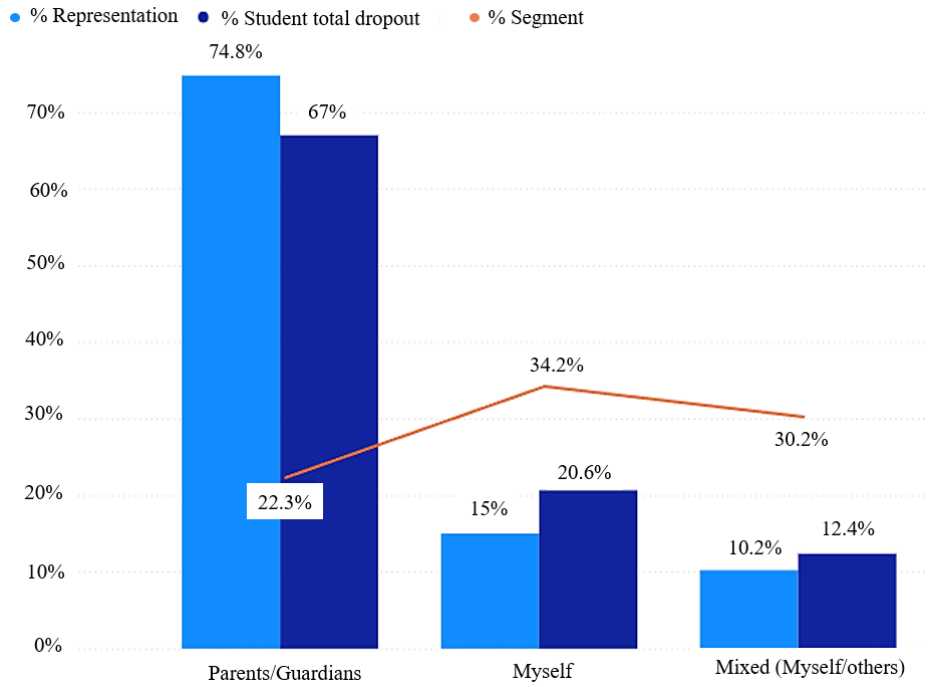

Figure 3. Dropout rates by funding source for educational studies.

Conversely, those whose education is financed by parents or guardians may have more stability and support, potentially reducing the pressure and probability of dropping out. This insight emphasizes the need for robust financial aid systems and targeted support for students in higher education, particularly those at risk of dropping out due to financial constraints.

According to Park and Holloway [23], students from economically disadvantaged families receive less parental support due to their parents' lack of understanding of the importance of

education. This results in less student involvement and a gradual loss of academic interest. It also calls for a deeper understanding of students' socio-economic backgrounds to tailor support systems to address their needs and challenges, thus improving retention rates and ensuring equal access to education. Wilson-Ihejirika et al. [24] stated that securing financial aid can boost persistence among high school students pursuing engineering degrees.

Parental status

Regarding the parental status of students, most dropouts are found among those without children (1222 students), with 306 dropping out. The segment of students with children comprises 44 individuals, of whom 9 discontinue their studies by the end of the third semester. The chart in Figure 4 compares these two segments, where the group of students without children accounts for 96.6% of the 1265 students. Considering only the students who drop out (315), the representation percentage of this segment increases to 97.1%. The dropout rate among students with children is 25%, higher than among those without children, at 20.9%.

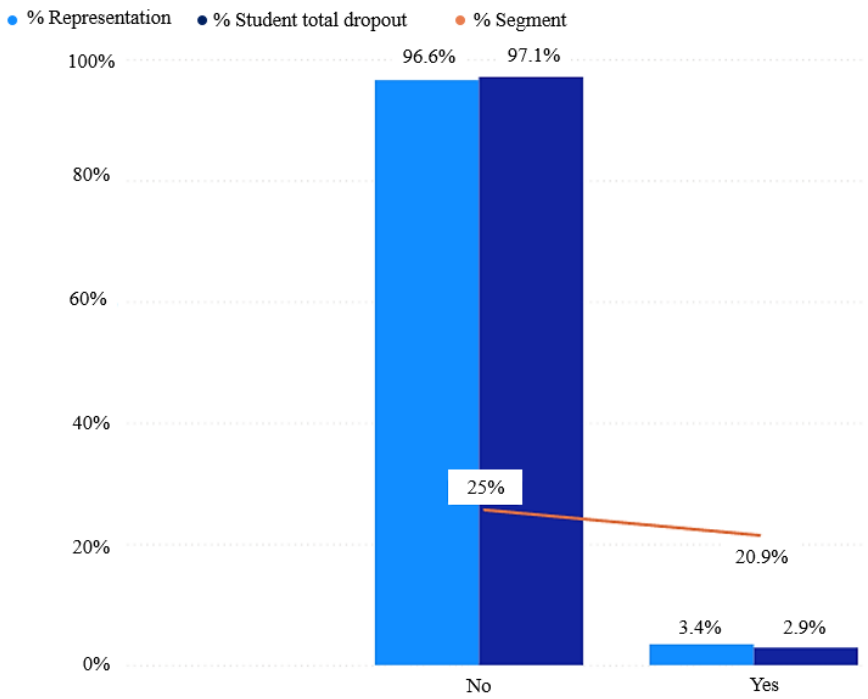

Figure 4. Comparative analysis of dropout rates among students with and without children, highlighting the higher dropout percentage among childless students.

Based on the results observed in the study group, having a child currently does not constitute a significant factor in dropping out. The dropout rate in this segment (20.9%) is lower than that of students without children (25.0%), see Table 7.

Table 7. Dropout and non-dropout rates by parenting situation.

X25.-Do you have children?	Dropout	Non-dropout	% Dropout
No	306	916	25%
Yes	9	34	20.9%

This trend may be attributed to increased familial support for students, enabling them to complete their higher education. Additionally, societal perspectives, which do not criticize

students with children, along with initiatives implemented by the university to support and integrate student-parents, have had a positive impact. These factors collectively contribute to enabling student-parents to continue and complete their studies.

Head of the family role

Variable X24, which relates to the role of the head of the family, impacts the likelihood of dropout. This is illustrated in Table 8.

Table 8. Analysis of student dropout rates concerning family role dynamics and shifts in the role of head of the family.

X24.- Who assumes the role of head of the family?	Dropout	Non-dropout	% Dropout
My grandparents	14	38	26.9%
My spouse or partner	1	8	11.1%
My sibling	6	6	50%
My mother	153	455	25.2%
My father	97	368	20.9%
Another person	2	4	33.3%
Another relative	5	18	21.7%
Myself	36	53	40.5%
Not applicable	1	0	100%

The findings from Table 8 indicate that most students who drop out are those in families where the mother assumes the family role. Furthermore, the segments with the highest dropout probability are those where a sibling or the student undertakes the family role. This reflects a shift in the traditional family structure, where the mother increasingly supplants the father as the head of the household. This trend may be attributed to societal changes experienced over the last few decades, where family configurations are represented differently, moving away from a traditionally patriarchal structure. In situations where most children remain with the mother post-separation, it can lead to a perception of the mother as the primary family head over the father.

Gender

Concerning gender distribution, male students account for 1050 of the totals, with 260 experiencing dropouts. In contrast, female students total 215, with 55 discontinuing their studies by the third semester. For those identifying as "other," only one student has dropped out. A comprehensive analysis, however, necessitates a larger sample size in this category. Therefore, this analysis includes a total of 1265 students. Figure 5 illustrates that male students represent the largest percentage of the total student body and the total number of dropouts, with 83% and 82.5%, respectively. Nevertheless, the difference in dropout probability between male and female students does not appear significant (24.8% vs. 25.6%, respectively).

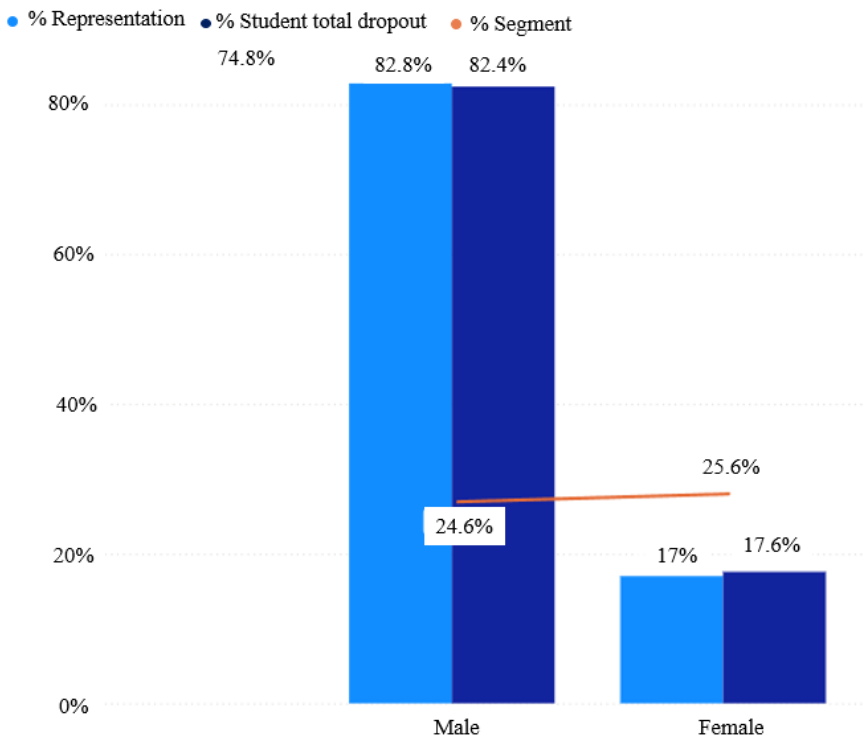

Figure 5. Comparative analysis of student participation and dropout rates regarding gender.

The results shown in the graph highlight a concerning trend of low female participation in the School of Engineering, currently at 17%. This rate slightly increases to 17.6% when considering the total dropout proportion. Female students have a marginally higher dropout rate than male students, indicating a one-in-four chance of female students dropping out by their second year. The subsequent table details the dropout rates across these gender categories.

Table 9. Analysis of student dropout rates concerning student's gender.

X34- Gender	Dropout	Non-dropout	% Dropout
Male (M)	260	790	24.8%
Female (F)	55	160	25.6%

This situation underscores the critical need for initiatives encouraging women's participation in STEM fields and reducing dropout rates. Supporting this, Fowler and Meadows (2013) found that expectancy for success is a more effective predictor of GPA for men than for women, potentially contributing to the higher dropout risk among female engineering students. Additionally, it is essential to consider other sociocognitive factors that could help decrease early-stage dropout rates among women in engineering. These include sense of belonging, as discussed by Emigh et al. [25] and Quezada-Espinoza et al. [26], and self-efficacy, highlighted in the research of Andrews et al. [8]. These factors are pivotal in shaping female students' educational experiences and engineering retention.

Level of contact with friends

Finally, regarding the variable "Level of Contact with Friends" (X20), student responses were categorized into groups of "High Level of Contact" (34 and 79), "Medium Level of Contact"

(127 and 364), and “Low or Almost No Contact” (154 and 507) for students who dropped out and those who did not, respectively (Figure 6).

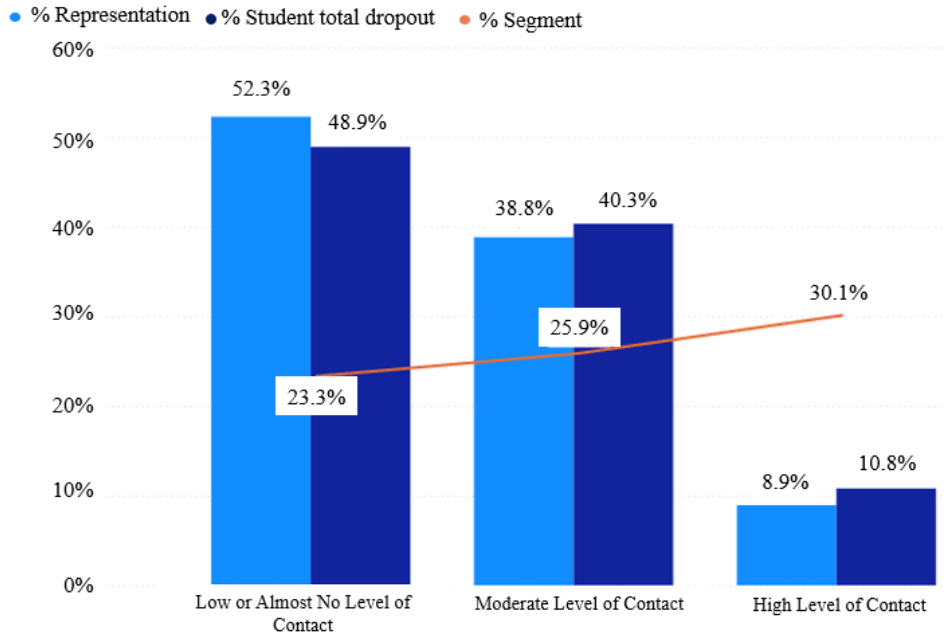

Figure 6. Comparison of dropout rates among students based on their level of contact with friends.

As illustrated in Figure 6, most students who drop out belong to the group that reports having a “Low or Almost No Level of Contact” with friends. However, among the students who report “Having a High Level of Contact with Friends,” there is a higher percentage of dropouts than other students. Table 10 shows more details about students’ rates on the level of contact with friends.

Table 10. Analysis of student dropout rates concerning students’ contact level with friends.

X20-Friends	Dropout	Non-dropout	% Dropout
High level of contact	34	79	30.1%
Moderate level of contact	127	364	25.9%
Low or almost no level of contact	154	507	23.3%

In this regard, Martin et al. [27] mention that the relationships students have outside the academic context are relevant, both for first-generation students and those who are not. If these students have a support system, whether from family, teachers, or friends, they are more likely to persist in their studies. This is especially true if their close circle is involved in STEM careers.

Conclusions

This study proposes a model to analyze student dropout rates at a Chilean private university's School of Engineering. It examines sociodemographic factors such as family background, economic status, and employment from an entrance survey. The focus is on dropouts within the first two years, using data from a survey of 1266 new students in 2022. The study employed a quantitative, non-experimental method to explore and construct a model for

student dropout causes without altering variables. It utilized Machine Learning and the Knowledge Discovery in Databases (KDD) method, tailored for the context of Higher Education.

This research demonstrates the feasibility of identifying engineering students at risk of dropout in the early stages of their studies. With an accuracy of 0.7065, the neural network algorithm stands out, showing its potential to predict student dropout effectively. The analysis revealed that dropout's most influential socioeconomic factors are financial responsibility, parenthood, family head, gender, and social contact level. Higher dropout rates were observed among male students without children, those financed by parents, with mothers as family heads, and those exhibiting low social contact. This analysis pertains to the overall dropout trend within the total student population 1265.

Conversely, a different pattern emerges when examining dropout rates within specific segments of the student population. For instance, among students who self-finance, come from families led by siblings or themselves, and maintain high social contact, the analysis reveals a relatively higher dropout rate within these groups. Specifically, within the self-financed segment of 125 students, 65 dropped out. This nuanced approach highlights that while the general population trend points to certain factors contributing to higher dropout rates, a segmented analysis reveals that different dynamics may influence dropout rates within specific student groups.

Female participation was only 17% in 2022, with a dropout rate of 25.6%. This highlights the need for strategies to increase female participation and success in engineering, thus avoiding labor and economic gaps. Early dropout prediction is a valuable tool for management teams, allowing them to focus support efforts on high-risk students and improve academic management indicators.

Machine Learning, particularly neural networks, has proven to be a significant advancement in predicting student dropout in engineering. This technology effectively analyzes large data sets and accurately identifies key dropout factors. By modeling the complexities of student behavior and circumstances, neural networks offer a deeper perspective than traditional analytical methods. This approach improves early intervention and highlights the importance of ethical technology for educational success. This research underscores the usefulness of machine learning in education and paves the way for future innovations in higher education.

Limitations, future work, and ethics considerations

This study's focus was limited to a specific group of students enrolled in the first semester of 2022. To broaden and deepen the findings, future research will cover subsequent entrance cohorts. Including a wider and more diverse sample will complement current data and provide a more exhaustive comparative analysis. This will enable the development of a more robust database, significantly enhancing our understanding and the initial research results. A longitudinal approach will identify trends and patterns over time and verify the consistency of the results with different student cohorts.

Due to the research's defined scope, the exploration of interactions involving two or more variables was not undertaken in this study. Consequently, informed by the outcomes of this investigation, future work will be dedicated to examining the impact of interactions among the ten primary variables identified herein. This subsequent study aims to provide a more

nuanced understanding of these variables' interplay and collective influence on the research subject.

Recognizing the inherent limitations of employing historical data to forecast future events is essential, as such data may not always accurately project future trends due to the potential impact of unforeseen external factors. Despite these limitations, models based on historical data can still offer significant insights into the factors influencing student persistence in STEM careers, a finding supported by existing literature.

In this context, expanding the study's scope to assess the relevance of its findings to other countries emerges as a valuable direction for future research. It is conceivable that, while the foundational results might have universal implications for student persistence in STEM fields, the influence of specific factors could vary according to different educational systems, cultural contexts, and economic conditions. Therefore, conducting international comparisons could corroborate the initial results and highlight the particularities distinguishing the effects of these factors in diverse environments. This expanded inquiry would greatly enhance our global understanding of the factors affecting STEM student persistence, offering a more nuanced and comprehensive perspective.

During the research on the predictive model, we carefully took ethical considerations into account. Student confidentiality was maintained, ensuring all data were anonymized and untraceable to specific individuals. Regarding data privacy, the research team ensured secure handling and storage of data to protect student privacy. Moreover, we declare that we used the data solely for research purposes. Finally, it is important to note that the study's findings do not impact the integrity of the students involved, either directly or indirectly.

References

[1] M. Becerra-Cid, M. Quezada-Espinoza, M. E. Truyol. (2023). Belongingness of Chilean Engineering Students: A Gender Perspective Approach. 2023 ASEE Annual Conference & Exposition, 37306. https://orcid.org/0000-0002-0383-0179

[2] S. Cwik y C. Singh. “Students’ sense of belonging in introductory physics course for bioscience majors predicts their grade.” *Phys. Rev. Phys. Educ. Res*. vol. 18. n.º 1. p. 010139. May 2022. doi: 10.1103/PhysRevPhysEducRes.18.010139. Available in: https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.18.010139.

[3] L. Ainscough, E. Foulis, K. Colthorpe, K. Zimbardi, M. Robertson-Dean, P. Chunduri, and L. Lluka. “Changes in Biology Self-Efficacy during a First-Year University Course.” *LSE*. vol. 15. n.º 2. p. ar19. jun. 2016. doi: 10.1187/cbe.15-04-0092. Available in: https://www.lifescied.org/doi/10.1187/cbe.15-04-0092

[4] S. Hammad. T. Graham. C. Dimitriadis. y A. Taylor. “Effects of a successful mathematics classroom framework on students’ mathematics self-efficacy. motivation. and achievement: a case study with freshmen students at a university foundation programme in Kuwait”. *International Journal of Mathematical Education in Science and Technology*. vol. 53. n.º 6. pp. 1502-1527. jun. 2022. doi: 10.1080/0020739X.2020.1831091. Disponible en: https://www.tandfonline.com/doi/full/10.1080/0020739X.2020.1831091

[5] A. S. Huffmyer. T. O’Neill. and J. D. Lemus. “Evidence for Professional Conceptualization in Science as an Important Component of Science Identity.” *LSE*.

vol. 21. n.º 4. p. ar76. dic. 2022. doi: 10.1187/cbe.20-12-0280. Available in: https://www.lifescied.org/doi/10.1187/cbe.20-12-0280.
[6] S. Rodriguez. K. Cunningham. and A. Jordan. “STEM Identity Development for Latinas: The Role of Self- and Outside Recognition”. *Journal of Hispanic Higher Education*. vol. 18. n.º 3. pp. 254-272. jul. 2019. doi: 10.1177/1538192717739958. Available in: http://journals.sagepub.com/doi/10.1177/1538192717739958
[7] J. J. VanAntwerp and D. Wilson. “Differences in motivation patterns among early and mid-career engineers”. *J Women Minor Scien Eng*. vol. 24. n.º 3. pp. 227-259. 2018. doi: 10.1615/JWomenMinorScienEng.2018019616. Available in: http://www.dl.begellhouse.com/journals/00551c876cc2f027.1c5a21585279c945.435f72dd3e2d259a.html
[8] M. E. Andrews. M. Borrego. and A. Boklage. “Self-efficacy and belonging: the impact of a university makerspace”. *IJ STEM Ed*. vol. 8. n.o 1. p. 24. dic. 2021. doi: 10.1186/s40594-021-00285-0. Available in: https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-021-00285-0.
[9] H. Taimoory. D. B. Knight. & W. C. Lee. "Exploring the Relationship Between Undergraduate Students’ Level of Engagement and Perception of Support." presented at the *ASEE Annual Conference. June 26. 2022*. [Online]. Available: https://peer.asee.org/exploring-the-relationship-between-and-undergraduate-students-level-of-engagement-and-perception-of-support.pdf
[10] W. C. Lee. D. B. Knight. A. Godwin. L. Ann Moyer. & I. M. Hasbun. "EAGER: Student Support in STEM: Developing and Validating a Survey Instrument for Assessing the Magnitude of Institutional Support Provided to Undergraduate Students at a College Level." presented at the *ASEE Annual Conference. Salt Lake City. Utah. USA. 23-27 June 2018*. [Online]. Available: https://peer.asee.org/board-110-eager-student-support-in-stem-developing-and-validating-a-survey-instrument-for-assessing-the-magnitude-of-institutional-support-provided-to-undergraduate-students-at-a-college-level.pdf
[11] A. Olewnik. Y. Chang. and M. Su. “Co-curricular engagement among engineering undergrads: do they have the time and motivation?”. *IJ STEM Ed*. vol. 10. n.o 1. p. 27. abr. 2023. doi: 10.1186/s40594-023-00410-1. Available in: https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-023-00410-1.
[12] S. Atwood. S. Gilmartin. A. Harris. and S. Sheppard. “Defining First-generation and Low-income Students in Engineering: An Exploration”. in *2020 ASEE Virtual Annual Conference Content Access Proceedings*. Virtual Online: ASEE Conferences. jun. 2020. p. 34373. doi: 10.18260/1-2--34373. Available in: http://peer.asee.org/34373
[13] M. R. Bamberger and T. J. Smith. "First-Generation College Students: Goals and Challenges of Community College." *Community College Review*. vol. 51. no. 3. pp. 445-462. Jul. 2023. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/00915521231163903
[14] B. Helmbrecht and C. Ayars. "Predictors of Stress in First-Generation College Students." *J. Stud. Aff. Res. Pract.* vol. 58. no. 2. pp. 214-226. Mar. 2021. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/19496591.2020.1853552
[15] V. J. S. Hernández. “Estudiantes de primera generación en Chile: una aproximación cualitativa a la experiencia universitaria”. *Revista Complutense de Educación*. vol. 27. n.º 3. pp. 1157-1173. 2016. doi: 10.5209/rev_RCED.2016.v27.n3.47562. Available in: https://revistas.ucm.es/index.php/RCED/article/view/47562
[16] W. c. Lee & H. M. Matusovich. "A Model of Co-Curricular Support for Undergraduate Engineering Students." *J. Eng. Educ.*. vol. 105. no. 3. pp. 406–430. 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jee.20123

[17] H. Boone y A. Kirn. “First Generation Students Identification with and Feelings of Belongingness in Engineering”, in *2016 ASEE Annual Conference & Exposition Proceedings*. New Orleans. Louisiana: ASEE Conferences. jun. 2016. p. 26903. doi: 10.18260/p.26903. Available in: http://peer.asee.org/26903.
[18] C. Chaka. “Educational data mining student academic performance prediction. prediction methods. algorithms and tools: an overview of reviews”. *Journal of e-Learning and Knowledge Society*. pp. 58-69 Pages. ago. 2022. doi: 10.20368/1971-8829/1135578. Available in: https://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/1135578.
[19] P. De-La-Cruz. R. Rojas-Coaquira. H. Vega-Huerta. J. Pérez-Quintanilla. y M. Lagos-Barzola. “A Systematic Review Regarding the Prediction of Academic Performance”. *Journal of Computer Science*. vol. 18. n.o 12. pp. 1219–1231. dic. 2022. doi: 10.3844/jcssp.2022.1219.1231. Available in: https://thescipub.com/abstract/10.3844/jcssp.2022.1219.1231.
[20] *Hidden for blind review*
[21] C. M. Bishop, "Title of the section," in *Pattern Recognition and Machine Learning*, Springer, 2006, pp. xxx-xxx.
[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
[23] S. Park y S. Holloway, “Parental Involvement in Adolescents’ Education: An Examination of the Interplay among School Factors, Parental Role Construction, and Family Income,” *School Community Journal*, vol. 28, n.o 1, pp. 9-36, 2018, Available in: https://eric.ed.gov/?id=EJ1184925.
[24] D. J. Wilson-Ihejirika, Q. Liu, J. Meihui Li, M. Nisar, y J. Lin, “Engineering Pathways from High School to Workplace: A Review of the Literature,” in *2023 ASEE Annual Conference & Exposition*, Baltimore, Maryland, Jun. 2023. Available in: https://peer.asee.org/43335
[25] P. J. Emigh, S. Krishna, J. Liao, K. Kita, J. Casey, and J. Nissen, "Student belonging in STEM courses that use group work," presented at the *Physics Education Research Conference 2023, Sacramento, California, Jul. 2023.* [Online]. Available: https://www.per-central.org/perc/2023/detail.cfm?ID=14367
[26] Quezada-Espinoza, M., Silva, M., & Alvarado, C. (2023). Sense of Belonging of Women in Construction: Insights from Focus Groups. *2023 ASEE Annual Conference & Exposition*, 38336. https://orcid.org/0000-0002-0383-0179
[27] J. P. Martin, S. K. Stefl, L. W. Cain, y A. L. Pfirman, “Understanding first-generation undergraduate engineering students’ entry and persistence through social capital theory”, *International Journal of STEM Education*, vol. 7, n.º 1, p. 37, ago. 2020, doi: 10.1186/s40594-020-00237-0. Available in: https://doi.org/10.1186/s40594-020-00237-0.