[1,2] *Victor Hugo Guadalupe-Mori

[1,3] Ciro Rodríguez

[1]José Antonio Ogosi-Auqui

Predictive Model Based on Machine Learning for the Reduction of Student Desertion in Private Universities in Peru: The Case of Universidad Privada San Juan Bautista

Journal of Electrical Systems

Abstract: - This study addresses dropout issues in private university institutions in Peru by developing a predictive model using ML techniques. It discusses the relationship between Artificial Intelligence (AI) and Machine Learning, emphasizing the latter's role in enabling computer systems to learn and improve from data. The study explores various reasons for dropout, categorizing forms such as complete, partial, early, and late dropouts, emphasizing the need to address this issue due to its impact on students' academic and professional progress. Using a descriptive and explanatory method, the research analyzes 30 cases from a specific private university institution in Peru. It describes the variables, dimensions, and indicators of the predictive model, highlighting personal factors' substantial impact on the model's predictive capacity. Results reveal significant relationships between variables, with the CatBoostClassifier achieving 78.62% accuracy in early dropout detection. The study underscores the importance of considering personal aspects in preventive and support strategies and presents the predictive model as a valuable tool for addressing student dropout in the Peruvian university context.

Keywords: Machine Learning, Student Desertion, Private Universities, Predictive Modeling, Universidad Privada San Juan Bautista.

I. INTRODUCTION

Student dropout is seen as a systemic problem within the larger educational system. The rates of students leaving universities present challenges not just for the students and their families but also for academic institutions, society, and more. This problem is widespread globally, affecting different parts of the world. However, despite its prevalence, there is a lack of effective solutions utilizing technology and available information [1]. Furthermore, the consequences go beyond just education, affecting a nation's economy due to the substantial investments made by governments in the educational sector. To tackle this issue proactively, it is crucial to pinpoint students who are in danger of early academic dropout.

[2] statistics show how this phenomenon affects student retention and graduation. In the international context, for university education, according to data presented in the report Education at the glace [3], university dropout in OECD countries reaches 31 %. For the case of the European Higher Education Area (EHEA), made up of 47 countries, the dropout rate varies from 20 % to 55 %. In Latin America, dropout rates range from 8% to 48%. [4].

At the international level, university dropout is the act of abandoning studies performed by the student without fulfilling the objective of completing their professional career, likewise [5], also states that student dropout affects education systems globally and Chile is no exception, especially in the university sector. There is no single reason that leads students to drop out, but rather it is a multicausal phenomenon, the conditions of each institution take relevance when trying to explain this phenomenon.

Moreover, as highlighted by [6], university student dropout stands out as a significant challenge confronting the higher education system. Concerns among university authorities have escalated due to the surge the need for higher education. The actual quantity of students successfully completing their higher education falls short of expectations, indicating a substantial dropout rate, particularly in the initial semesters. This trend not only poses academic challenges but also engenders financial difficulties for universities. The mismatch between the increasing demand for higher education and the lower-than-anticipated completion rates underscores the gravity of the issue.

For [7], mentions that, most universities present indifference to sexual violence of which mainly students are objects in the university environment and none of them has a protocol to prevent, attend and punish this practice in

[1] Universidad Nacional Federico Villarreal

[2] Universidad Privada San Juan Bautista

[3] Universidad Nacional Mayor de San Marcos

universities despite the cases that have been reported, a key indicator for desertion in universities Similarly, as outlined by [8], in response to the escalating instances of sexual harassment within university settings, there is a pressing need for the establishment of comprehensive protocols. These protocols should encompass preventive measures, investigative procedures, effective management strategies, mitigation efforts, explicit rejection mechanisms, and appropriate sanctions for individuals engaged in such behaviors. Importantly, these instruments must include provisions for actions that facilitate an equitable process for both the accuser and the accused, thereby mitigating the risk of victimization. Additionally, considerations should be made to safeguard the presumption of innocence, ensuring a fair and just adjudication process. This proactive approach seeks to address the serious challenges posed by the increasing prevalence of sexual harassment within academic institutions.

Mental health is of vital importance to everyone, everywhere. Worldwide, mental health needs are considerable, but responses are insufficient and inadequate. By drawing on the latest available data, displaying examples of good practice from around the world and expressing people's direct experience, [9].

In the context of Peru, educational institutions are categorized into two main types: public and private universities. Public universities operate as legal entities under public law, while private universities function as legal entities under private law. Universities are envisioned as academic communities with a dual focus on research and teaching. They play a pivotal role in delivering comprehensive training encompassing humanistic, scientific, and technological aspects. Additionally, these institutions maintain a strong awareness of the country's multicultural reality, emphasizing education as a fundamental right and an essential public service. The university community comprises teachers, students, and graduates, and in accordance with legal regulations [10], representatives of the promoters are actively involved in its governance. Nevertheless, certain indicators highlight high dropout rates as a consequence. Therefore, it's imperative to meticulously analyze every factor contributing to students' decisions to discontinue their studies. The abrupt shift to virtual education forced students to reorganize their academic routines [11].

Other indicators that are considered include the following: school career (entrance grade, entry modalities and readiness for academia), sociodemographic background (age of incoming students), academic abilities, and involvement in extracurricular activities (such as sports and arts participation upon enrollment), as well as the quantity of credits completed, and the mean grade achieved by the end of the first academic semester. [1].

At the local level, the phenomenon the phenomenon of students leaving higher education is no longer a problem that only involves higher education institutions; it involves society. According to [12], dropout is understood as the early cessation of a study program, before reaching the degree. The issue of student dropout is currently the main problem faced by universities when validating their educational offerings in the field of higher education [10]. Likewise [13], states that student dropout has always been a matter of concern due to its multiple implications with the different pattern recognition techniques to expose useful information and formulate inference rules in automatic diagnostic systems.

In the contemporary landscape, the global issue of university dropout is multifaceted, encompassing a myriad of causes. To specifically delineate the scenario of student attrition at a Peruvian university, an extensive examination was conducted, considering diverse factors. The primary focus was on the variable of university dropout, and the analysis delved into several dimensions, namely individual aspects, academic variables, institutional elements, and financial considerations. This meticulous exploration aimed to comprehensively characterize the complex nature of student dropout at the specified Peruvian university [14].

Student attrition at Peru's private universities, including the Universidad Privada San Juan Bautista (UPSJB), has been identified as a multifaceted challenge that affects not only student retention, but also educational quality and institutional stability. UPSJB, like many other educational institutions, is faced with the task of understanding and addressing the reasons behind student attrition to ensure a thriving and sustainable academic environment.

The causes of student dropout are diverse and can include academic, personal, financial and social factors. From academic difficulties to adjustment problems or lack of financial support, these reasons can be complex and multifaceted. Students often face challenges beyond academics, such as mental health issues, family difficulties, or financial stress, which can influence their decision to drop out. By analyzing large amounts of historical student data, this model can identify patterns and trends that help predict which students are at highest risk of dropping out.

This data may include information on academic performance, class attendance, participation in extracurricular activities, socioeconomic and demographic status, among other relevant factors.

The implementation of a predictive model at UPSJB can provide the university with a powerful tool to intervene early and effectively in cases of risk of student dropout. By identifying students who may need additional support and designing tailored interventions, UPSJB can significantly improve its retention rates and promote a more inclusive and supportive educational environment for all of its students. The study and validation of this model in the specific context of UPSJB will allow the university to adapt retention strategies that reflect the unique needs and characteristics of its student population. Additionally, this data-driven approach can help UPSJB optimize the use of its resources and focus its efforts on areas where they can have the greatest impact on reducing student attrition and fostering the academic and personal success of its students.

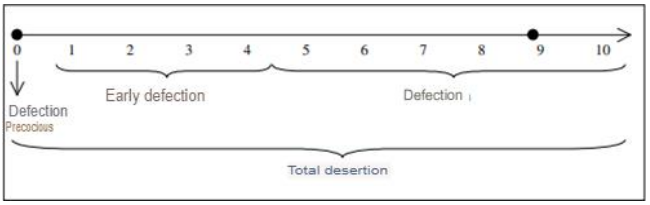


FIGURE 1. Classification of attrition according to time

In relation to time, we can observe that dropout is divided into early dropout, early dropout, dropout and total dropout, each of them with their respective determinants or factors involved.

Hence, this study endeavors to construct a Machine Learning-based predictive model aimed at informing strategies to diminish student attrition rates in private universities across Peru. The particular goals include assessing the precision of the Machine Learning predictive model in gauging the impact of personal factors on student dropout within private universities in Peru, evaluating the dependability of the Machine Learning predictive model in estimating the impact of academic factors on student dropout within private universities in Peru, and appraising the accuracy of the Machine Learning predictive model in determining the influence of socioeconomic factors on student dropout within private universities in Peru.

Student attrition in higher education institutions, particularly in private universities in Peru, has emerged as a significant challenge that affects not only the financial viability of the institutions, but also the educational experience and future of students. In a country where higher education is increasingly important to access job opportunities and improve social mobility, understanding and addressing the reasons behind student dropout has become imperative for universities, including Universidad Privada San Juan Bautista (UPSJB).

The causes of student dropout are diverse and complex. From academic problems, such as difficulties adapting to the pace of university study or facing overly challenging subjects, to personal factors, such as mental health or family problems, the reasons behind a student's decision to abandon their studies can vary widely. Additionally, economic factors, such as the inability to pay tuition or the need to work to support oneself, can also play a significant role in student dropout.

In the specific context of UPSJB, a private university with a broad offering of academic programs and a diverse student population, understanding the dynamics that contribute to student attrition is crucial to implementing effective retention strategies. This involves analyzing historical student data to identify patterns and trends that may predict dropout risk, as well as understanding individual student experiences and needs that may influence their decision to drop out.

The development of a predictive model based on machine learning represents an innovation in this field, as it allows educational institutions to proactively anticipate and prevent student dropouts. By using sophisticated algorithms to analyze large volumes of data, this model can identify early signs of attrition risk, allowing universities to intervene before it is too late. This may involve implementing tutoring programs, academic advising, or financial support to help students overcome the challenges they face and move forward with their studies.

In addition to the predictive aspect, it is also important to address the underlying causes of student attrition by creating a more inclusive and supportive educational environment. This may involve implementing policies and

programs that promote equity and diversity, as well as creating support networks and student communities that help students feel connected and engaged in their college experience.

Ultimately, reducing student attrition at UPSJB requires a holistic approach that combines data analysis, tailored interventions, and an ongoing commitment to improving educational quality and the student experience. By working collaboratively with students, faculty, staff, and other stakeholders, UPSJB is able to create an environment where all students have the opportunity to reach their full potential and achieve their academic and career goals.

In addition to the aforementioned approaches, it is crucial to consider the importance of student guidance and counseling as an integral part of efforts to reduce student attrition at UPSJB. Implementing effective orientation programs can help students adjust to college life, understand available resources, and set clear academic and career goals. By providing academic and personal guidance, advisors can help students overcome obstacles and challenges, as well as identify opportunities for growth and development.

Additionally, it is essential to address socioeconomic and cultural disparities that may influence student dropout. This may involve implementing inclusion and equity policies that ensure that all students have equal access to educational resources and opportunities. Additionally, it is important to recognize and value the cultural and background diversity of students, and foster an inclusive environment where all voices are heard and respected.

Collaborating with external partners, such as local businesses, nonprofit organizations, and government agencies, can also be beneficial in efforts to reduce student attrition. These partners can provide additional resources, such as employment opportunities, scholarships, and mentoring programs, that can help support students and improve their academic and career success.

Furthermore, it is essential to carry out continuous monitoring and evaluation of the programs and strategies implemented to reduce student dropout. By closely monitoring results and adjusting interventions as necessary, UPSJB can ensure its efforts are effective and aligned with the institution's student retention goals.

In summary, reducing student dropout at UPSJB requires a comprehensive approach that addresses the various causes and factors that contribute to this problem. By combining data analytics, personalized interventions, student coaching, and collaboration with external partners, UPSJB can create a more inclusive, supportive, and student-success-oriented educational environment. This will not only benefit individual students by enabling them to achieve their educational and career goals, but will also strengthen UPSJB's position and reputation as an institution committed to the success and well-being of its students.

II. THEORETICAL FRAMEWORK

According to [14], the field of Artificial Intelligence itself lacks, in my opinion, a definition that is at the same time clear, that marks the boundaries well, that is easy to understand and that, in addition, is commonly accepted. Sometimes it is also not easy to define and explain the exact relationship between AI and ML. Approximately, it is understood, and we can consider without much risk, that ML is a subset of AI, although personally I have a minor hesitation regarding this classification. As mentioned by [15], ML involves algorithms that have the ability to learn and enhance their performance autonomously through experience. It is worth noting that the term "on their own" is enclosed in quotation marks to underscore that this autonomous learning occurs through the utilization of data and experiences. This stands in contrast to models where a business expert assigns rules and frameworks based on their knowledge and experience. In statistical models and machine learning models, the emphasis is on allowing the data itself to communicate and automatically derive relationships without explicit human-defined rules.

Machine learning refers to a sub-area of AI focused on providing computational systems (programs and algorithms) with the ability to automatically learn and improve from a specific data set. In short, the objective of machine learning is to make "good" predictive models for "new" data.

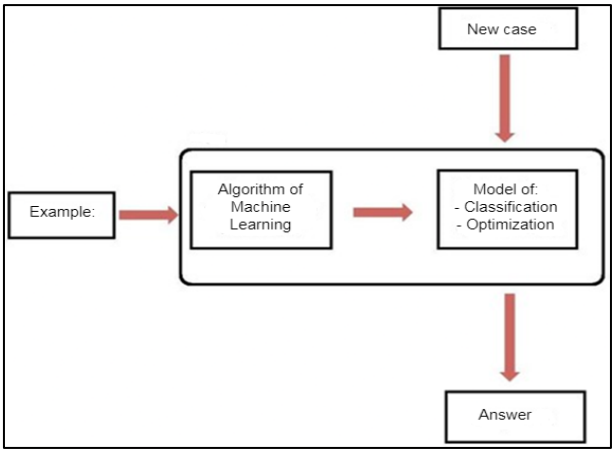


FIGURE 2. Machine learning, training model that adapts this artificial intelligence technique

[16] Indicates that school dropout is the definitive abandonment of studies by a student before the end of their academic period; it may be due to various factors, such as economic, family or socio-contextual problems.

Identifying students prone to premature departure from school is vital for addressing this concern effectively, facilitating personalized interventions suited to everyone's situation. According to the [17], contemporary education confronts numerous hurdles, with significant focus placed on the efficacy of educational systems. Among these obstacles, one of the most daunting is the consistent prevalence of low rates of academic attainment observed across numerous institutions, characterized particularly by elevated rates of student attrition and inadequate student achievement in classes. These hurdles arise from various factors beyond instructional methods, encompassing student characteristics and their ability to manage time autonomously. Although this term is commonly used to refer to dropping out of high school, it can also be applied to any level of education. On the other hand, this problem not only affects students, but also society since it represents a loss of money on the part of potential human capital. Within the domain of higher education, the concept of dropout refers to students discontinuing their academic pursuits, influenced by a multitude of circumstances, opportunities, or obstacles [18]. Exploring the issue of dropout involves examining why many students cease their university studies and fall short of achieving their professional aspirations. In recent years, distance education or online learning has experienced significant growth, facilitated by technological advancements leading to an expansion of academic offerings. However, this mode of education requires refinement, particularly regarding dropout rates, which tend to be higher compared to traditional face-to-face instruction. It's essential to recognize that dropout in higher education encompasses students abandoning their studies due to various circumstances, opportunities, or challenges.

- Economic problems: one of the main reasons, since lacking money makes it difficult to access some necessities or services such as school supplies, transportation, or food.

- Adolescent pregnancy: assuming motherhood in the middle of your school development can mean a great responsibility and, in many cases, results in a partial or total school dropout.

- Teenage pregnancy: assuming motherhood in the middle of your school development can mean a great responsibility and, in many cases, results in a partial or total school dropout.

- Health problems: some health conditions can affect your performance when studying. In some cases, students will be more prone to sleepiness or simply will not focus on their studies. This would generate a desertion on the part of the student.

- Social-contextual problems: These problems are present in the different contexts that students have. These are the following: bullying, citizen insecurity, lack of opportunities, family violence, violence in the neighborhood, problems with classmates, family problems, etc.

Infrastructure problems: some students, especially in rural areas, live far from their educational centers; therefore, they stop attending classes before traveling several kilometers to their schools.

TYPES OF SCHOOL DROPOUTS

This is the classification of school dropout:

- Complete dropout: situation in which the student completely abandons an academic cycle and does not return to study again.
- Partial dropout: scenario in which the student leaves school for a while, but then resumes his or her studies. This case may occur through a leave of absence or special permission - Early dropout: the student stops attending school during the first months of the academic cycle.
- Late dropout: the student abandons his or her lessons after the middle of the school year.
- Early dropout: the child or adolescent decides not to attend any classes, even though he/she is enrolled in his/her grade.

[5] mentions that university student desertion is not a new problem, nor is it exclusive to Peru. This phenomenon occurs all over the world, it is an old problem that has many variables and which is not an exclusive concern of the academic world. University student desertion results in an increase in the number of students with incomplete higher education who enter the labor market and become underemployed without obtaining the desired income, which is detrimental to the students themselves, their families, the country and the university because its budget is affected.

The examination of student attrition at the university level is intricate and holds significant importance, as it is increasingly recognized as a metric reflecting the quality of university administration. Indeed, the rate of university dropout serves as a benchmark in various models employed for evaluating the efficacy of university institutions. Assert that elevated rates of student attrition are indicative of suboptimal quality, suggesting that the university failed to implement essential measures for ensuring the successful completion of degrees by its students.

Every university has developed its unique programs aimed at easing the transition of new students into university life. However, in the majority of instances, these programs are distributed across diverse departments or academic realms, each possessing distinct organizational frameworks. Consequently, the guidance provided to newly enrolled students emanates from various perspectives. Regrettably, instead of offering effective assistance, this decentralized approach often results in heightened confusion for the students, ultimately thwarting the intended objective of facilitating their seamless integration into the university environment.

The complexity of the analysis of dropout lies in the fact that it is a problem of several variables, which, according to [17], can be grouped into those belonging to the pedagogical area and those belonging to the non-pedagogical area. An adequate program of insertion into university life should contemplate the variables of both areas. On the other hand, we believe that these programs, in their design stage, should count on the participation of school authorities since they are the ones who have had our future students for 12 years, on average, in their schools.

Machine Learning represents a branch of artificial intelligence that has gained prominence in various fields, including education. This computational approach relies on the ability of computer systems to automatically learn and improve from experience without specific programming. In the context of student attrition, Machine Learning offers powerful tools to analyze historical student data and predict patterns and trends that can influence a student's decision to drop out.

Student attrition, a phenomenon affecting higher education institutions around the world, is a significant concern for private universities. This problem refers to the situation in which a student abandons his studies before completing his academic program, which can have negative consequences for both the student and the educational institution. The reasons behind student dropout are diverse and can include academic, personal, financial and social difficulties, among others.

Private universities, which rely heavily on student tuition for funding, face additional challenges when it comes to student attrition. The loss of students not only affects your income, but can also have an impact on your reputation and institutional stability. Therefore, these institutions have a significant interest in understanding and addressing the factors that contribute to student attrition to ensure the retention and success of their students.

In this context, predictive modeling has been highlighted as an effective strategy to address student dropout in private universities. Using Machine Learning techniques, researchers and institutions can analyze large sets of student data and develop predictive models that identify students who are at highest risk of dropping out. These models can take into account a variety of variables, such as academic performance, class attendance, socioeconomic and demographic status, among others, to provide accurate predictions about a student's probability of dropping out.

By implementing predictive models based on machine learning, private universities can intervene early and effectively to support students at risk of dropping out. This may involve implementing tutoring programs, academic advising, financial support, and other interventions designed to address individual student needs and foster a more inclusive and supportive educational environment. Ultimately, the use of Machine Learning techniques in predictive modeling offers private universities a powerful tool to improve student retention and promote the academic and personal success of their students.

Machine Learning, as a discipline within the field of artificial intelligence, has become an invaluable tool for addressing complex challenges in a variety of industries and sectors, including higher education. In the context of private universities, where student retention is critical to institutional and financial stability, the use of Machine Learning techniques to predict and mitigate student attrition has emerged as a promising strategy.

A multifaceted and worrying phenomenon, student dropout can be attributed to a variety of factors, ranging from academic difficulties to personal and financial problems. At private universities, where tuition costs are often higher and students may face additional pressures to succeed, understanding and addressing these causes is crucial to ensuring student retention and success.

Machine Learning-based predictive modeling leverages large sets of historical student data to identify patterns and trends that can predict a student's likelihood of dropping out. By analyzing variables such as academic performance, class attendance, participation in extracurricular activities, socioeconomic and demographic status, among others, these models can offer an accurate view of a particular student's risk of dropping out.

The implementation of predictive models in private universities such as the Universidad Privada San Juan Bautista (UPSJB) allows institutions to intervene proactively and personalized to support students at risk of dropping out. This may include allocating additional resources, such as academic tutoring, one-on-one advising, or financial aid, as well as developing programs and policies that promote a more inclusive and supportive educational environment.

In addition to the predictive aspect, Machine Learning can also be used to analyze the effectiveness of existing interventions and retention programs. By continually monitoring student data and evaluating the impact of retention initiatives, private universities can adjust and improve their strategies to address the changing needs of their students.

In conclusion, the use of Machine Learning techniques in predicting and mitigating student attrition represents an exciting opportunity for private universities like UPSJB. By harnessing the power of data analytics and artificial intelligence, these institutions can significantly improve their student retention rates and promote an educational environment more conducive to the success of their students.

III. METHOD

A. Design and operationalization

In this thesis project the type of research is applied at a descriptive and explanatory level, taking as a sample 30 processes of reduction of student desertion in private universities in Peru: The case of San Juan Bautista Private University, taking into account the following operationalization table.

TABLE 1. Variables, dimensions and indicators of this study

Variables	Dimensions	Indicators
Independent: Predictive Model Based on Machine Learning	• Model quality	• Model Prediction Level
		• Attrition rate
	• Operating time	

		• Retention rate
	• Model performance	• Model validation • Prediction accuracy
Dependent: Student Attrition	• Process efficiency	• Degree of influence on prediction variables in determining attrition. • Average permanence • Factors involved in determining attrition.

B. Procedures

The Minitab program will be used for the data analysis, which will be carried out according to the following phases:

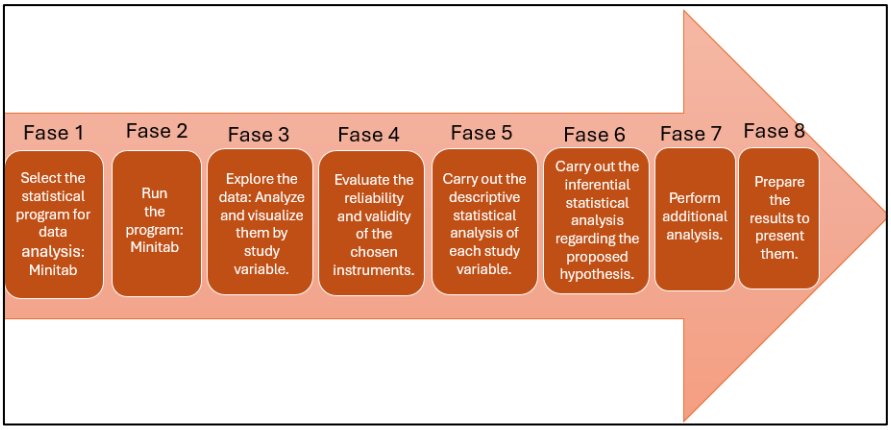


FIGURE 3. Phases for data analysis

IV. RESULTS

Influence of Personal Factors

TABLE 2. Table of factors that are related to the dropout of undergraduate students

Factors	Description
Marital status	Marital status of the student. (categorical)
Application mode	The application method used by the student. (categorical)
Application order	(Order of Application): This is the order in which students submitted their application. (number)
Course	The course the student took. (categorical)
Daytime/evening attendance	Whether the student attends classes during the day or at night. (categorical).
Previous qualification	A grade earned before the student enrolled in a postsecondary institution. (categorical).
Nacionality	Student nationality. (categorical).
Mother's qualification	Student nationality. (categorical)
Father's qualification	Student's Parent Rating: Student's Parent's Rating. (categorical)
Mother's occupation	Occupation of the student's mother. (categorical)
Father's occupation	Occupation of the student's father. (categorical).
Displaced	If the student is a displaced person. (categorical).
Educational special needs	If the student has special educational needs. (categorical)
Debtor	If the student is a debtor. (categorical)
Tuition fees up to date	If the student's registration is up to date. (categorical).
Gender	Student gender. (categorical).
Scholarship holder	If the student is a scholarship recipient. (categorical).
Age at enrollment	The student's age at the time of enrollment. (number).
International	If the student is an international student. (categorical).
Curricular units 1st sem (credited	Number of curricular units for which the student received credit in the first semester. (number)
Curricular units 1st sem (enrolled)	Number of curricular units in which the student is enrolled in the first semester. (number).
Curricular units 1st sem (evaluations)	Number of curricular units evaluated by the student in the first semester. (number).
Curricular units 1st sem (approved)	Number of curricular units approved by the student in the first semester. (number).
Unemplyment rate	Unemployment rate
Inflation rate	Inflation index
GDP	GDP

Delving deeper into the findings, a marked impact of factors of a personal nature on the predictive ability of student dropout is evident. Elements such as family environment, socioeconomic conditions and previous experiences emerge as substantive contributors to the predictive ability of the model. This analysis has made it possible to discern the presence of many fundamental personal variables that play a determining role in the dynamics of this phenomenon. The identification of these variables not only reinforces the intrinsic complexity of student dropout, but also provides a more nuanced and detailed understanding of how personal aspects crucially affect the model's projections and anticipations.

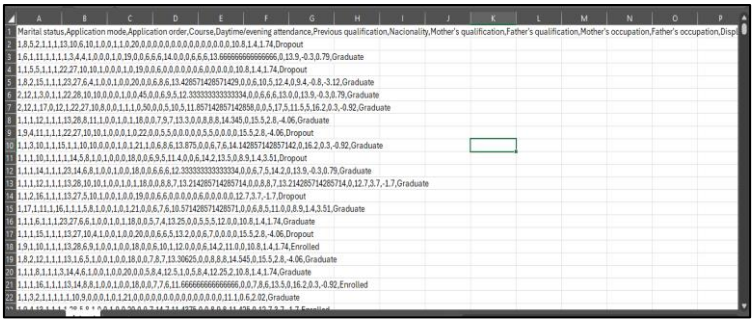


FIGURE 4. Data set for the quantification of dropout factors for undergraduate students, data used for training the Machine Learning model.

TABLE 3. General Securities Data

Marital status values	Mother's and Father's values
- 1: Single - 2: Married - 3: widower - 4: Divorced - 5: De facto union - 6: Legally separated	- 1: Secondary education: 12 years of education or equivalent - 2: Higher education—Bachelor's degree - 3: Higher education—degree - 4: Higher education—Master - 5: Higher education—Ph.D. - 6: Frequency of higher education - 7: Grade 12—Incomplete - 8: Grade 11—Incomplete - 9: 7th year (old) - 10: Others: 11th year of school - 11: 2nd complementary baccalaureate course - 12: 10th year of education - 13: General commerce course - 14: 3rd cycle of basic education (9/10/11 years) or equivalent level - 15: Complementary high school course - 16: Technical-Expert Course - 17: High school remedial course: inconclusive - 18: 7th grade - 19: 2nd cycle of the general baccalaureate course - 20: 9th year of training not completed - 21: 8th grade - 22: General administration and commerce courses. - 23: Accounting and additional management - 24: Unknown - 25: I can't read or write - 26: Can read without having 4th grade education. - 27: 1 cycle of basic education (4th/5th year) or equivalent level - 28: 2nd cycle of basic education (6th/7th/8th year) or equivalent level - 29: Specialized technology course - 30: Higher education—bachelor's degree (1 cycle) - 31: Specialized baccalaureate program - 32: Professional higher technical courses - 33: Higher education—Master (2 cycles) - 34: Higher education—Doctorate (3rd cycle)

TABLE 4. Categorization of the Nationality values factor

Nationality values
1: Portuguese
2: German
3: Spanish
4: Italian
5: Dutch
6: English
7: Lithuanian
8: Angolan
9: Cape Verdean
10: Guinean
11: Mozambican
12: Santo Tomense
13: Turkish

TABLE 5. Categorization of the Application mode values factor

Application mode values
1: 1st Phase - General Conditions
2: Ordinance No. 612/93
3: 1st Phase - Special Division (Azores Island)
4: Other holders of higher courses
5: Ordinance No. 854-B/99
6: Foreign student (total student)
7: 1st Phase - Special Division (Madeira Island)
8: 2nd Phase - General Division
9: 3rd Phase - General Division
10: Ordinance No. 533-A/99, point b2 (Separate Plan) 11: Ordinance No. 533-A/99, point b3 (Other institutions)
12: Over 23 years old
13: Transfer
14: Change of course
15: Holders of technical specialization diplomas
16: Change of institution/course
17: Short-term diploma holders
18: Change of institution/course (International)

TABLE 6. Factor Categorization

Course values
1. Biofuel production technology
2. Animation and Multimedia Design
3. Social Service (Night Assistance)
4. Agriculture
5. Communication Design
6. Veterinary Nursing
7. Information Engineering
8. Cultural Equality
9. Management
10. Social Services
11. Tourism
12. Nursing
13. Dental Hygiene
14. Advertising and Marketing Management
15. Journalism and Communication
16. Basic Education
17. Management (Night Assistance)

TABLE 7. Factor categorization previous qualification values

Previous qualification values
1. Secondary education
2. Higher education—Bachelor's degree
3. Higher education—degree
4. Higher education—Master
5. Higher education—Ph.D
6. Frequency of higher education
7. Grade 12—Incomplete
8. Grade 11—Incomplete
9. Others—11th year of school
10. 10 year school
11. Year 10—Incomplete
12. 3rd cycle of basic education (9th/10th/11th year) or equivalent level
13. 2nd cycle of basic education (6th/7th/8th year) or equivalent level
14. Technology specialization course
15. Higher education—bachelor's degree (1 cycle)
16. Professional higher technical courses
17. Higher education—Master (2 cycles)

TABLE 8. Categorization of the Gender values factor

Gender values
0: woman
1: Male

TABLE 9. Categorization of the Attendance regime values factor

Attendance regimen values
0: night class
1: day class

TABLE 10. Categorization of the Yes/No attributes factor:

Yes/No attributes (yes-1/ no-0)
- Displaced people
- Special educational needs
- debtor
- Tuition fees to date.
- Scholarship holder
- International

It is obtained that the following categories: 1: Students, 12: Other circumstances, - 13: (blank) have a high student dropout rate.

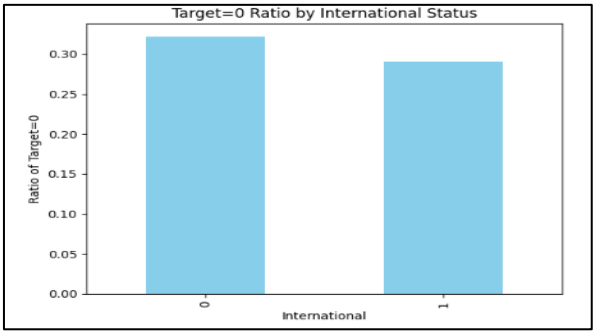


FIGURE 5. International Status vs student dropout (target = 0) check chart.

It is reported that the student, regardless of whether he or she has international status, there is not much difference in the student dropout rate.

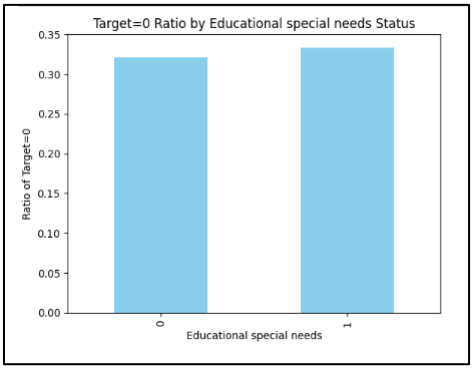


FIGURE 6. Educational special needs vs student dropout (target = 0) check chart.

It is revealed that the student, regardless of the special education he or she has received, does not make much difference in the student dropout rate.

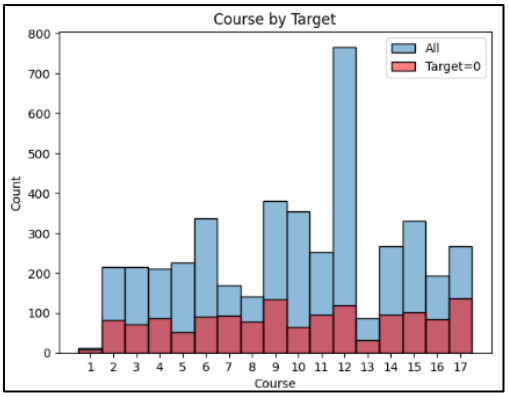


FIGURE 7. Course vs student dropout check chart (target = 0).

Course	Total Count	Target=0 Count	Target=0 Ratio
1	12	8	0.666667
8	141	78	0.553191
7	170	92	0.541176
17	268	136	0.507463
16	192	85	0.442708
4	210	86	0.409524
13	86	33	0.383721
2	215	82	0.381395
11	252	96	0.380952
14	268	95	0.354478
9	380	134	0.352632
3	215	71	0.330233

FIGURE 8. Course vs student dropout data check chart (target = 0).

It is obtained that the values: 1,7,8,17 have high dropout rates. These numerical values mean the following: 1 (Biofuel production technology), 7 (Information Engineering), 8 (Culture of Equality) and 17 (Night Assistance Management). It can be seen that the dropout rate is high in these areas.

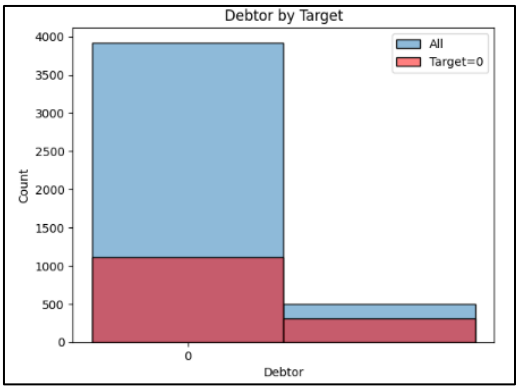


FIGURE 9. Debtor vs student dropout check chart (target = 0)

Debtor	Total Count	Target=0 Count	Target=0 Ratio
1	503	312	0.620278
0	3921	1109	0.282836

FIGURE 10. Debtor vs student dropout check chart (target = 0)

Given what is shown in the graph, it is observed that, whether a student is a debtor or not, there is a large significant difference in the student dropout rate. In other words, the student is a debtor, it would mean that there would be a great possibility that he could drop out of his studies.

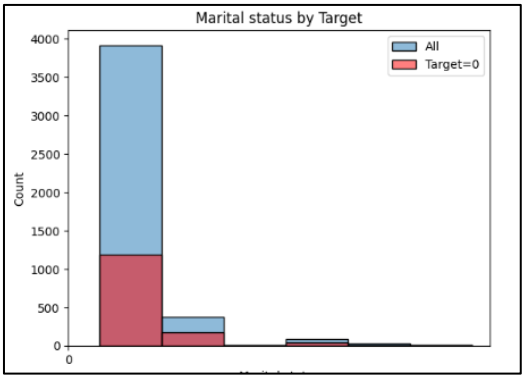


FIGURE 11. Marital status vs student dropout check chart (target = 0)

Marital status

Marital status	Total Count	Target=0 Count	Target=0 Ratio
6	6	4	0.666667
2	379	179	0.472296
4	91	42	0.461538
5	25	11	0.440000
1	3919	1184	0.302118
3	4	1	0.250000

FIGURE 12. Marital status vs student dropout check chart (target = 0)

This leads to the conclusion that: The state of legal separation has a very high abandonment rate.

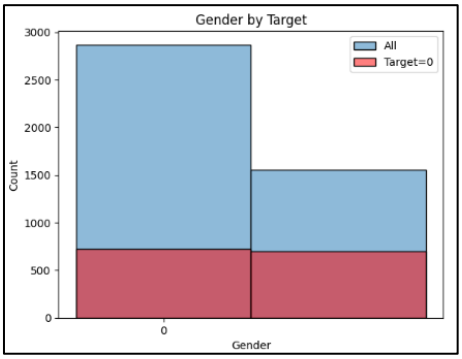
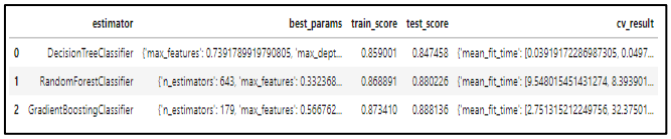


FIGURE 13. Gender vs student dropout (target = 0) check chart

Gender	Total Count	Target=0 Count	Target=0 Ratio
1	1556	701	0.450514
0	2868	720	0.251046

FIGURE 14. Marital status vs student dropout check chart (target = 0)

It turns out that the inflation rate does not seem so relevant compared to the student dropout rate.



	estimator	best_params	train_score	test_score	cv_result
0	DecisionTreeClassifier	{'max_features': 0.7391789919790805, 'max_dept...	0.859001	0.847458	{'mean_fit_time': [0.0391917228698730...
1	RandomForestClassifier	{'n_estimators': 643, 'max_features': 0.332368...	0.868891	0.880226	{'mean_fit_time': [9.548015451431274, 8.393901...
2	GradientBoostingClassifier	{'n_estimators': 179, 'max_features': 0.566762...	0.873410	0.888136	{'mean_fit_time': [2.751315212249756, 32.37501...

FIGURE 15. Comparison between predictive ML models: DecisionTreeClassifier vs RandomForestClassifier vs GradientBoostingClassifier

In our classification study, we evaluate three machine learning models: DecisionTreeClassifier (DT), RandomForestClassifier (RF), and GradientBoostingClassifier (GB). We use cross-validation to obtain a comprehensive view of the performance of each model. All models demonstrated competitive performance on the classification task, with GradientBoostingClassifier (GB) standing out slightly with the highest test score (88.81%). Furthermore, RandomForestClassifier (RF) also showed good performance, outperforming DecisionTreeClassifier (DT) in both training and testing scores. Additionally, it should be noted that cross-validation supports consistency of performance across different partitions of the data set.

The CatBoost algorithm was used as part of the development of a predictive model based on ML. The code introduces the import of the CatBoostClassifier class from the CatBoost library and the definition of a parameter dictionary (param_distributions) for hyperparameter optimization using techniques such as RandomizedSearchCV. The parameters to adjust include the depth of the model (depth) and the number of iterations (iterations). This configuration allows exploring various combinations of hyperparameters with the intention of enhancing the predictive capacity of the model in the early identification of factors associated with student dropout in the Peruvian university context.

In the source code, a classification model is being configured using the CatBoostClassifier algorithm, designed to efficiently handle categorical features and prevent overfitting. The model configuration is established, including the choice to use the CPU for the task (task_type="CPU"), the evaluation metric such as accuracy (eval_metric='Accuracy'), and a random seed for reproducibility (random_seed=42). Subsequently, RandomizedSearchCV is used to carry out a random search of hyperparameters with the objective of finding the optimal combination that maximizes the precision of the model. The search is performed over the parameter space defined in param_distributions with 16 iterations, evaluating precision using 5-fold cross-validation. The final configuration seeks to obtain an efficient and accurate model for the specific task under consideration.

```
cb = CatBoostClassifier(task_type="CPU",
                        eval_metric='Accuracy',
                        random_seed=42,
                        verbose=False)
S
random_search = RandomizedSearchCV(cb,
                                   param_distributions=param_distributions,
                                   n_iter=16, # 하이퍼파라미터 조합의 수 (실험 횟수)
                                   scoring='accuracy',
                                   cv=5,
                                   verbose=2,
                                   random_state=42,
                                   n_jobs=-1)
```

FIGURE 16. Classification model training code using RandomizedSearchCV

A high accuracy, near 1.0 or 100%, indicates that the model is achieving accurate predictions and is able to correctly classify the vast majority of instances in the test set. However, it is important to consider other model evaluation metrics depending on the problem context, especially if there are unbalanced classes or if certain types of errors are more critical than others.

After a meticulous testing process and exhaustive evaluations, the results obtained from the applied Machine Learning model highlight an exceptional accuracy of 78.62% in the ability to foresee student withdrawal within the given context of private universities in Peru. This significant accuracy rate not only evidences the model's ability to anticipate dropout situations, but also underlines its capacity to classify these cases with remarkable reliability. The data reveal that the model has demonstrated a high degree of efficacy in the early identification of student dropout scenarios, which is essential for implementing preventive strategies and providing adequate support to at-risk students. This achievement highlights the robustness and reliability of the model, positioning it as a valuable tool to address the problem of student dropout in the private university environment in Peru.

After finding the optimal model through random hyperparameter search, the feature importance's of the best model are extracted. These importances are obtained by evaluating the relative contribution of each feature to the model predictions. Then, the features whose importance is equal to or greater than 0.5 are filtered, establishing a threshold to highlight the most influential ones. The names of these features are retrieved from the training set. Finally, a visualization is presented in the form of a bar graph, where the selected features are displayed in order of decreasing importance. This analysis provides an intuitive understanding of the most relevant decision-making characteristics of the model, facilitating the interpretation and identification of key factors for the task.

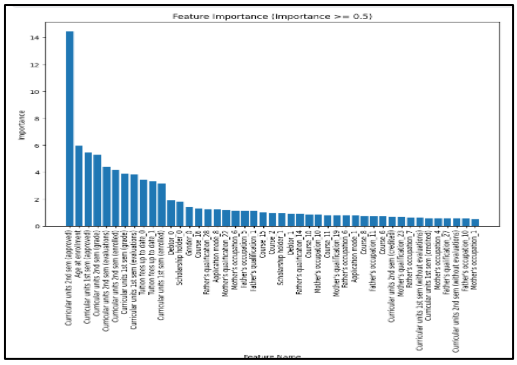


FIGURE 17. Feature Importance Outline

V. CONCLUSION

The research shows that personal aspects, such as family context, socioeconomic conditions, and experiences, fulfill a vital function in the predictive capacity of the student dropout model. The identification of key personal variables offers a more thorough and nuanced understanding of how these elements affect the model's projections, highlighting the inherent complexity of the dropout phenomenon.

A strong association is evident between the educational level of the parents and the student dropout rate, especially with regard to the academic background of both the father and the mother. This finding indicates that the educational and family environment plays a crucial role in the decisions students make about staying in college. The positive connection between parental income and the negative correlation with tuition upgrading underscores the influence of economic conditions on the choice to continue studies.

The prediction-based Machine Learning model, specifically using CatBoostClassifier, evidences an outstanding accuracy of 78.62% in anticipating student dropout in private higher education institutions in Peru. This high accuracy underlines the effectiveness of the model in identifying early dropout situations, thus offering a valuable tool to implement preventive strategies and provide adequate support to at-risk students. The robustness and reliability of the model establish it as a significant contribution to address the problem of student dropout in the private university setting in Peru.

ACKNOWLEDGMENT

I want to express my gratitude to the University for furnishing the essential resources for crafting this article. Special thanks to my colleagues who shared valuable insights, contributing to the creation of the Machine Learning model. Lastly, heartfelt appreciation to my family for providing unwavering support, empowering me to persevere in this research endeavor.

REFERENCES

[1] M. Metropolitana and D. E. Lima, "GOBIERNOS LOCALES," [Online]. Available: [URL].

[2] H. E. Viale Tudela, " A THEORETICAL APPROACH TO THE COLLEGE STUDENT DROP OUT" *Revista Digital de Investigación En Docencia Universitaria*, 2014. [Online]. Available: https://doi.org/10.19083/ridu.8.366.

[3] R. C. Dávila Morán, E. C. Agüero Corzo, H. Portillo Rios, and O. R. Quimbita Chiluisa, "Deserción universitaria de los estudiantes de una universidad peruana," *Braz Dent J.*, vol. 33, no. 1, 2022. [Online]. Available: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202022000200421&lng=es&tlng=es.

[4] S. I. Quintero Solis, "El Acoso y hostigamiento sexual escolar, necesidad de su regulación en las Universidades," *Revista de Estudios de Género, La Ventana*, vol. 6, no. 51, 2020. https://doi.org/10.32870/lv.v6i51.7083.

[5] G. V. Romero, J. S. Toranzo Calderón, S. E. Jaremczuk, J. C. Gómez, and C. Verrastro, "Predictor de deserción universitaria," *Proyecciones*, vol. 19, no. 1, 2021. [Online]. Available: https://ria.utn.edu.ar/xmlui/handle/20.500.12272/5587

[6] "Factors Of Student Desertion And Its Relationship With The Development Of Virtual Classrooms In A Private University In Metropolitan Lima", JNS, vol. 33, pp. 1201–1214, May 2023. [Online]. Available: https://doi.org/10.59670/jns.v33i.2056.

[7] E. Núñez, "Acoso sexual: una realidad invisible en las universidades en Paraguay," *Revista Científica Estudios e Investigaciones*, 2018. https://doi.org/10.26885/rcei.foro.2017.42.

[8] F. J. Mantilla Lozano y P. N. Vilca Yataco, "Machine Learning utilizando el Método Boosting de ensemble para la deserción estudiantil en EBR", Universidad César Vallejo, 2023. [Online]. Available: https://hdl.handle.net/20.500.12692/133709

[9] C. A. Rodríguez Vásconez, "Aplicación de algoritmos de Machine Learning para predecir la deserción estudiantil en alumnos de primer y segundo semestre en universidades públicas del Ecuador", UNIVERSIDAD TÉCNICA DE AMBATO, Ecuador, 2023. [Online]. https://repositorio.uta.edu.ec/jspui/handle/123456789/38615

[10] W. C. Vargas, L. R. F. Gómez, y W. Pineda-Ríos, "Detección de alertas tempranas para la prevención de la deserción estudiantil en una institución de educación superior a partir de un modelo de clasificación y su predicción por medio de técnicas de machine learning", Conocimiento global, vol. 6, núm. S2, pp. 408–426, 2021. [Online]. Available: https://conocimientoglobal.org/revista/index.php/cglobal/article/view/243

[11] N. E. Borjas Ramos y C. J. P. Saqui Marin, "Modelo de machine learning para disminuir la tasa de deserción de estudiantes antiguos en centro de estudio superior", UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS, Perú, 2023. [Online]. Available: http://hdl.handle.net/10757/672020

[12] C. Márquez-Vera, A. Cano, C. Romero, A. Noaman, Y. M. Mousa Fardoun,and S. Ventur. (2016) Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33: 107–124. [Online]. Available: https://doi.org/10.1111/exsy.12135.

[13] E. M. Queiroga et al., "A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course," Appl. Sci., vol. 10, no. 11, p. 3998, 2020, [Online]. Available: https://doi.org/10.3390/app10113998.

[14] Q. Li, R. Baker y M. Warschauer, "Using clickstream data to measure, understand, and support self-regulated learning in online courses", Internet Higher Educ., vol. 45, p. 100727, abril de 2020. Accedido el 31 de marzo de 2024. [Online]. Available: https://doi.org/10.1016/j.iheduc.2020.100727

[15] H. Villarreal Torres, W. Marín Rodriguez, J. Ángeles Morales, J. Cano Mejía, y C. Mejía Murillo, "Classification model for student attrition in a Peru public university", Salud, Ciencia y Tecnología - Serie de Conferencias, vol. 2, núm. 2, p. 175, 2023. [Online]. Available: https://doi.org/10.56294/sctconf2023175

[16] E.G. Añez López and C.A. Añez López, "Factores de Deserción Estudiantil en Institutos Universitarios," Revista Arbitrada Del Centro de Investigación y Estudios Gerenciales (CIEG), pp. 344–358, 2021.

[17] A. Nuñez-Naranjo. Deserción y estrategias de retención: un análisis desde la universidad particular. 593 Digital Publisher CEIT, 5(5-2), 79-87. [Online]. Available: https://doi.org/10.33386/593dp.2020.5-2.306

[18] C. Núñez-Hernández and J. Buele, "Factors Influencing University Dropout in Distance Learning: A Case Study", JHETP, vol. 23, no. 14, Sep. 2023. [Online]. Available: https://doi.org/10.33423/jhetp.v23i14.6379