Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine

Omar Jiménez
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
jimenezramirezomar02@gmail.com

Ashley Jesús
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
ashleyalejandra56@gmail.com

Lenis Wong
Universidad Peruana de Ciencias Aplicadas
Lima, Perú
lwongpuni@gmail.com

Abstract. **University dropout is a problem that not only affects students, but also families, universities, society, and others. This problem has a global character, so it is common to identify it in different parts of the world. However, there are few solutions that efficiently take advantage of available technology and information. Therefore, this study implements a predictive analysis model to identify students at risk of dropout in Peruvian universities and the variables that influence it. For this purpose, the Cross Industry Standard Process for Data Mining (CRISP - DM) methodology is used to develop the model and four Machine Learning algorithms. The methodology consists of five phases: business understanding, data understanding, data preparation, modeling, and evaluation. The experiment was carried out by conducting a survey to 385 students from different public and private universities in Peru, where cognitive, affective, family environment, pre-university, career and university variables were considered. The results showed that the most influential variables in the prediction of university dropout were "age", "term" and the student's "financing method". We also found that the Random Forest algorithm obtained the best performance, with an AUC of 0.9623 in the prediction of college dropout.**

I. Introduction

In Peru, one out of ten university students decides to interrupt their studies [1]. University dropout is a phenomenon that consists in the partial or definitive interruption of a pupil's studies. The phases that a student goes through before making this decision are: perception of not belonging, thoughts of withdrawal/change, deliberation, search for information and final decision [1]. The reasons for initiating a dropout process are diverse. Usually, it is stated that academic performance is not necessarily a determining factor in the dropout process, but that other psychological and social factors of greater complexity are involved [2]. Currently, universities include the dropout rate as a quality indicator, since this problem has been detected in more than 180 countries [3]. In Europe, an average dropout rate of 16% is estimated, and Spain has one of the highest dropout rates, with values of as high as 30%. As a result, university dropout generates a financial impact of more than 1.5 billion euros per year [4]. Due to the global nature of the problem, this topic has generated multiple researchers to explore exhaustively the causes in order to put forward proposals capable of placating the social, economic and organizational impacts that dropout causes [3].

University desertion has different levels and forms: short interruptions, in which a student decides to interrupt his or her studies for a period of up to one academic year, changes of specialization or career, and finally, definitive interruption and non-continuation of studies. Each of these modalities has specific causes that should be studied. According to Tinto's theory (1975), the family work environment or background is an important factor in the prediction of dropout [3]. This means that the student is influenced by the profession and educational level of his or her parents. However, it also mentions social and institutional components. This is one of the most representative theories in academic dropout research. Another important theory on this problem is Spady's theory (1970), which gives more importance to students' links with their classmates, teachers and institution [5].

To solve this problem, research have been presented that exploit different techniques and technology based on data mining. In this study we will review works that contemplate different Machine Learning (ML) algorithms [2], [5]–[13] to address the problem of university dropouts.

For example, in [14], a prediction model was designed by applying Data Mining in a public university in Peru, with the purpose of predicting and determining dropout factors. The objective of the research was to identify students at academic risk in order to propose retention strategies by applying the Support Machine Vector (SMV) algorithm. In [10], the phenomenon of student dropout is studied by using different ML algorithms and, consequently, achieving different results in prediction accuracy.

However, there is still limited research on the peruvian student environment. It has been perceived in the literature reviewed that they have identified the causes of student

dropout, but in other realities. Therefore, this study proposes the implementation of a predictive analysis model applying the Cross Industry Standard Process for Data Mining (CRISP - DM) methodology and four ML algorithms to identify students at risk of dropout and identify the variables that influence peruvian university dropout.

This article is organized as follows: Section 2 presents related work, where the existing literature is reviewed and compared. Section 3 presents details of the implementation methodology and validation results. Finally, section 4 presents the conclusions of the research and recommendations for future work.

II. Related works

For this section, in the first place, research focused on conceptualizing university dropout is considered in order to understand the factors of dropout. The research is divided by types of variables: cognitive, affective, environmental, admission and quality (see Table I).

The factors referring to "cognitive variables" are self-regulation of learning (activities performed by the student to optimize his own learning) [3], [4], [15] and academic performance (number of credits enrolled in the first semester vs. number of credits passed in the first semester) [16], [17]. In the "affective variables", we consider motivation (types of motivation) [3], [18], [19], stress [20], satisfaction with the career (curriculum, study content, and others) [3], [4], [15], satisfaction with the services offered by the university [16], [18] and expectations of self-efficacy and perception of academic performance [3]. In [21], dropout is measured by identifying a student's anxiety. As for the "environmental variables", the research emphasizes the family environment [17], [21], in which the maximum professional degree of the parents and whether the student has domestic obligations at home are consulted. [16]. The integration of the student with his/her teachers and classmates is also evaluated [4], [16], [18], [19]. In [4], social integration, self-regulation of learning and career satisfaction are the factors that determine a student's dropout. Concerning the "admission variables", the student's school context is considered, i.e., the student's preparation and study habits before choosing a program of study [15]. Likewise, if the student participated for vocational training [17], [22]. And with respect to "quality variables", mention is made of the amount of scientific production from 2012 to 2017 at a university in Peru [23].

Secondly, it cites research that built a predictive analytical model to reduce college dropout. These studies define procedures and models that can be extrapolated to this study (see Table II).

TABLE I. Related works grouped by type of feature used

Type of feature	Related works
Cognitive	[3], [4], [15]–[17]
Affective	[3], [4], [16]–[21]
Environmental	[4], [16]–[19], [22]
Admission	[15], [17], [22]
Quality	[23]

TABLE II. Related works and its applied method

Reference	Algorithm	Method	Dataset (sample)
[2]	C4.5 Trees	Selects one variable and distribute over five (05) classes.	Bangor University (4970)
[6]	Decision tree (DT)	Compares academic information in four (04) semesters and develops a predictive model.	Prague University of Economics and Business (3339)
[7]	RF	Recognizes attendance patterns to identify students at risk of dropping out.	Korean National Education Information System (165,715)
[8]	SVM	Using admission and academic performance information, it applies SMOTE to define the dropout result.	Public University of Spain (1418)
[9]	RF	Uses academic context variables, admission and social factors applying SMOTE.	National Autonomous University of Moquegua – UNAM (4365)
[10]	RF	Uses variables of student's academic performance.	Constantine University (261)
[11]	Logistic regression (LR)	Uses information on the student's context (work, class attendance) and academic performance.	Datos de una universidad taiwanesa (3552)
[12]	C4.5 Trees	Relieves quantitative and qualitative student information and applies five (05) different classification algorithms.	Oviedo University (1055)
[13]	RF and XGBoost	Uses the academic results of first term subjects to predict academic dropout.	Universidade de Trás-os-Montes e Alto Douro – (331)
[5]	RF	Applying the CRIPS-DM methodology, it collects general and academic student data.	University of South Pacific (963)

In the work of [2], C4.5 Trees algorithm is applied to identify early on students at risk of dropping out and thus, universities intervene in a timely manner. Based on attendance variables, they show that the model predicts with data from the third week of fall semester classes with 97% accuracy. Similar to [2], in [7] emphasizes that academic attendance is a predominant variable in the identification of at-risk students using the Random Forest (RF) algorithm, a model that obtained an accuracy of 95%.

On the other hand, in [10]–[13] the authors highlight "academic performance" as a variable of great impact during dropout. The model developed in [13] allows inferring a student's lifestyle and social status based on his or her grades. Similar to [13], in [10] different algorithms are compared in an e-learning environment to study the causes that motivate a student to drop out, with the RF algorithm being the most adequate of the model with an accuracy of 93%. In [12] the most important factor is academic performance during the first

year in a statistical analysis model to recognize students at risk of dropping out

In the work of [8], [9], the authors use the Synthetic Minority Over-sampling Technique (SMOTE) methodology to provide a solution to college dropout. This method consists in the use of a surplus of the class of least occurrences. The algorithms used for this purpose are RF and Support Vector Machine (SVM). There are other studies [5], [6], [10] that make use of the CRISP-DM methodology, which consists of understanding the business in order to be more effective in dealing with a problem.

III. Methodology

This section explains the implementation of the predictive model based on the five phases of the CRISP-DM methodology. [24]: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling and (5) evaluation (Fig. 1).

A. Business Understanding

The objective of this phase is to understand the college dropout problem from a business perspective. This allows to correctly identify the data that needs to be collected to effectively address the problem. In this sense, after having reviewed the existing literature, the objective is defined as the early identification of students at risk of dropping out. In addition, the use of different categories of variables that will be explained later in this research is based on this objective.

Fig. 1. Adapted sequence of the CRISP – DM process [24]

B. Data Understanding

In this phase, the categories of variables that make up the predictive model are defined. Spady's theory is considered as a criterion for the identification of variables [5], Tinto's exchange theory [3] and Bean's conceptual model of undergraduate student dropout [21]. Spady emphasizes that a university student interrupts his or her studies when he or she loses connection with the institution in which he or she is educated, in other words, social relationships and affection can be critical variables in the dropout process [5]. Tinto has his own approach to the causes and motivations of the dropout process. Family environment factors, pre-university experience, university environment, among other characteristics, can give indications of how susceptible a university student is to the process of dropping out [3]. Unlike the aforementioned authors, Bean develops a model of student dropout that contemplates four main sets of variables: academic performance, psychological outcomes, previous academic experiences and environmental variables [21].

Table III shows the categories that group the variables and what each one consists of.

TABLE III. Feature type definition

Type of feature	Description
General data	It refers to the personal data of the student such as age, sex and others [9].
Career and university	It refers to the student's profession [18].
Pre-college education and admission of the student	The relationship between dropout and pre-university training of the student is analyzed [9].
Family environment	The relationship betwenn dropout and family environment of the student is analyzed [3].
Cognitive variables	It is about student's performance and class attendance. [3].
Affective variables	It is about student's affections with certain situations like quality of education and others [3].

Concerning the statistical sample used for data collection, it is defined that the participants are students of Peruvian universities who are currently pursuing a degree. The information collected allows us to identify the dropout risk for each student. Since it was not possible to use the personal data of students due to legal issues, a survey was carried out.

Eq. (1) was used to determine the sample:

$$sample\ size = \frac{z^2 \times p \times (1-p)}{c^2} \quad (1)$$

A sample size of 385 records was defined, considering the following parameters:

- Confidence level (z) = 95%
- Margin of error (c) = 5%
- Mean deviation (p)= 1.96

Table IV shows the survey questions asked to students from different universities in the second half of 2022. The questions are grouped by variable categories. This consists of twenty-seven questions referring to general data (age, sex, marital status and residence), career being studied, university, semester to which he/she belongs, mode of entry, family environment and financing of studies, integration and academic performance, whether the student has thought of changing career and/or withdrawing from the career or withdrew from the career.

TABLE IV. LIST OF QUESTIONS GROUPED BY TYPE OF FEATURE

Question		Type
General data		
Q01	Which is your gender?	Open
Q02	Which is your marital status?	Close
Q03	Which district do you live in?	Open
Q04	How old were you when you entered college?	Close
Q05	What is your age?	Open
Q24	Do you have a disability?	Close
Career and university		
Q06	Where do you study?	Open
Q07	What career are you studying?	Open
Q08	Which semester are you in?	Close
Q10	What was the reason for choosing your career?	Open
Q11	How satisfied are you with your choice of career and college?	Close
Pre-college education and admission		
Q09	What was your admission modality?	Close
Family environment		
Q12	What is your parent's or legal guardian's highest level of education?	Close
Q13	Do you have obligations in your family circle?	Close
Q14	What is the marital status of your parents?	Close
Q15	Are you currently working or interning?	Close
Q16	If you answered Yes to the previous question, does this work have an impact on your study habits?	Close
Q17	Do you consider that your family environment encourages or promotes adequate study habits?	Close
Q18	What type of financial support do you have to finance your studies?	Open
Q19	Do you have debts with your university?	Close
Cognitive variables		
Q20	Do you regularly attend most of the classes?	Close
Q22	In relation to your classmates, how do you consider your academic performance?	Close
Affective variables		
Q21	How do you consider your relationship with your teachers and classmates?	Close
Q23	Have you ever been the target of harassment, discrimination, prejudice, among others while in college or high school?	Close
Q25	Have you thought about or have you interrupted or abandoned your studies temporarily or permanently?	Close
Q26	Have you thought about or decided to change careers?	Close
Q27	How closely do you identify with the following sentences? • College is expensive. • College is frustrating. • My career provides me with useful knowledge. • I value my time at university.	Close

C. Data Preparation

Data cleansing

Once the 385 responses were received, a data cleaning and coding process was carried out. These data were treated in such a way that the model can be trained by this information and make predictions possible (for example: coding qualitative variables into numbers).

On the other hand, although the survey is mostly composed of multiple-choice questions, there are some free-text questions. For this reason, it was necessary to group these responses into categories and eliminate records that did not meet the conditions of the sample we were looking for. The affective and cognitive variables, such as "Relationship with professors", "Class attendance", "Frustrating experience" and "Satisfaction" (Table V), reflect the university experience from the student's perspective. For example, "Relationship with professors" included responses such as "Very bad", "Bad", "Good" and "Very good". Therefore, each option was assigned a number from 1 to 4, respectively, to convert these variables into numerical variables. As a result of this data cleaning, a total of 322 student records were obtained that met the conditions defined for the sample.

Among the 322 respondents, 186 were male, while the remaining 136 were female. The survey collected information from different faculties. In addition, it was found that the age of the respondents was typically between 16 and 31 years old. The survey also collected funding information for the studies, although numerically accurate data was not requested.

Feature Selection

For the feature selection, two statistical concepts are introduced that will increase the predictive accuracy indicator achieved by the model: permutation feature importance and correlation of variables. Table V shows the list of twelve variables considered for the development of the predictive analysis model.

TABLE V. FEATURE DEFINITION

Código	Feature	Descripción
F01	Age	Indicates the respondent's age in years.
F02	Term	Indicates the current academic period.
F03	Relationship with teachers	Indicates the quality of the relationship between teachers and student.
F04	Work or internship	Indicates the student's employment status
F05	Financing method	Indicates the mechanism by which the student's degree was financed.
F06	Class attendance	Indicates the frequency with which the student is absent from class.
F07	Frustrating experience	Indicates the frustration level of the student in dealing with college
F08	Satisfaction	Indicates the student's satisfaction with his or her career choice
F09	District	Indicates the student's district of residence.
F10	Admission Mode	Indicates the student's mode of admission to college
F11	Father's Grade	Indicates the father's level of education.
F12	Mother's Grade	Indicates the mother's level of education.

The importance of variables or "permutation feature importance" [8] is a concept that makes it possible to identify the relationship between variables and the outcome of a prediction. For example, Fig. 2 shows that "age" (F01), "term" (F02) and "financing method" (F05) are the variables that have the greatest impact on the prediction. This concept aims to reduce uncertainty by identifying variables that, far from being related to the final result, are values that could well be random and the prediction would remain the same. In this sense, a permutation with 10 repetitions that randomly order the

variables was considered in order to identify those that detract from the accuracy of the model.

Correlation of variables is a concept that defines how influential one variable is for another variable [5], [10]. Fig. 3 shows the Variables Correlation Matrix. This graph is generated by studying the relationship between variables. A cleansing of correlated variables was performed leaving the most important variable among the pairs. Consequently, a low degree of correlation was achieved.

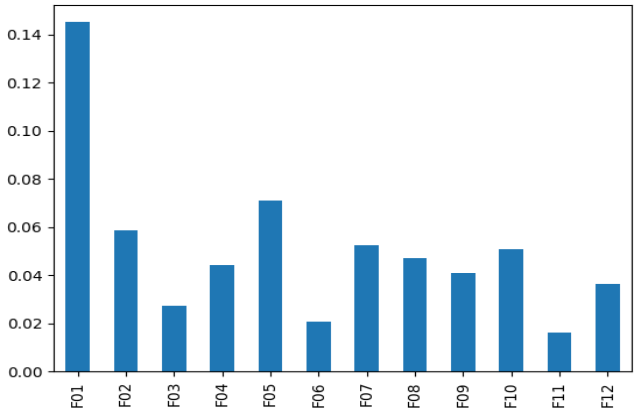

Fig. 2. Feature importance

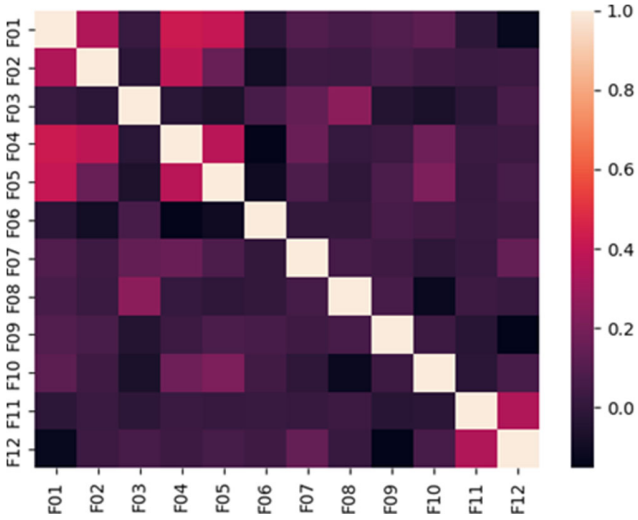

Fig. 3. Feature correlation matrix

For example, with the inclusion of this concept in the dataset, it is observed that there is a relatively high implication between the variable "Age" (F01) and "Work or internship" (F04). This relationship makes sense, since it has been noted in the data collection that the older the respondents are, the more common it is to see that they have work or professional training obligations.

In a first iteration, without considering these two statistical concepts, the prediction model achieved an accuracy of approximately 70%. This value was improved with the inclusion of the importance of variables and the correlation matrix, since it allowed identifying and discarding variables that are not significant for the prediction. After calibrating the variables, a result of around 93.8% was achieved with the algorithm with the highest accuracy.

D. Modeling

In this stage, ML algorithms are applied to the information collected from the survey data. To predict the permanence or dropout of a university student, the proposed model will use information from cognitive, affective, family environment, pre-university, career and university variables to identify behavioral patterns that allow identifying students at risk of dropping out.

Currently, there are different binary classification algorithms that can be used to predict academic dropout. Taking into account the works reviewed and knowing the types of data collected, the following classification algorithms were applied [10]:

1) Decision Tree (DT): This is a supervised learning algorithm consisting of a tree structure. It is a data structure in which, at each node, an evaluation is placed according to a certain characteristic that yields two possible results. This evaluation allows to continue to the next node until the tree is completed.

2) Random Forest (RF): It is an algorithm that uses multiple decision trees for classification tasks. For classification, the prediction result will be based on the majority result of multiple decision trees.

3) Support Vector Machines (SVM): Similar to RF, this algorithm is applied to classification and regression tasks. In classification tasks, the algorithm builds or sets up a hyperplane that divides a dataset into classes.

4) Neural Network (NN): It is an algorithm that is composed of neurons and artificial nodes. As input, it takes data to structure them and maximize the prediction accuracy. Finally, the result will be the processed data divided into different layers.

Regarding the implementation of models with ML algorithms, for example, it was determined that the model using RF would have 100 estimators and a quality criterion based on 'gini' as in DT. On the other hand, it was defined that the NN model would have three layers with 150, 100 and 50 neurons, in addition to 300 iterations. Finally, it was decided that the SVM model would have a 'linear' kernel and probability estimation enabled.

To compare the potential of the algorithms in student dropout prediction, metrics indicating the performance of each algorithm are calculated. The metrics used are accuracy, recall (or true positive rate), precision, fall out (or false positive rate) and ROC represented by equations (1), (2), (3), (4) and (5) respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Fall\ out = \frac{FP}{FP + TN} \quad (5)$$

The following terms should be considered:

- True positive (TP): correct predictions of college student dropout.
- True negative (TN): correct prediction of college student non-dropout.
- False positive (FP): incorrect prediction of college student dropout.
- False negative (FN): incorrect prediction of college student non-dropout.

Metrics are defined as:

- Accuracy (ACC): is the overall proportion of correctly made predictions against the total number of predictions made [10].
- Recall: is the proportion of positive reals that are true positives [10].
- Precision: is the proportion of true negative that are expected to be negative [10].
- Fall out: is defined as the proportion of incorrect predictions [25].

These values are used to calculate the metrics mentioned above and are the result of analyzing the confusion matrices generated from the performance of each algorithm.

On the other hand, the ROC curve is a commonly used metric in models applying classification tasks. The purpose of AU-ROC (Eq. 6) is to distinguish the accuracies of each model according to its classes [13].

$$AUC = TPR \times d(FPR) \quad (6)$$

According to [24], when the AUC score is 1, it means that the classification model is able to distinguish all positive and negative outcomes correctly. If the AUC value is between 0.5 and 1, the model recognizes more positive than negative results.

To validate the performance of our models, samples of 193 students are used and classified into "dropout" and "non-dropout". Then, by calculating the metrics using the confusion matrices (Fig. 6), the algorithms that achieve the best results in correct and incorrect predictions in college dropout are identified. Table VI shows the number of correct and incorrect results by outcome and algorithm. In that sense, the model using the RF algorithm is the one that achieved the highest number of correct predictions with 181, followed by DT, NN and SVM with 177, 165 and 151 correct predictions respectively.

TABLE VI. CORRECT AND INCORRECT DROPOUT PREDICTIONS BY ALGORITHM

Algorithm	Result	Correct predictions	Incorrect predictions
RF	Non dropout	115	7
	Dropout	66	5
NN	Non dropout	108	8
	Dropout	57	20
DT	Non dropout	103	8
	Dropout	74	8
SVM	Non dropout	108	11
	Dropout	43	31

Fig. 4 shows the comparison of the ROC curve result of the four algorithms. It is observed that the algorithm that obtained the best AUC value is RF with a value of 0.9623. The values collected from each algorithm are considered as good in capturing positive values; however, it is observed that NN, SVM and DT generate more false positive predictions, which reduces the accuracy of the final predictive performance. According to the percentage of false positives, the RF algorithm has few failures in incorrect predictions (12). Also, Fig. 5 shows the ROC curve obtained in each algorithm.

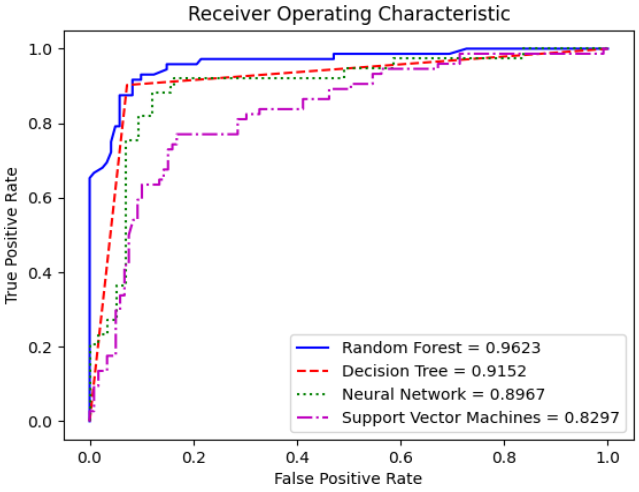

Fig. 4. ROC Curve Comparison

E. Evaluation

Table VII shows the comparison of the algorithms with respect to the previously defined metrics. It is observed that the highest dropout *accuracy* was obtained by the RF algorithm (92.9%). Regarding the recall metric, the algorithms that achieve the best results are RF and DT. It is observed that these algorithms, in addition to achieving the best results, also share a tree-based structure. If we compare these results with those obtained in [9], the RF algorithm obtains very similar results with around 96% for the *recall* of students who "do not drop out". On the other hand, the algorithms that obtain the lowest recall values are NN and SVM between 80% and 88%. Regarding the *fall out* metric, RF and SVM show the lowest values.

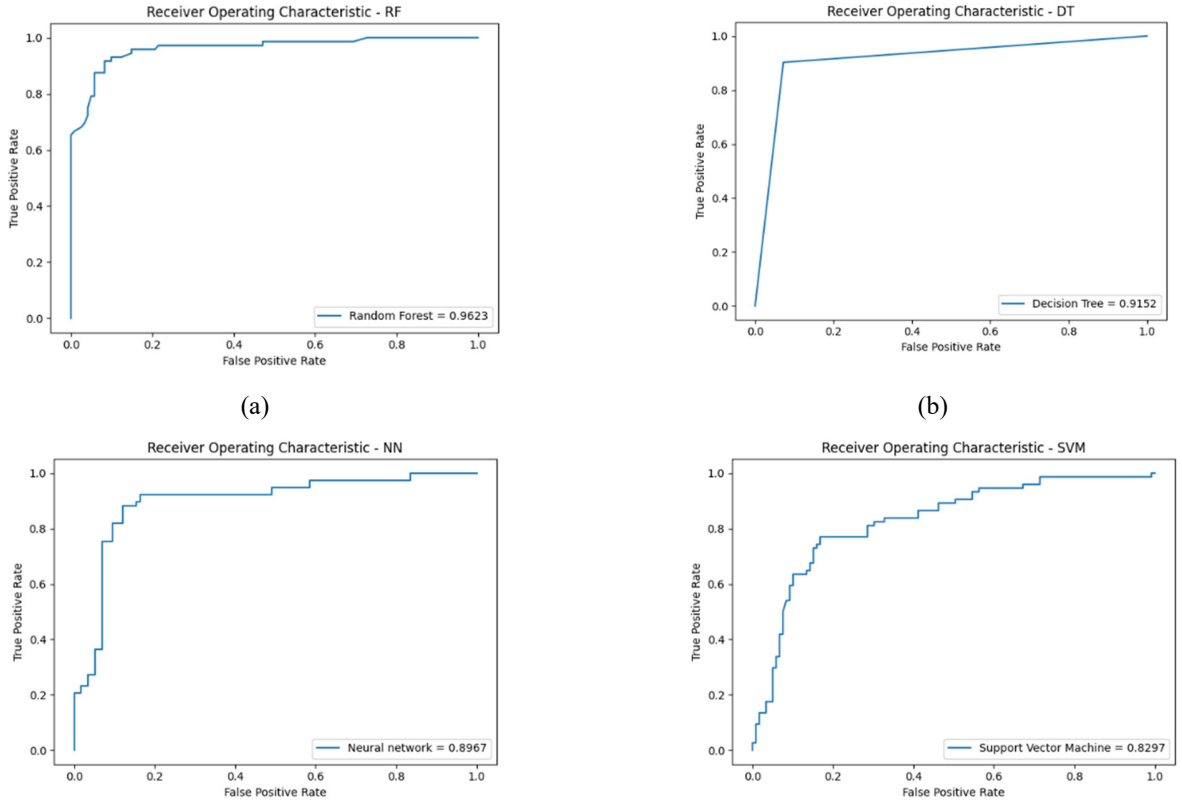

Fig. 5. ROC curves of algorithms used in model development (a) RF, (b) DT, (c) NN, (d) SVM

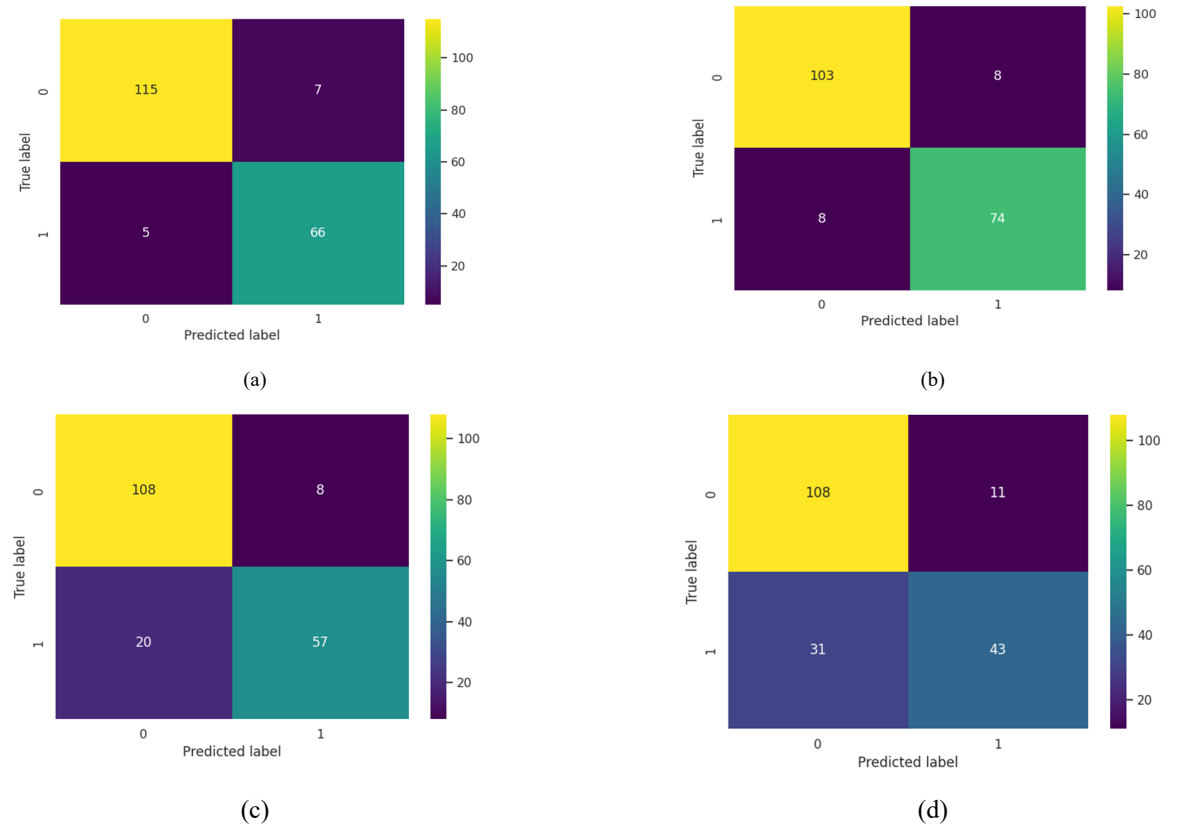

Fig. 6. Confusion matrices on the predictions made by (a) RF, (b) DT, (c) NN, (d) SVM

As for the *precision* metrics, all algorithms achieved a result greater than 90% in the prediction of students at risk of 'non-dropout'. Finally, the algorithm that obtained the best result in accuracy, which indicates the proportion of correct predictions of the entire validation dataset, was RF. The results presented in [10] show accuracy metrics between 77% and 93%, the latter also belonging to the RF algorithm.

TABLE VII. COMPARISON OF METRICS BY ALGORITHM

Alg	AUC	Result	Recall	Fall out	Precision	ACC
RF	0.9623	No dropout	0.958	0.096	0.943	0.938
		Dropout	0.904	0.042	0.929	
DT	0.9152	No dropout	0.928	0.098	0.928	0.917
		Dropout	0.902	0.072	0.902	
NN	0.8967	No dropout	0.844	0.123	0.931	0.855
		Dropout	0.877	0.156	0.740	
SVM	0.8297	No dropout	0.777	0.204	0.908	0.782
		Dropout	0.796	0.223	0.581	

IV. CONCLUSIONS AND FUTURE PROJECTS

Academic dropout rates are often a key indicator in the analysis of the educational quality of a country. In this paper we proposed a predictive analysis model applying four ML algorithms to monitor dropout in Peruvian universities using the CRISP - DM methodology. To make predictions, four ML algorithms were used: RF, DT, NN and SVM.

The dataset was obtained through a survey of students from various public and private universities in Peru. A statistical analysis was applied to detect the most influential variables and the correlation between them. The most influential variables were "age" (F01) and "admission mode" (F10). Also, a relationship was found between the student's "age" (F01) and "work or internship" (F04). As the student is older, it is common for him/her to assume greater responsibilities and the same is true for "term" (F02) and "work or internship" (F04).

On the other hand, during the implementation of the 4 ML algorithms, different evaluation measures were used: Confusion Matrix, ROC Curves, accuracy, precision, recall and fall out. The results showed that the RF algorithm obtained the best results, obtaining an AUC of 0.9623 in the prediction of college dropout.

As future work, it is recommended to continue exploring opportunities for extrapolation of the model, for example, applying the model in technical training institutions, primary and secondary education or even training schools.

ACKNOWLEDGEMENTS

We deeply thank the students at the universities to participate in this research and the Research Department of the Universidad Peruana de Ciencias Aplicadas for the support provided to carry out this research work.

REFERENCES

[1] Andina, "Minedu: tasa de deserción en educación universitaria se redujo a 11.5 %," *Andina*, 2021, [Online]. Available: https://andina.pe/agencia/noticia-minedu-tasa-desercion-educacion-universitaria-se-redujo-a-115--868634.aspx

[2] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Comput. Educ.*, vol. 131, no. January, pp. 22–32, 2019, doi: 10.1016/j.compedu.2018.12.006.

[3] A. D. Mujica, M. V. P. Villalobos, A. B. Bernardo Gutiérrez, A. C. Fernández-Castañón, and J. A. González-Pienda, "Affective and cognitive variables involved in structural prediction of university dropout," *Psicothema*, vol. 31, no. 4, pp. 429–436, 2019, doi: 10.7334/psicothema2019.124.

[4] A. Castro-Lopez, A. Cervero, C. Galve-González, J. Puente, and A. B. Bernardo, "Evaluating critical success factors in the permanence in Higher Education using multi-criteria decision-making," *High. Educ. Res. Dev.*, vol. 00, no. 0, pp. 1–19, 2021, doi: 10.1080/07294360.2021.1877631.

[5] M. Naseem, K. Chaudhary, B. Sharma, and A. G. Lal, "Using Ensemble Decision Tree Model to Predict Student Dropout in Computing Science," in *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Dec. 2019, pp. 1–8. doi: 10.1109/CSDE48274.2019.9162389.

[6] P. Berka and L. Marek, "Bachelor's degree student dropouts: Who tend to stay and who tend to leave?," *Stud. Educ. Eval.*, vol. 70, no. January, 2021, doi: 10.1016/j.stueduc.2021.100999.

[7] J. Y. Chung and S. Lee, "Dropout early warning systems for high school students using machine learning," *Child. Youth Serv. Rev.*, vol. 96, pp. 346–353, Jan. 2019, doi: 10.1016/j.childyouth.2018.11.030.

[8] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sanchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEE Access*, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[9] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electron.*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030457.

[10] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Appl. Sci.*, vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/app11073130.

[11] S. C. Tsai, C. H. Chen, Y. T. Shiao, J. S. Ciou, and T. N. Wu, "Precision education with statistical learning and deep learning: a case study in Taiwan," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-00186-2.

[12] L. J. Rodríguez-Muñiz, A. B. Bernardo, M. Esteban, and I. Díaz, "Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?," *PLoS One*, vol. 14, no. 6, pp. 1–20, 2019, doi: 10.1371/journal.pone.0218796.

[13] D. E. Moreira da Silva, E. J. Solteiro Pires, A. Reis, P. B. de Moura Oliveira, and J. Barroso, "Forecasting Students Dropout: A UTAD University Study," *Futur. Internet*, vol. 14, no. 3, pp. 1–14, 2022, doi: 10.3390/fi14030076.

[14] M. Albán, "Contribuciones a la predicción de la deserción universitaria a través de minería de datos," 2019.

[15] A. Bernardo, M. Esteban, A. Cervero, R. Cerezo, and F. J. Herrero, "The Influence of Self-Regulation Behaviors on University Students' Intentions of Persistance," *Front. Psychol.*, vol. 10, no. October, pp. 1–8, 2019, doi: 10.3389/fpsyg.2019.02284.

[16] E. Tuero Herrero, A. Cervero, M. Esteban, and A. Bernardo, "Why do university students drop out? influencing variables regarding the

approach and consolidation of drop out," *Educ. XX1*, vol. 21, no. 2, pp. 131–154, 2018, doi: 10.5944/educXX1.20066.

[17] J. R. Casanova, R. Vasconcelos, A. B. Bernardo, and L. S. Almeida, "University dropout in engineering: Motives and student trajectories," *Psicothema*, vol. 33, no. 4, pp. 595–601, 2021, doi: 10.7334/psicothema2020.363.

[18] C. Truta, L. Parv, and I. Topala, "Academic engagement and intention to drop out: Levers for sustainability in higher education," *Sustain.*, vol. 10, no. 12, pp. 1–11, 2018, doi: 10.3390/su10124637.

[19] L. Bardach, M. Lüftenegger, S. Oczlon, C. Spiel, and B. Schober, "Context-related problems and university students' dropout intentions—the buffering effect of personal best goals," *Eur. J. Psychol. Educ.*, vol. 35, no. 2, pp. 477–493, 2020, doi: 10.1007/s10212-019-00433-9.

[20] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 104–112, 2020, doi: 10.1080/02673843.2019.1596823.

[21] L. Bäulke, C. Grunschel, and M. Dresel, "Student dropout at university: a phase-orientated view on quitting studies and changing majors," *Eur. J. Psychol. Educ.*, 2021, doi: 10.1007/s10212-021-00557-x.

[22] N. Tieben, "Non-completion, Transfer, and Dropout of Traditional and Non-traditional Students in Germany," *Res. High. Educ.*, vol. 61, no. 1, pp. 117–141, 2020, doi: 10.1007/s11162-019-09553-z.

[23] R. Benites, "La Educación Superior Universitaria en tiempos del COVID-19," *Estado Int. S.a*, p. 1, 2020, [Online]. Available: https://www.estadointernacional.com/la-educacion-superior-universitaria-en-tiempos-del-covid-19/

[24] V. Galán, "Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario," *Bibl. la Univ. Carlos III Madrid*, p. 120, 2015, [Online]. Available: https://e-archivo.uc3m.es/handle/10016/22198

[25] IBM, "False positive rate (FPR)," *IBM*, 2022, [Online]. Available: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.6.x?topic=overview-false-positive-rate-fpr