Predicting Academic Performance in a Subject Using Classifier Algorithms

Edwar Abril Saire Peralta[1], Giovanni Rolando Cabrera Málaga[1] and Sonia Benilda Calloapaza Pari[1]

[1] *Universidad Nacional de San Agustín de Arequipa, Santa Catalina 117, Arequipa Perú*

Abstract

The objective of this research is to determine the academic performance of students starting a Systems Engineering course. The discrete structure I course which is considered a course that is difficult to pass. The population is represented by 827 students, the research was approached from a quantitative approach, non-experimental design and at a correlational level. The methodology implemented is CRISP-DM (Cross Industry Standard Process for Data Mining) through the supervised learning technique using binary classification models based on random forest algorithms, xgboost and support vector machines. The results have allowed predicting if a student will pass or fail the course. The classification models that have shown the best results are based on random forest and xgboots algorithms with an accuracy of 82.5%.

Keywords

Academic performance, Data Mining, Classification Algorithms, Supervised Learning, Data Mining.

1. Introduction

Academic performance has been a concern in university education for many years. The biggest challenge has always been to provide quality education, which means to improve academic performance [1]. One of the consequences of low academic performance of students is failing courses, and one of the ways to solve the problem is to analyze the academic background of the most influential data of students entering the university by data mining.

The transition from high school to college can be a difficult transition for entering students. In Engineering of Systems of the National University of San Agustin de Arequipa - Peru, the course of discrete structures I has shown according to statistics provided by the same career, that on average 50% of entrants have failed the course in their first enrollment, being considered a course that shows difficulty to be approved. It is necessary to explore and analyze with what tendency of academic performance students enter the university.

Data mining allows to explore, analyze and find patterns in the data obtaining useful information with the objective of understanding the performance and the environment where the student performs [2]. The results of finding patterns of behavior through the application of data mining, allows decision making to solve problems in educational settings [3].

Both [4] and [5] point out that academic performance is influenced by a set of internal and external factors of the student, where the final result of the performance obtains a quantitative value, reflected in the status of the courses with labels of passed and failed. Furthermore [6] indicate that the numerical average obtained in a course is the most accurate signal of a student's academic achievement.

CISECT 2023: International Congress on Education and Technology in Sciences, December 04–06, 2023, Zacatecas, Mexico

esaire@unsa.edu.pe (E. A. Saire Peralta); gcabrerama@unsa.edu.pe (G. R. Cabrera Málaga); scalloapaza@unsa.edu.pe (S. B. Calloapaza Pari)

0000-0002-9526-0205 (E. A. Saire Peralta); 0000-0003-3257-5127 (G. R. Cabrera Málaga); 0000-0003-0684-2623 (S. B. Calloapaza Pari)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Most of the research on academic performance has been approached using supervised classification algorithms, where they point out that models can be built which can learn from experiences (historically recorded experiences), and the more experiences the model improves in its predictive learning [7]. [8] point out that supervised learning allows finding trends based on behavioral patterns, which have been obtained from large amounts of data.

The present research aims to predict the status of the discrete structure course I (pass or fail), applying the CRISP-DM methodology through supervised classification algorithms. The final result of the research will allow to identify in advance if a student passes or fails the course before the beginning of the semester. Identifying in advance who will fail the course will allow alerting teachers and educational authorities to take tutoring actions to improve academic performance.

2. Related works

[9] have developed an investigation to predict students' academic performance based on academic, demographic and sociodemographic data. The algorithms used are decision tree, K-Nearest-Neighbor, support vector machines and naive bayes implemented with the Python programming language. The research approach is quantitative and has worked with 4738 students of Industrial Engineering and Electronic Engineering. The data were collected from 2008 to 2018, with 324 variables for each student. In view of so many variables, the variables that have the most influence on academic performance have been selected. Finally, the algorithm that showed the best results was KNN with an accuracy ranging from 78.5% to 80%. The reduction techniques of the most influential variables in the academic performance is a determinant work in the prediction. A strength of this work is the number of attributes and records available.

[10] developed a model to predict public college dropout in COVID-19 time. Their goal was to determine the most efficient Machine Learning algorithm that could classify students based on historical data from 2018 to 2021. They applied the algorithms to a population of 652 students with 106 variables with a descriptive type of research. In the end it was obtained as a result that the K-Nearest-Neighbor algorithm found better results with an accuracy of 91% having as inputs data related to the academic and socioeconomic aspect. With the results found, it was concluded that the model is useful to predict early, in the first semesters, who are the possible university students that could drop out. The dropout diagnosis shows early warnings for the university, so that it can support these students with tutoring or other academic programs in favor of the students. Reducing the dimension of 106 variables to the most significant ones indicates that the model works with the most influential variables, which is a decisive contribution to the model presented in the context presented.

[11] investigated the main predictor variables that influence the academic performance of students after six semesters have elapsed since they entered university. They worked with 622 students and applied twelve classification algorithms, where an ensemble was used based on the algorithms that showed the best results in the values of their metrics, which are logistic regression, naive Bayes and support vector machines. When applying the ensemble with optimal cut-off point, a specificity of 0.695 and a sensitivity of 0.947 were obtained. The grade obtained in mathematics was a determining factor and sociodemographic factors had no influence. An important fact in this research is that many of the sociodemographic variables did not have a strong influence on the result. The score obtained in the university entrance exam was not taken into account, which is something that is striking, since, in other research, it represents a decisive variable in academic performance.

[12] developed a model based on supervised machine learning with the purpose of predicting whether a student passes the leveling course. They used Gradient Boosting and Logistic Regression algorithms, where the inputs were the predictor variables grouped into demographic, socioeconomic, family, institutional and academic performance in the application. The population consisted of 7139 students. With the first algorithm, an accuracy of 96% was obtained in the cross-validation and 89% for predicting new data. The logistic regression algorithm indicates that the average grade of the first bimester, the average grade with which the student entered the

university and his geographical location of origin, among others, do affect the probability that the student will pass the course. Meanwhile, the variables that have determined that a student fails the course are the grade obtained when entering the university, the province of origin and the lack of academic support or tutoring. There remains the possibility of testing other machine learning algorithms to see their accuracy and verify which would be the most influential attributes in determining whether or not a student passes or fails the course.

[13] developed models with predictive ability of student academic risk, using educational data mining, for early detection of academic risk. In this research, sociodemographic data and the results of university entrance exams of 415 students of computer science majors enrolled between the years 2016 and 2019 were applied. The best classification model was based on the LMT algorithm with an accuracy of 75.42% and a value of 0.805 for the area under the ROC curve. The data that have shown the most influence, such as college entrance exam score, were identified. The research began with 65 attributes and when determining the most significant ones, 9 variables remained, since this depends on the quality of the data and the predictive power they have in relation to the target variable.

3. Application of the methodology

The research has had a quantitative approach and has worked with 778 students, where the CRISP-DM data mining methodology has been applied. Figure 1 shows the outline of the model to be applied in the research based on the data mining methodology.

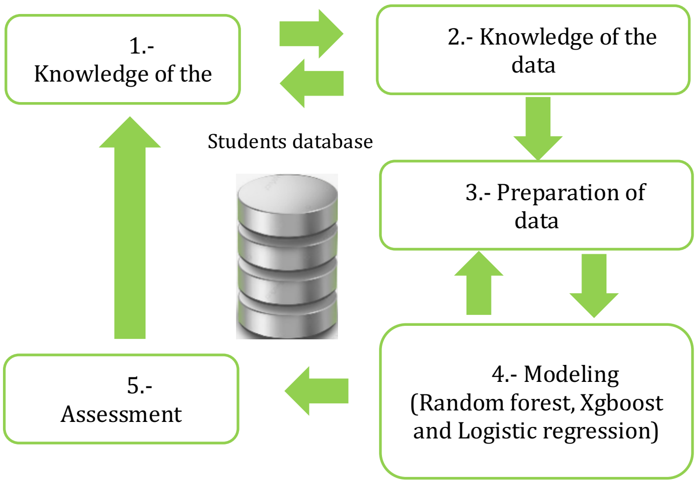

1.- Knowledge of the
2.- Knowledge of the data
Students database
3.- Preparation of data
4.- Modeling (Random forest, Xgboost and Logistic regression)
5.- Assessment

Figure 1: Stages of the CRISP-DM Methodology

Note: Source: Schematic generated based on [14].

The data used in the research are shown in Table 1. The proposed predictive model has as input the admission data, academic data and other data that have been calculated such as age at high school graduation, time elapsed before entering college and age at college entrance.

Table 1
Data input and output

I/O	Items
Input data	□ Admission Data □ Academic Data □ Calculated data
Output data	□ Classification of students: Pass/Fail

Below, in Figure 2 we can see the scheme proposed in the research.

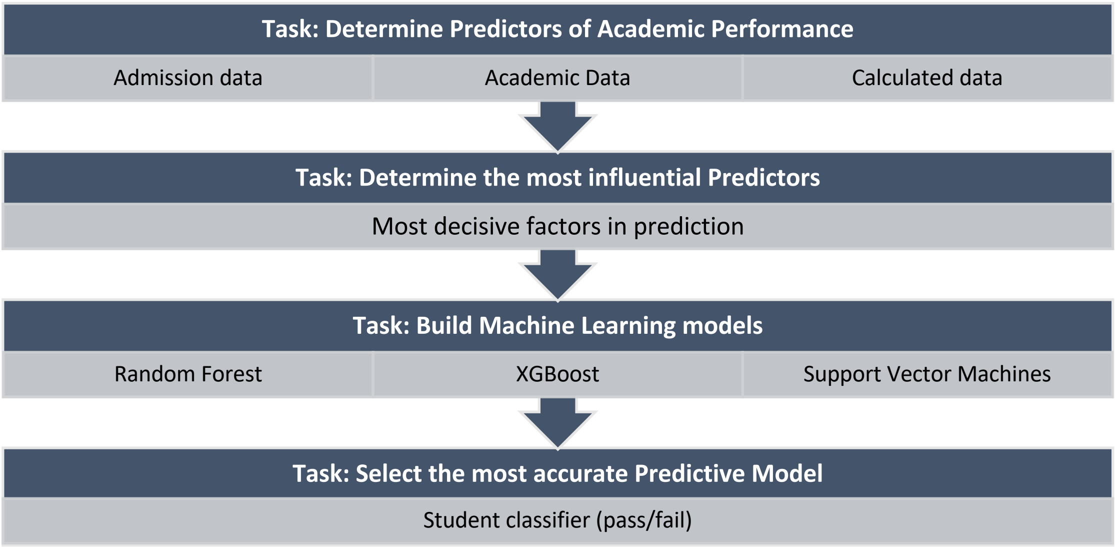

Task: Determine Predictors of Academic Performance
Admission data
Academic Data
Calculated data
Task: Determine the most influential Predictors
Most decisive factors in prediction
Task: Build Machine Learning models
Random Forest
XGBoost
Support Vector Machines
Task: Select the most accurate Predictive Model
Student classifier (pass/fail)

Figure 2: Research approach

3.1. Understanding the business

Universities have three objectives, which are teaching, research and social responsibility. Both licensing and accreditation contribute to achieving quality education. Entering students travel a difficult path from college to university, which must be gradual in order for them to adapt. The task of adaptation should be considered a priority for the university, since it must know what the student is facing in the first semesters.

Knowing in advance what the academic performance of incoming students will be is uncertain. The evaluation of academic performance is classified in this research by labeling the student as pass or fail. The problem of the present investigation is the lack of knowledge about their possible academic performance in the course of discrete structure I, of the systems career of the Universidad Nacional de San Agustin de Arequipa. According to data provided by the School of Systems, statistics show that from 2011 to 2020 there has been an average of approximately 50% of students who failed their first enrollment in the course. In Table 1 we can see the percentage of passed and failed students from 2011 to 2020.

Table 1
Pass and fail percentages by year

Year	Passed	Failed
2011	47%	53%
2012	41%	59%
2013	36%	64%
2014	40%	60%
2015	36%	64%
2016	53%	47%
2017	53%	47%
2018	65%	35%
2019	62%	38%
2020	69%	31%
	50.2%	49.8%

According to data provided by the school of systems in the discrete structures I course, there were 137 students who dropped out because they were never able to pass the discrete structures

I course, even though they tried to pass by taking the course up to 3 times. The population for the present research is made up of students from the graduating classes from 2011 to 2020, which consists of a total of 778 students.

3.2. Understanding the data

The data requested for the project come from two sources, the first source is related to the admission data of the students entering the systems career and the second source is related to the academic data of the students of the School of Systems who have enrolled. The data provided are shown in Table 2.

Table 2
Admission and academic data

ADMISSION DATA		
N°	Attribute	Description
1	Last Name and First Name	Last Name and First Name of student
2	Gender	Student's gender
3	Date of birth	Date of birth of student
4	Place of Birth (Department)	Department where the student was born
5	Place of birth (Province)	Province where the student was born
6	Place of birth (District)	Province where the student was born
7	Place of birth (code)	Place of birth (code)
8	School	Origin of high school
9	School code	Code of school
10	Location of school (Department)	Department where school is located
11	Location of school (Province)	Province where the school is located
12	Location of school (District)	District where school is located
13	Type of school	Type of school
14	Year of school leaving	Year of graduation from school
15	Admission mode	University entrance mode
16	Score	University entrance score
17	Extraordinary admission	Extraordinary mode of admission
ACADEMIC DATA		
N°	Attribute	Description
1	CUI	Student code
2	Entrance code	University entrance code
3	Last name and first name	Last name and first name of student
4	Course	Course in which the student is enrolled
5	Grade	Grade achieved in the course
6	Condition	Student's condition
7	#Enrollment	Number of enrollment

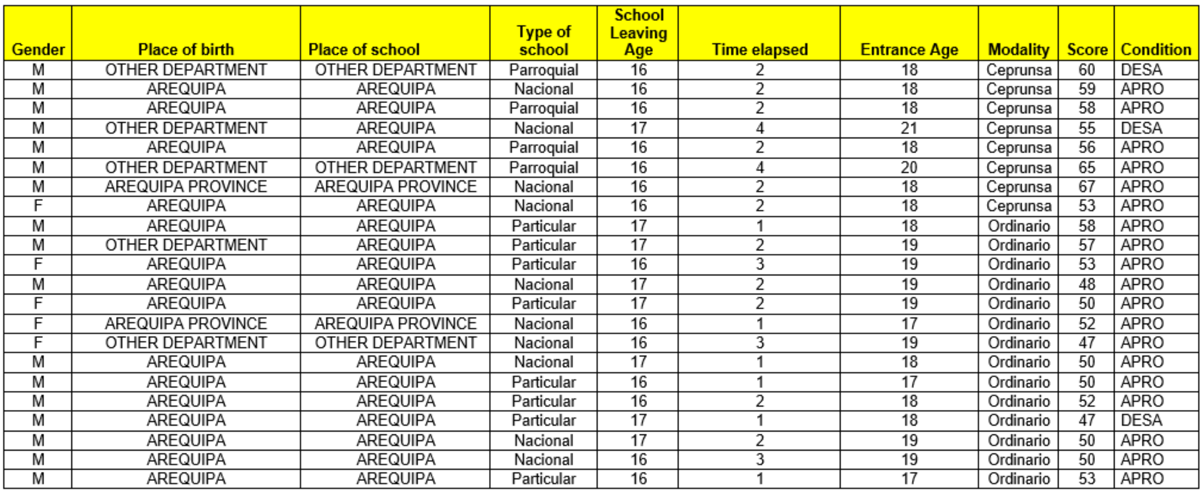
3.3 Data preparation

From all the attributes provided by the university, those to be used in the prediction models have been selected. We have excluded data that do not have any contribution, such as the student's entrance code, last names and first names, among others. New attributes have also been generated. Table 3 shows the final data that will be used to train the models.

Table 3
Final Data

	Data	Description
1	Gender	Gender of student
2	Placebirth	Place of birth
3	PlaceSchool	Place of school
4	Typeschool	Type of school
5	School leaving age	Age of leaving school
6	ElapsedTime	Elapsed time from school to university
7	AgeEntrance	Age of university entrance
8	Modality	University entrance modality
9	score	University entrance test score
10	Condition	Pass/Fail Condition

Figure 3 shows a view of the data to be used in the classification algorithms. The status column is the objective to be predicted (pass or fail the course) and the other columns are the predictor attributes. For the modeling stage the status column will only take two values which is represented by DESA (fail) and APRO (pass).



Gender	Place of birth	Place of school	Type of school	School Leaving Age	Time elapsed	Entrance Age	Modality	Score	Condition
M	OTHER DEPARTMENT	OTHER DEPARTMENT	Parroquial	16	2	18	Ceprunsa	60	DESA
M	AREQUIPA	AREQUIPA	Nacional	16	2	18	Ceprunsa	59	APRO
M	AREQUIPA	AREQUIPA	Parroquial	16	2	18	Ceprunsa	58	APRO
M	OTHER DEPARTMENT	AREQUIPA	Nacional	17	4	21	Ceprunsa	55	DESA
M	AREQUIPA	AREQUIPA	Parroquial	16	2	18	Ceprunsa	56	APRO
M	OTHER DEPARTMENT	OTHER DEPARTMENT	Parroquial	16	4	20	Ceprunsa	65	APRO
M	AREQUIPA PROVINCE	AREQUIPA PROVINCE	Nacional	16	2	18	Ceprunsa	67	APRO
F	AREQUIPA	AREQUIPA	Nacional	16	2	18	Ceprunsa	53	APRO
M	AREQUIPA	AREQUIPA	Particular	17	1	18	Ordinario	58	APRO
M	OTHER DEPARTMENT	AREQUIPA	Particular	17	2	19	Ordinario	57	APRO
F	AREQUIPA	AREQUIPA	Particular	16	3	19	Ordinario	53	APRO
M	AREQUIPA	AREQUIPA	Nacional	17	2	19	Ordinario	48	APRO
F	AREQUIPA	AREQUIPA	Particular	17	2	19	Ordinario	50	APRO
F	AREQUIPA PROVINCE	AREQUIPA PROVINCE	Nacional	16	1	17	Ordinario	52	APRO
F	OTHER DEPARTMENT	OTHER DEPARTMENT	Nacional	16	3	19	Ordinario	47	APRO
M	AREQUIPA	AREQUIPA	Nacional	17	1	18	Ordinario	50	APRO
M	AREQUIPA	AREQUIPA	Particular	16	1	17	Ordinario	50	APRO
M	AREQUIPA	AREQUIPA	Particular	16	2	18	Ordinario	52	APRO
M	AREQUIPA	AREQUIPA	Particular	17	1	18	Ordinario	47	DESA
M	AREQUIPA	AREQUIPA	Nacional	17	2	19	Ordinario	50	APRO
M	AREQUIPA	AREQUIPA	Nacional	16	3	19	Ordinario	50	APRO
M	AREQUIPA	AREQUIPA	Particular	16	1	17	Ordinario	53	APRO

Figure 3: Data provided by the university

3.4. Modeling

The data flow to build the classifier predictive model are shown in Figure 4, which allows predicting whether a student passes or fails the course, for which there are nine inputs and one output.

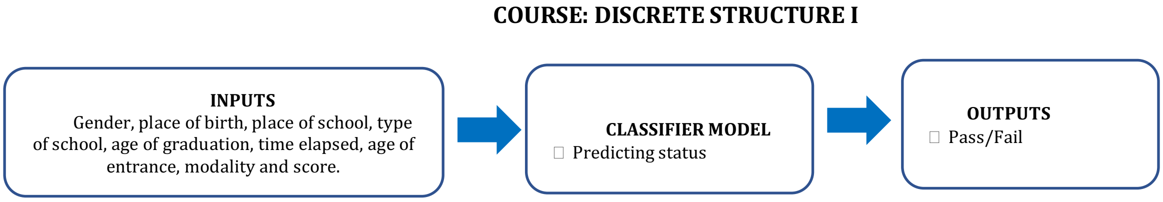
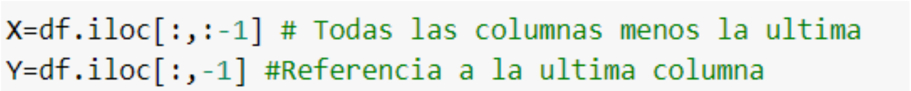

COURSE: DISCRETE STRUCTURE I
INPUTS
Gender, place of birth, place of school, type of school, age of graduation, time elapsed, age of entrance, modality and score.
CLASSIFIER MODEL
□ Predicting status
OUTPUTS
□ Pass/Fail

Figure 4: Data provided by the university

The classification models were implemented, where the dataset was uploaded to a Google Collaboraty repository in CSV format. If the proposed model does not show the best values in its metrics/indicators, the alternative is to return to the initial phases iteratively, until the most appropriate metric values for the model are achieved. The tasks that have been executed in the Google Colaboraty environment with Python are the next ones:

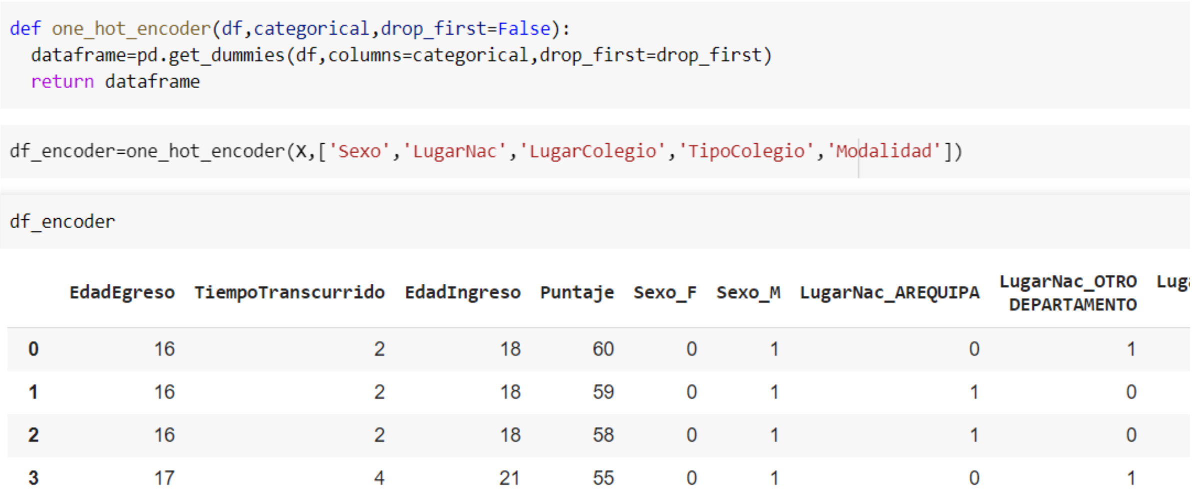
- Separate from the total columns or variables, which are the predictor variables and which is the target variable. The X variable represents the predictor variables, while the Y variable represents the objective variable, as shown in Figure 5.



```
X=df.iloc[:,:-1] # Todas las columnas menos la ultima
Y=df.iloc[:,-1] #Referencia a la ultima columna
```

Figure 5: Predictor and objective variables

- In the initial test, no balancing techniques were applied, because the data were found to be 50% balanced in both groups (pass and fail students in course 1 of their first enrollment).
- Converting categorical variables to dummy variables. The categorical variables to be converted are: sex, place of birth, place of school, type of school and mode of entry, as shown in Figure 6.



```
def one_hot_encoder(df,categorical,drop_first=False):
  dataframe=pd.get_dummies(df,columns=categorical,drop_first=drop_first)
  return dataframe
```

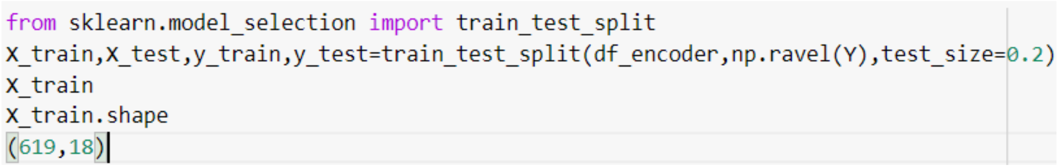
```
df_encoder=one_hot_encoder(X,['Sexo','LugarNac','LugarColegio','TipoColegio','Modalidad'])
```

```
df_encoder
```

	EdadEgreso	TiempoTranscurrido	EdadIngreso	Puntaje	Sexo_F	Sexo_M	LugarNac_AREQUIPA	LugarNac_OTRO DEPARTAMENTO	Lug
0	16	2	18	60	0	1	0	1	
1	16	2	18	59	0	1	1	0	
2	16	2	18	58	0	1	1	0	
3	17	4	21	55	0	1	0	1	

Figure 6: Dummies Variables

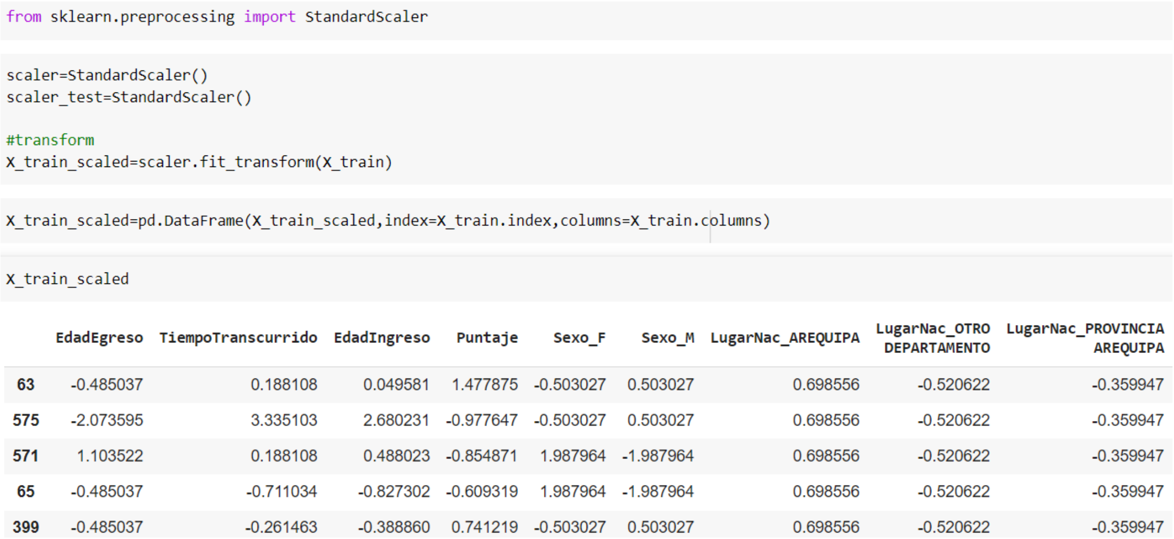
- Split data for training (80%) and data for validating the model (20%). We use the Split function of python which allows us to split the data, the value of 0.2 in the variable test_size represents 20% for testing the model and the remainder or complement represents 80% for training the model, as shown in Figure 7.



```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(df_encoder,np.ravel(Y),test_size=0.2)
X_train
X_train.shape
(619,18)
```

Figure 7: Split of the training and test data

- Scaling the data, has the objective of transforming the values of the features so that they are within a range domain. In the research it was necessary to scale the data, because there are machine learning algorithms that have problems when finding outliers or values that show bias. In Figure 8 we can see the results of the data scaling process



```
from sklearn.preprocessing import StandardScaler
```

```
scaler=StandardScaler()
scaler_test=StandardScaler()

#transform
X_train_scaled=scaler.fit_transform(X_train)
```

```
X_train_scaled=pd.DataFrame(X_train_scaled,index=X_train.index,columns=X_train.columns)
```

```
X_train_scaled
```

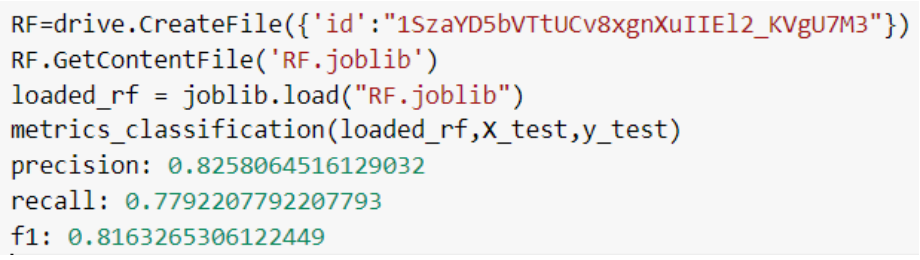
	EdadEgreso	TiempoTranscurrido	EdadIngreso	Puntaje	Sexo_F	Sexo_M	LugarNac_AREQUIPA	LugarNac_OTRO DEPARTAMENTO	LugarNac_PROVINCIA AREQUIPA
63	-0.485037	0.188108	0.049581	1.477875	-0.503027	0.503027	0.698556	-0.520622	-0.359947
575	-2.073595	3.335103	2.680231	-0.977647	-0.503027	0.503027	0.698556	-0.520622	-0.359947
571	1.103522	0.188108	0.488023	-0.854871	1.987964	-1.987964	0.698556	-0.520622	-0.359947
65	-0.485037	-0.711034	-0.827302	-0.609319	1.987964	-1.987964	0.698556	-0.520622	-0.359947
399	-0.485037	-0.261463	-0.388860	0.741219	-0.503027	0.503027	0.698556	-0.520622	-0.359947

Figure 8: Scaler data

□ Same conversion procedure was done for the test data, which is represented by X_test, which is data that the model has never seen and will be used to validate the model.

The following is a summary of the tests that have been performed with some classification algorithms. After several experiments and tests it has been determined that the best result has been achieved by using the first 8 variables shown with the mutual information technique. The models were experimented, first with four more determinant or significant variables (college entrance score, college entrance age, time elapsed since leaving school until entering college and finally the age at which they left school), and because it did not show results of improvement in the prediction, it was experimented with three variables, four, five, six, seven, eight, nine, etc. until a line that determines the point of improvement could be found. After several attempts, it was determined that the first eight variables showed the best predictions, reflected in the values of the metrics.

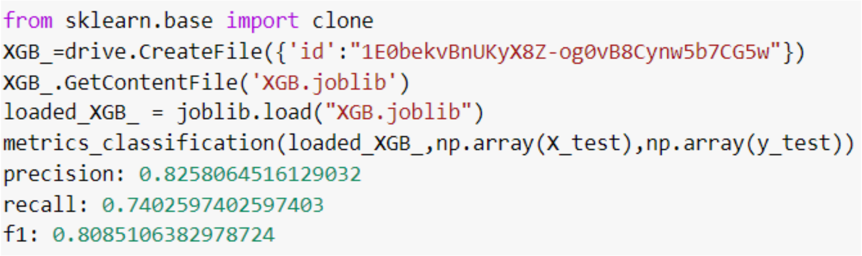
We worked several times in the search for the most suitable values for the hyperparameters, using Bayesian Optimization, which is very similar to GridSearch. We worked and tested again with the random forest, xgboost and support vector machines algorithms, with the eight predictors and the hyperparameters found, we retrained the models. The source code of the tested algorithms is shown below. Figure 9 shows the random forest algorithm.



```
RF=drive.CreateFile({'id':"1SzaYD5bVTtUCv8xgnXuIIEl2_KVgU7M3"})
RF.GetContentFile('RF.joblib')
loaded_rf = joblib.load("RF.joblib")
metrics_classification(loaded_rf,X_test,y_test)
precision: 0.8258064516129032
recall: 0.7792207792207793
f1: 0.8163265306122449
```

Figure 9: Random Forest

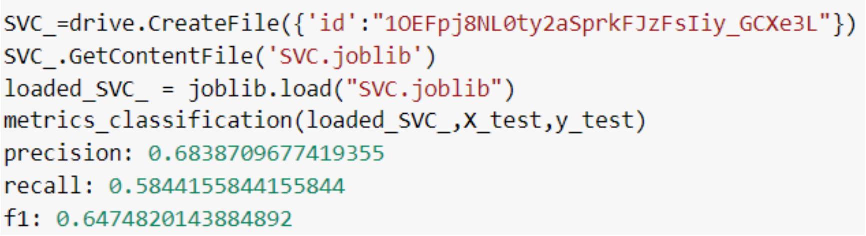
Figure 10 shows the xgboost algorithm.



```
from sklearn.base import clone
XGB_=drive.CreateFile({'id':"1E0bekvBnUKyX8Z-og0vB8Cynw5b7CG5w"})
XGB_.GetContentFile('XGB.joblib')
loaded_XGB_ = joblib.load("XGB.joblib")
metrics_classification(loaded_XGB_,np.array(X_test),np.array(y_test))
precision: 0.8258064516129032
recall: 0.7402597402597403
f1: 0.8085106382978724
```

Figure 10: Xgboost

In Figure 11 we can see the SVC (Support Vector Machines) algorithm.



```
SVC_=drive.CreateFile({'id':"1OEFpj8NL0ty2aSprkFJzFsIiy_GCXe3L"})
SVC_.GetContentFile('SVC.joblib')
loaded_SVC_ = joblib.load("SVC.joblib")
metrics_classification(loaded_SVC_,X_test,y_test)
precision: 0.6838709677419355
recall: 0.5844155844155844
f1: 0.6474820143884892
```

Figure 11: Vector Support Machines

3.5. Validation of the approach

All the proposed supervised models that have been implemented in classification tasks were evaluated, the values of the performance metrics of the models were taken into account and the most optimal classifier model was selected. The evaluation of the obtained models will verify the efficiency with the test data, which represents 20% of the total, in other to say, those data that were separated and that the obtained model does not know and was not taken into account in the training of the models. In order to validate the models presented in this research, two aspects have been taken into account, which are cross-validation (training data) and testing with the test. Metrics represent values to measure the efficiency of a classifier model.

3.5.1. Internal validation

Internal validation refers to cross-validation, which is the training that the models underwent. The accuracy metric of the models has shown values between 0 and 1, at the beginning the values were from 0.4 to 0.6 depending on the model. However, after many tests and experiments we have been able to reach up to 0.82. In Table 4 we can see the test results in the training stage.

Table 4
Metric values

Algorithm/Model	Accuracy	Recall	F1
Random Forest	0.8511	0.7877	0.8260
XGBoost	0.8477	0.7611	0.8112
Support Vector Machines	0.7030	0.6047	0.6533

In this testing stage the results showed that random forest is the algorithm that has shown the best results in predicting if a student passes or fails the Discrete Structures I course in his first enrollment.

3.5.2. External validation

The external validation consisted of testing if the model works with new data, this has been tested by entering the data that were initially separated, which corresponds to 20% of records. The tests indicated that the algorithms that have shown the highest prediction accuracy are the model implemented with random forest and the model implemented with xgboost with an accuracy of 82.5% as shown in Table 5. It is important to highlight that the values of the metrics vary, due to the fact that the training and test data are selected randomly.

Table 5
Metric values

Algorithm/Model	Accuracy	Recall	F1
Random Forest	0.8258	0.7779	0.8163
XGBoost	0.8258	0.7402	0.8085
Support Vector Machines	0.6838	0.5844	0.6474

4. Results and Discussion

Based on the research developed and the results obtained, the subfield of Machine Learning related to supervised learning has shown great advances when applied in the field of education, not only to predict the academic performance of students, but also to predict student desertion, student dropout, learning patterns, among others, as shown in the literature consulted.

The reduction of dimensionality through the technique of mutual information and permutation of the random forest algorithm have allowed improving the results, showing that the most determinant variables in this context are college entrance score, age of graduation from high school, time elapsed and age of university entrance among the admission data. The research of [15] approached the aspect of dimensionality and determined that the most influential variable was the age at which they began their studies, which is a result that is common to this research. However, there are other investigations such as that of [16] which, by collecting historical data from a public institution and applying the decision tree algorithm, has shown that the admission score was not significant in the prediction of academic performance, since other variables to be considered were present, such as credits approved in relation to theoretical credits that should have been approved; these changes are due to the fact that other additional data were available, in comparison with the present investigation, where the score was the most determining variable in the prediction.

Another research with which we can compare is that of [17], which also worked with historical data from a public institution, and determined that the number of failed courses and the level of education of the father were determinant. These comparisons are mentioned because it is different to work with historical data that the educational institution has been recording without the intention of using it in research, compared to those institutions that have planned it for research purposes, which increases the richness of the results.

In the literature it has been found that predictions are sought by classifying students as pass/fail, dropout/non-dropout, low performance/high performance among others, using algorithms such as neural networks, random forest, decision trees, support vector machines, logistic regression among others, seeking the best prediction accuracy as seen in the research of [14], [18], [19] and [20] where they have managed to obtain predictions with an average accuracy of 80%, even with more data compared to the research presented here. The literature reviewed regarding this line of research and educational contexts, shows predictions of binary classification of an object or event, however, the contribution of the present research work lies in predicting whether a student passes or fails the course of discrete structures I with few attributes which were not intended to be collected for research purposes, in addition to having few attributes, for which it has been necessary to filter and select the most decisive attributes to achieve better results.

5. Conclusions

The model that achieved the highest accuracy was implemented with the random forest algorithm with an accuracy of 82.5% when tested with the test data and achieved an accuracy of 85% when tested with the training data. The quality of the data are determinant to achieve a higher accuracy in the predictions of the models. The present research worked with data that were not intended for that purpose, however, they were used and positive results have been found in the use of them based on trial and error, looking for the most influential attributes and also looking for the best values for the hyperparameters of the algorithms. Another important aspect in the training of the models is the small amount of records, which is determinant for an optimal training of the models, which could bring as a consequence low accuracies in the models. It is important to have balanced data for model training, since this way we will avoid developing biased models and achieve reliable model predictions. Finally, the most important conclusion that has been deduced is that the quality of the attributes related to the subject to be predicted is determinant for the success of the classification models.

References

[1] V. Kumar & A. Chadha (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. International Journal of Advanced Computer Science and Applications, 2 (3):80-84.

[2] M. Ramaswami (2009). A Study on Feature Selection Techniques in Educational Data Mining. International Working Group On Educational Data Mining, Vol. 1, Issue 1.

[3] C. Heiner, R. Baker & K. Yacef (2006). Proceedings of Educational Data Mining workshop. 8th International Conference on Intelligent Tutoring Systems.

[4] A. Pérez-Luño, J. Ramón Jerónimo & J. Sánchez Vázquez, "Análisis exploratorio de las variables que condicionan el rendimiento académico " Sevilla, España: Universidad Pablo de Olavide, 2000.

[5] M. Vélez Van & N. Roa, "Factors associated with academic performance in medical students" PSIC. Educación Médica, 2005.

[6] S. Rodriguez, F. Eva and T. Mercedes, "Rendimiento académico en la transición secundaria-universidad" Revista de Educación, 2003. http://hdl.handle.net/11162/67356

[7] J. Bell, "Machine learning: hands-on for developers and technical professionals," (J. W. & Sons. (Ed.); Second Edi), 2020.

[8] J D. Kelleher, B. Mac Namee and A. D'arcy, "Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies," (M. Press. (Ed.)), 2020.

[9] L. Contreras Bravo, N. Nieves-Pimiento & K. Gonzalez-Guerrero (2023). Predicción del rendimiento académico universitario mediante mecanismos de aprendizaje automático y métodos supervisados. Ingeniería, 1, 1–25. https://doi.org/https://doi.org/10.14483/23448393.19514

[10] J. Valero Cajahuanca, A. Navarro Raymundo, A. Larios Franco & J. Julca Flores (2022). Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción. Revista de Ciencias Sociales, 28(3), 362–375. https://doi.org/10.31876/rcs.v28i3.38480

[11] J. E. Gamboa Unsihuay & J. W. Salinas Flores (2022). Predicción de la situación académica en alumnos se pregrado usando algoritmos de Machine Learning. Perfiles, 1(27), 4–10. https://doi.org/10.47187/perf.v1i27.142

[12] K. Calva, M. Flores & H. Porras (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado. Latin American Journal of Computing, VIII(1). https://doi.org/10.5281/zenodo.5770905

[13] E. Ayala Franco, R. E. López Martínez & V. H. Menéndez Domínguez (2021). Modelos predictivos de riesgo académico en carreras de computación con minería de datos

educativos. Revista de Educación a Distancia (RED), 21(66), 1–36. https://doi.org/10.6018/red.463561

[14] P. Chapman, C. Julian, K. Randy, K. Thomas, R. Thomas & S. C. Wirth (1999). CRISP-DM 1.0: Step-by-step data mining guide.

[15] A. Reinoso Quijo (2023). Desarrollo de un modelo para predecir el rendimiento académico de estudiantes de la EPN en base a su nivel de acceso a TICS y factores socioeconómicos. [tesis maestria, Escuela Politecnica Nacional, Quito]. http://bibdigital.epn.edu.ec/handle/15000/23615

[16] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén & V. Cornejo-Aparicio (2020b). Analysis of the academic performance of systems engineering students, desertion possibilities and proposals for retention. Ingeniare, 28(4), 668–683. https://doi.org/10.4067/S0718-33052020000400668

[17] L. Quiñones & Y. L. Carrasco (2020). Rendimiento académico empleando minería de datos. Espacios, 41(44), 277–285. https://doi.org/10.48082/espacios-a20v41n44p17

[18] P. Mejía Zamora (2023). Modelo matemático para predecir el grado de deserción de los estudiantes en el Instituto Superior Tecnológico Bolívar [[tesis de maestria, Universidad Técnica de Ambato, Ecuador]]. In Repositorio Institucional de la Universidad Técnica de Ambato. https://repositorio.uta.edu.ec/bitstream/123456789/37204/1/t2153mma.pdf

[19] A. J. Camargo (2020). Modelo Para La Predicción De La Deserción De Estudiantes De Pregrado, Basado En Técnicas De Minería De Datos. Universidad de la Costa - Pregrado.

[20] H. E. C Gismondi (2021a). Modelo predictivo basado en machine learning como soporte para el seguimiento académico del estudiante universitario. tesis doctor, Universidad Nacional del Santa, Perú. https://hdl.handle.net/20.500.14278/3804