Predicting Academic Performance Path Using Classification Algorithms

Edwar Abril Saire-Peralta* and Maria del Carmen Córdova-Martínez

***Abstract*—The objective of this research is to determine the academic performance route of students entering the Systems Engineering program. The academic performance route is defined by three courses, which develop sequentially in the first semesters, where students show difficulty to be approved. The population is represented by 827 students, the research was approached from a quantitative approach, the research design is non-experimental and the scope or level of research is correlational. The methodology implemented is CRISP-DM (Cross Industry Standard Process for Data Mining) using machine learning algorithms, through binary classification models using logistic regression algorithms, random forests and XGboost. The results have allowed predicting whether a student would pass or fail in each of the courses, determining their academic performance path. The classification models have been able to achieve an accuracy between 87% and 93%.**

***Index Terms*—Classification algorithms, supervised learning, data mining, academic performance**

I. Introduction

Pérez-Luño *et al*. [1, 2] indicate that academic performance is influenced by a set of factors specific to the student, where performance is assigned a quantitative value, which is reflected in the subjects passed and failed. On the other hand, Rodriguez *et al*. [3] indicate that student's grades represent the most accurate indicator of the achievements obtained, which is influenced by a set of personal and social aspects, among others. According to Vargas [4], there are different aspects associated with academic performance, including both internal and external components of the individual. Espinar [5] indicates that the score is an excellent predictor indicator of academic performance in college, since this factor has a determining weight to understand and achieve an explanation to the fact.

According to the aforementioned authors, they state that the final average obtained at the end of the course is a consequence of the final result of the course and not of the teaching and learning process. There is a history of students who at the beginning of university fail courses, drop out of courses and possibly end up dropping out of the degree course. Knowing in advance what the academic performance of entering students will be is uncertain. The outcome of academic performance is classified by labeling the student as passing or failing. Research to analyze academic performance has been addressed by machine learning algorithms, which is a subfield of artificial intelligence, which allows building systems with the ability to learn from the experiences (data) with which it is trained, where the system is improving according to experiences [6]. Kelleher *et al*. [7] indicate that machine learning allows learning from data, finding patterns, trends and behaviors in large amounts of data, which represent historical data. The patterns found within the data are used to create predictive models. Many of these models have been implemented using the CRISP-DM methodology which allows us to guide us in the development of this reserach which allows us to guide us in the execution of projects, describing the approaches and experiences commonly used by experts to address data mining problems [8].

The problem addressed in the present research is the lack of knowledge about the possible academic performance of students entering the university of the school of systems. The researches cited will allow us to know the state of the art of the problem to be addressed, related to the predictor variables and their availability. In some cases, the researches work with few variables and others with many variables, others with few student records and others with many student records, and finally all the works cited predict a single classification, which is generally the result of a course.

In this research, the university entrance exam score is related as a predictor of their academic performance, as evidenced in the research developed by Calva *et al.* [9]. Another determinant predictor is the average obtained during the semester or course, as pointed out by Unsihuay *et al.* [10]. The literature does not show research on predictions of academic performance based on three sequential courses, which currently represents a gap, but it does show predictions of a course or event. For this reason, the task of analyzing 827 students, each one with 9 characteristics such as sex, university entrance score, school of origin, entrance modality, among others, was undertaken. We have the limitation of having few students and few characteristics when using classification algorithms, however we have requested all the data available to the university from 2011 to 2021.

The present research will work on the problem of academic performance of new entrants in the courses of Discrete Structures I, Discrete Structures II and data structure and algorithms, which are initial and sequential courses of the career, where students have shown difficulty in passing them. In this context of the problem, it has been found that approximately 50% of students have failed the first course in their first enrollment. Being immersed in this problem, in this research we want to achieve the objective of predicting the academic performance path of entering university students using classification algorithms, labeling the student as passed or failed according to Peruvian university classification criteria. The academic performance pathway is defined by

Manuscript received June 5, 2023; revised July 17, 2023; accepted September 18, 2023.

The authors are with Universidad Nacional de San Agustin de Arequipa, Peru.

*Correspondence: esaire@unsa.edu.pe (E.A.S.P.)

three sequential courses. The research will be approached from a quantitative approach and at a correlational level.

II. Literature Review

The literature review shows research projects to date related to predicting academic performance. Bravo *et al.* [11] have predicted students' academic performance based on academic, demographic and sociodemographic data, using decision tree, KNN, support vector machines and naive Bayes algorithms. They worked with 4738 students, where the best performing algorithm was the KNN with an accuracy ranging from 78.5% to 80%. The gender variable had no impact on the prediction, however variables such as grade point average and place of residence were determinant. The research focused on predicting the academic performance of a course, which is a gap that has not yet been filled.

Cajahuanca *et al.* [12] have succeeded in predicting college dropout in COVID-19 times using an artificial neural network. The research was done using data from 392 students, where data were collected through surveys grouped by academic, demographic, social and institutional data. The neural network correctly classified 73% of the test data. Among the variables used were the level of education of the father and mother, extra class payments, alcohol consumption, among others. In the end, the most determining variables in the prediction were study time, absences and time spent on social networks. The strength of the research is that it worked with 29 variables and the weakness is that it has few students, where there is still a gap in predicting a set of sequential courses.

Unsihuay *et al.* [10] investigated the main predictor variables that influence the academic performance of students after six semesters of university entrance. They worked with 622 students and applied twelve classification algorithms, where an ensemble was used based on the algorithms that showed the best results, 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 variables sex, age at entry and type of school of origin were not determinant, in contrast to the course grade, which was determinant. The variables were analyzed in parallel for several courses and academic performance was determined; however, there is still the opportunity to work with sequential courses and to show predictive results in each one of them.

Calva *et al.* [9] implemented a model based on supervised machine learning with the purpose of predicting whether a student passes the remedial course. They used gradient boosting and logistic regression algorithms, where the inputs were predictor variables grouped into demographic, socioeconomic, family, institutional and academic performance in the application. The population consisted of 7139 students. The first algorithm obtained an accuracy of 96% in 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. Twenty predictor variables were used, which was determinant for the variety of data with which we worked, however, there remains the opportunity to work with several sequential courses and to see which variables are more determinant in each course.

Gil-Vera *et al.* [13] present a model based on a Neural Network, which has allowed predicting the academic performance of students, using academic, demographic, social and institutional data of 395 students with 39 variables or characteristics. The model classifies with an accuracy of 73%, the strength is based on analyzing 39 variables, which gives the opportunity to explore who are the most determinant variables in academic performance. There remains the opportunity to test with other supervised algorithms.

Franco *et al.* [14] developed models with predictive ability of student academic risk, using educational data mining, for early detection of academic risk. In that research, sociodemographic data and the results of university entrance exams of 415 students of computer science majors enrolled between 2016 and 2019 were applied. The best classification model was based on the LMT algorithm with an accuracy of 75.42%. The reduction of variables reached up to 9 out of 65 attributes, which were the most determinant ones worked with weka software, 5 classification algorithms were used. A strength is the availability of 65 attributes per student, which the institution has been storing over the years.

Castrillón, Sarache *et al.* [15] were able to predict the academic performance of higher education students from educational, family, socioeconomic, habits and customs data, among others, using Bayesian algorithms. With all these data, a model capable of early classification of a new student was achieved. The results allowed the educational institution to identify students with academic performance problems in advance. The population consisted of 460 students with 22 variables, where an accuracy of 91.7% was obtained. The opportunity to have several attributes of the students is determinant in the accuracy of the classification models, on the contrary, if we have few attributes on the part of the institutions

Bedregal-Alpaca *et al.* [16] studied the academic performance of Systems Engineering students between the years 2011 and 2016, they worked with 976 students. They used data such as university entrance score, course grades, credits theoretically allowed and credits carried in practice, and some personal data. Neural networks, decision trees and clustering were used to predict students' academic performance. The average accuracy result of the algorithms was 61% to be able to know in an early way the academic performance of the students, being the decision tree algorithm the one that had the best results. In this research we only worked with the data available to the institution, which were 24 attributes, where many of them were discarded for not having implications such as names, surnames, code, among others.

It has been found in the state of the art regarding the problem being addressed, the opportunity to predict a set of sequential courses, where the output of each course is a new input to predict the next course, which is the objective of the present research.

III. Research Methodology

Through the CRISP-DM data mining methodology and the application of supervised machine learning algorithms, a proposed solution to the problem will be developed, which consist of predicting the academic performance pathway of incoming students, based on three courses where students have difficulty in passing them. Models have been developed based on logistic regression, random forest and XGboost algorithms,

Initially, we have worked with 778 students and 24 attributes, with data collected from 2011 to 2021, which are the only available data that the university possesses and has provided us with for the present research. The scheme showing the general steps to be followed is shown in Fig. 1.

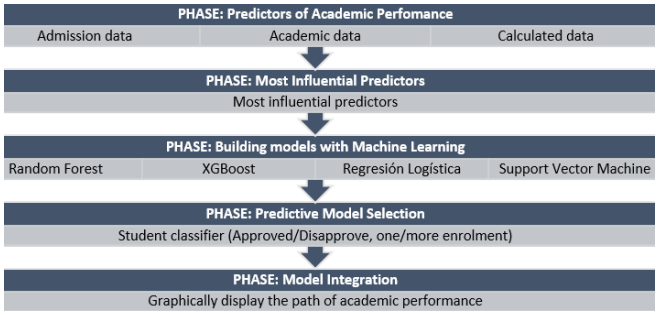

Fig. 1. Outline of the proposal.

The predictive model has as input the admission data and academic data provided by the university. In addition, other data have been calculated, such as the age at which the student left high school, the time elapsed since leaving high school until entering university, the age at which the student entered university, the number of attempts the student had before passing a course, and the number of attempts the student made to pass a course but did not pass it. The outputs are considered by a binary value, which represents whether the student passed or failed the course, and another binary value which represents whether the student passed the course in one enrollment attempt or in more than one enrollment attempt. Input and output data for the models are shown in Table I.

TABLE I: Input and Output Data

I/O	Elements
Input Data	• Admission data • Academic data • Calculated data
Output Data	• Classify the course status : Pass/Fail. • Classify enrollment number : Passed course on one enrollment attempt/Passed course on more than one enrollment attempt

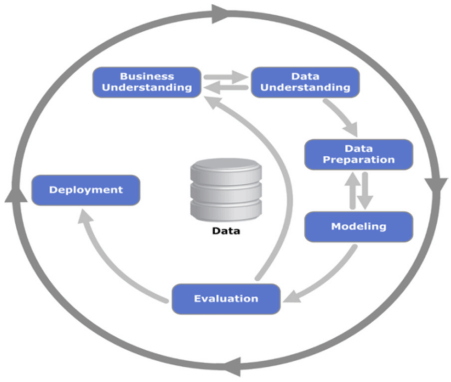

Fig. 2. Phases of the CRISP-DM methodology.

The proposal was developed based on the application of the CRISP-DM methodology, which was worked with the phases and tasks proposed in each of them, except for the deployment or implementation phase. The CRISP-DM methodology is shown in Fig. 2.

A. Business Understanding

In the systems engineering career, incoming students have been characterized by difficulties in passing a set of sequential courses. A first exploration of the data shows that approximately 50% of students have failed the first course in their first enrollment and in other cases, some students have dropped out of the course. Table II shows the three sequential courses where students have shown difficulty in passing, according to data provided by the school systems.

TABLE II: Courses Used in the Prediction

Semester	Courses that are difficult to pass
First year	Discrete Structures I (first semester) Discrete Structures II (second semester)
Second year	Data structure and algorithms (first semester)

B. Understanding the Data

The data requested from the university come from two sources, the first refers to the admission data of the students entering the school of systems and the second source refers to the academic data of the students of the school of systems who have enrolled. Next, in Table III we can see the total number of variables that have been provided by the university.

TABLE III: Admission and Academic Data

	No.	Attribute	Description
Admission Data	1	Last Name and First Name	Student's last name and first name
	2	Sex	Student's gender
	3	Date of Birth	Student's date of birth
	4	Department	Department where born
	5	Province	Province of birth
	6	District	Province of birth
	7	Ubigeo Birth	Ubigeo by birth
	8	College	School of origin
	9	Ubigeo school	Ubigeo of the school
	10	College Department	School department
	11	Province School	Province of the school
	12	College District	School district
	13	Type of school	Type of school
	14	Year of graduation	Year of graduation
	15	Entry mode	University Admission Modality
	16	Score	College Entrance Score
	17	Extraordinary Modality	Extraordinary modality
Academic Data	1	CUI	Student code
	2	Entry code	Student's access code
	3	Surname and first name	Student's last name and first name
	4	Course	Course in which enrolled
	5	Note	Grade obtained in the course
	6	State	Student status
	7	#enrollment	Registration number

C. Data Preparation

Many of the data provided by the university were not taken into account in this prediction process, since they do not contribute any significant value, as is the case of the variables last and first names, CUI, entrance code, among others. However, new data were created, such as the student's age at leaving school, the student's age at university entrance, and the time elapsed since leaving school until entering university, these new data were calculated based on the data on the date

of leaving school and the date of entering university. The variables department, province and district where the student was born, were replaced by the variable place of birth of the student, the same happened with the department , province and district of the school where the student studied, it was replaced by place of the school, for a better management and compression of the data Table IV initially shows the variables that were used in the algorithms, in addition to the creation of new variables.

TABLE IV: DATA PREPARATION

	Data	Description
1	Sex	Student's gender
2	Place of Birth	Student's place of birth
3	Location College	Student's school location
4	School Type	Type of student's school
5	Age of graduation	School leaving age
6	Elapsed Time	Time from high school graduation to university entrance
7	Age Income	Age of college entrance
8	Modality	University Admission Modality
9	Score	Entrance exam score
10	First	First enrollment
11	Approved	Course approved
12	Disapproved	Failed course
13	Abandoned	Abandoned course
14	Attempt	Number of attempts

The data were integrated into a single source, consisting of admission data, academic data and the creation of new variables. Fig. 3 shows data regarding each student's history for the first course called Discrete Structures I.

Elapsed Time	Entrance Age	Modality	Score	First	Disapproved	Abandoned	Approved	Attempt
2	18	Ceprunsa	60	Disapproved	3	3	0	6
2	18	Ceprunsa	59	Approved	0	0	1	1
2	18	Ceprunsa	58	Disapproved	1	0	1	2
4	21	Ceprunsa	55	Abandoned	0	3	0	3
2	18	Ceprunsa	56	Disapproved	2	0	1	3
4	20	Ceprunsa	65	Approved	0	0	1	1
2	18	Ceprunsa	67	Disapproved	2	1	1	4
2	18	Ceprunsa	53	Disapproved	1	0	1	2
1	18	Ordinary	58	Approved	0	0	1	1
2	19	Ordinary	57	Approved	0	0	1	1
3	19	Ordinary	53	Disapproved	2	0	1	3
2	19	Ordinary	48	Approved	0	0	1	1
2	19	Ordinary	50	Approved	0	0	1	1
1	17	Ordinary	52	Approved	0	0	1	1
3	19	Ordinary	47	Disapproved	5	0	1	6
1	18	Ordinary	50	Approved	0	0	1	1
1	17	Ordinary	50	Approved	0	0	1	1
2	18	Ordinary	52	Abandoned	2	1	1	4
1	18	Ordinary	47	Abandoned	0	3	0	3
2	19	Ordinary	50	approved	0	0	1	1
3	19	Ordinary	50	Disapproved	4	0	1	5

Fig. 3. Student history—course 1.

Fig. 3 shows the following data for the case or the first student: the column disapproved equals 3 indicates that the student has failed the course 3 times, the column dropped equals 3 indicates that the student has dropped the course 3 times , the column passed equals 0 indicates that the student has passed the course 0 times and finally the column attempted equals 6 indicates the student has had 6 enrolments or attempts to pass the course, but never succeeded, as the final value of the column passed equals 0. Finally, the variables used in the classification algorithms are those shown in Table IV, except for failed and dropped out. Based on the history of the students in each of the courses, statistics were obtained on whether they passed the course in one enrollment attempt or passed the course in more than one enrollment attempt, which is determinant for classifying the enrollment number. Table V shows the statistics based on numbers of students who passed the course and in which enrollment attempt they passed the course. The statistics shown have allowed the researchers to generate a binary classification, based on the following classes: whether the student passed the course in one enrollment attempt or the student passed the course in more than one enrollment attempt. The number of students or enrollments to generate the machine learning models is a determining factor, however, in this research it is a limitation since we only have an average of 700 students.

TABLE V: APPROVED WITH THEIR REGISTRATION NUMBER APPROVED-NUMBER OF ENROLMENTS

Course	Approved	Registration	Quantity
Discrete Structures I	641	One license plate	331
		More than one license plate	310
Discrete Structures II	591	One license plate	481
		More than one license plate	110
Data Structure and Algorithms	458	One license plate	327
		More than one license plate	131

D. Modelling

The architecture designed to determine the academic performance path based on the predictive classifier model is based on the three sequential courses already mentioned. The data flow of the proposed architecture for each course is described below, as shown in Fig. 4.

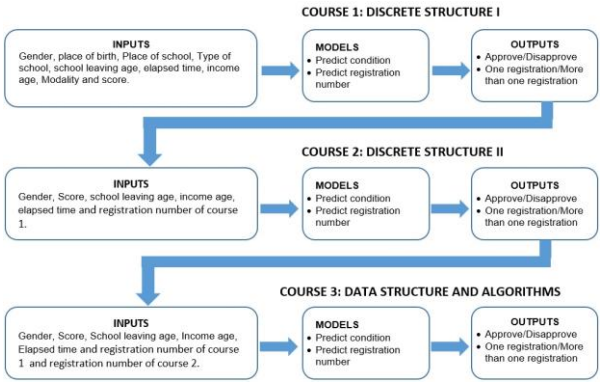

Fig. 4. Proposed academic performance pathway.

For the first course called Discrete Structures I there are two predictive models: The first model allows predicting whether a student passes or fails the course, for which nine input variables were used. The second predictive model will allow predicting whether the student passed the course in one enrollment attempt or in more than one enrollment attempt; for which the same input variables were used.

For the second course called Discrete Structures II, only students who passed the course Discrete Structures I are present, for which we have two predictive models: The first model allows predicting whether a student passes or fails the course, for which we have a total of six input variables: five input variables that were the most determinant in this model plus one input which is the number of enrolment attempts in which the first course was passed five input variables that were the most determinant in this model and the sixth input variable is the number of enrollment attempts in which the student passed the first course. The second predictive model will allow predicting whether the student passed the course in one enrollment attempt or in more than one enrollment attempt; for this model we have the five most determinant input variables. For the third course we also worked with five

more deterministic variables, as shown in Fig. 5.

```
from sklearn.feature_selection import mutual_info_classif
mi=mutual_info_classif(df_encoder,Y,discrete_features=True)
mi=pd.Series(mi)
mi.index=df_encoder.columns
```

```
EdadEgreso                          0.010021
TiempoTranscurrido                  0.019258
EdadIngreso                         0.019403
Puntaje                             0.037197
Sexo_F                              0.001483
Sexo_M                              0.001483
LugarNac_AREQUIPA                   0.001338
LugarNac_OTRO DEPARTAMENTO          0.000010
LugarNac_PROVINCIA AREQUIPA         0.002635
LugarColegio_AREQUIPA               0.001190
LugarColegio_OTRO DEPARTAMENTO      0.000043
LugarColegio_PROVINCIA AREQUIPA     0.001614
TipoColegio_Nacional                0.001655
TipoColegio_Parroquial              0.001503
TipoColegio_Particular              0.000459
```

Fig. 5. Most determinant variables in prediction.

Finally, for the third course called data structure and algorithms, only the students who passed the discrete structures II course are present, for which we have two predictive models: The first model allows predicting whether a student passes or fails the course, for which we have a total of seven input variables which are: five input variables that were more determinant in this model, one input which is a number of enrolment attempts in which they passed the first course and one input which is the number of enrolment attempts in which they passed the second course. Five input variables that were the most determinant in this model, the sixth input is the number of enrollment attempts in which the student passed the first course and the seventh input is the number of enrollment attempts in which the student passed the second course. The second predictive model allows predicting whether the student passed the course in one enrollment attempt or in more than one enrollment attempt, for this model we have the five most determinant inputs.

For the implementation of the models, the Python tool provided by Google Colaboratory has been used, where the following tasks have been developed: convert the categorical variables to dummy variables, determine the independent variables(predictors) and the dependent variable (target), divide the data for training and testing, implement the models by testing algorithms such as logistic regression, random forest, XGboost, among others, search for the best hyperparameters for the algorithms and improve the values of the metrics. Fig. 6 and Fig.7 show the source code of the dummy variables and percentage of data to train the model and test the model.

```
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	

Fig. 6. Variables dummies.

```
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)
```

Fig. 7. Data for training and testing.

The following are the tasks developed by each course to build the predictive models.

Course 1: discrete structures I. The predictive model has been generated with all the variables and it has also been tested to generated the model with the most relevant variables. The results in both cases were very similar in the values of their metrics, for this reason in the end it was decided to work with all the variables. For the selection of the most determinant characteristics or variables, the techniques of mutual information and permutations of the random forest algorithm were used. To make a more complete validity of the final model, 50 students were selected before implementation, consisting of 25 students from each class. In Fig. 8 we can see the selection of the 50 students from each class that were used in the models of the other courses.

```
a=0
b=0
while a<25:
  df_temp=df.sample()
  if not df_temp[df_temp['INTENTO.1'].astype(int)==1.0].empty:
    #print(df_temp)
    validation_set=validation_set.append(df_temp)
    #print(validation_set)
    df.drop(df_temp.index,inplace=True)
    a=a+1
while b<25:
  df_temp=df.sample()
  if not df_temp[df_temp['INTENTO.1'].astype(int)==0.0].empty:
    validation_set=validation_set.append(df_temp)
    df.drop(df_temp.index,inplace=True)
    b=b+1
```

Fig. 8. Selection of 50 students.

Predicting the status of course 1. In Table VI we can see the number of students with respect to the Discrete Structures I course, which represent the entries in records for the classifier model of student status, i.e. whether they pass/fail the course, after having attempted in one or more enrollments. Predicting the enrollment number of course 1. The inputs to implement the model that predicts the enrollment number consist only of all students who have passed the Discrete Structures I course. Of the 641 students who passed the course, 331 students passed the course in one enrollment attempt and the remaining 310 students passed the course in more than one enrollment attempt.

TABLE VI: QUANTY OF STUDENTS IN THE 3 COURSES

Course 1: Discrete Structures I	**Quantity**
Number of approvals	641
Number of disapproved	131
Did not enroll	6
Total students	778
Course 2: Discrete Structures II	**Quantity**
Number of approvals	591
Number of disapproved	33
Did not enroll	17
Total students	641
Course 3: Data Structure and Algorithms	**Quantity**
Number of approvals	458
Number of disapproved	52
Did not enroll	81
Total students	591

Course 2: discrete structures II. In this second course it has been seen that the most determinant variables have shown better results in their metrics. The most determinant variables

found are gender, score, age at graduation, age at entry, time elapsed and number of enrollment in course 1.

Predicting the status of course 2. The students who took the discrete structures II course are those students who passed the discrete structures I course, which are 624. In this model, 50 students were also selected before the implementation, in order to run a more realistic test to the model. In Table VI we can see the number of students with respect to the discrete structures II course. Predicting the number of enrollment in course 2. Of the 591 students who passed the Discrete Structures II course, 481 students passed in the first enrollment and 110 students passed in more than one enrollment.

Course 3: Data Structure and Algorithms. In this third course we have worked with the most determinant variables, as explained in course 2, in addition to the number of enrollment in course 1 and number of enrollment in course 2. Predicting the student status of course 3. The students who took the data structure and algorithms course, are those students who passed the discrete structures II course, which are 591. In Table VI we can see the number of students with respect to the data structure and algorithms course. Predict the enrollment number of course 3. Of the 458 students who passed the Data Structure and Algorithms course, 327 students passed in the first enrollment and 121 students passed in more than one enrollment.

In total 6 classifier models have been generated, two models for each course, where one model predicts the student status and the other model predicts the enrollment number. Each model has generated a data flow representation with the corresponding weights that must accompany the predictor variables, in addition to the value of the hyperparameters that each algorithm handles, all this is shown in the source code developed in Python. The classification models were integrated and the results of the academic performance pathway are shown in Fig. 9.

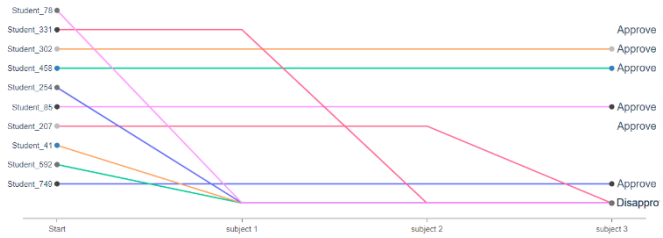

Fig. 9. Academic performance pathway.

IV. Proposal Validation and Result

This section evaluates all the proposed supervised models that have been proposed, taking into account the values of the performance metrics of the models. According to the problem and the objectives set, the most optimal model will be determined. The evaluation in the obtained models was verified the efficiency with the test data, which represents 20% of the total, that is to say those data that were separated and that the obtained model does not know them and was not taken into account in the training of the models, which is called test. Before the training of the models, 50 students have been separated, 25 students from each class (pass/fail), which is called set validation, and represents a more real and complete validation of the models,

Validation of the models generated from course 1 (Discrete Structures I), which consists of predicting the student ś state. Initially the tests were done with the test data, which represents 20%. Then the models were tested with the data called set validation. It is important to point out that the 25 students selected from each class have the characteristic of being balanced data with respect to both classes, i.e., 25 students who passed the course and 25 students who failed the course, compared to the test data which is 20% of students, There is a probability that within this 20% there are more pass than fail students or the other way around, as the selection is done randomly, which is why 25 students were chosen from each class to ensure that there is no bias The tests to the models are shown in Table VII.

TABLE VII: State Metrics—Course 1

Algorithm	Accuracy	Recall	F1
Tests with test data			
XGBoost	0.870	0.980	0.920
Random Forest	0.916	0.918	0.880
Testing with set validation data			
Logistic Regression	0.600	0.680	0.620
Random Forest	0.820	0.920	0.830
Model stacking	0.860	0.880	0.860

The best results were obtained with the random forest algorithm applied to the test data, and the tests with the set validation data obtained better results when the Logistic regression and random forest algorithms where stacked. To obtain better results, we applied a machine learning method called stacking, which is a meta-learning technique, where the combining algorithm is linear regression and the stacked algorithms are those shown in each test. Fig. 10 shows the source code of one of the tests performed with 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
```

Fig. 10. Random forest metrics.

Validation of the models of course 1 (discrete structures I), to predict the number of enrollment. The best results were found with the random forest algorithm, as shown in Table VIII. In this model we did not choose to develop the extra test with set validation due to the low results obtained. The test of the algorithms were only done with the test data and the results are also shown with a stacking of models of the random forest and XGboost algorithms.

TABLE VIII: Intent Metrics—Course 1

Algorithm	Accuracy	Recall	F1
Random Forest	0.56	0.55	0.58
XGBoost	0.53	0.98	0.69
Stacked models	0.53	0.98	0.69

Validation of the models of course 2 (discrete structures II), to predict student status. The best results were found with XGboost algorithm, and a stacked model between XGboost and logistic regression to test the test data and the

set_validation data. The results are shown in Table IX.

TABLE IX: COURSE 2 METRICS

Algorithm	Accuracy	Recall	F1
State metrics			
Tests with test data			
Random Forest	0.92	0.99	0.96
XGBoots	0.93	1.00	0.96
Stacks of models	0.93	1.00	0.96
Testing with set validation data			
Logistic regression	0.62	0.72	0.65
XGBoost	0.64	1.00	0.72
Model stacking	0.60	0.96	0.73
Intent metrics			
Tests with test data			
Random Forest	0.64	0.58	0.55
Testing with set validation data			
Random Forest	0.52	0.16	0.25
XGBoost	0.48	0.88	0.62

Fig. 11 shows the source code of one of the tests performed with the xgboots 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
```

Fig. 11. XGboost metrics.

Validation of the models of course 2 (discrete structures II), to predict the number of enrollment. The best results obtained to predict the number of enrollment were based on the random forest algorithm, as shown in Table IX.

Validation of the course 3 models (Data Structure and Algorithms), to predict the student's status: The best results obtained were found with the random forest model and a stacked model which is based on the logistic regression and XGboost algorithms, as shown in Table X.

TABLE X: COURSE 3 METRICS

Algorithm	Accuracy	Recall	F1
State metrics			
Tests with test data			
Random Forest	0.89	0.97	0.93
XGBoost	0.52	1.00	0.67
Testing with set validation data			
Logistic regression	0.60	0.60	0.60
XGBoost	0.66	0.88	0.72
Stacking	0.7	0.92	0.75
Intent metrics			
Tests with test data			
Random Forest	0.82	0.82	0.87
Logistic Regression	0.77	0.25	0.29
XGBoost	0.85	0.88	0.90
Testing with set validation data			
XGboost	0.54	0.28	0.37
Random Forest	0.61	0.42	0.52
Logistic Regression	0.62	0.60	0.61

Validation of the models of course 3 (data structure and algorithms), to predict the number of enrollment. The results are shown in Table X, where the random forest algorithm showed the best results.

The results obtained for predicting the state were very acceptable, however, the results were not very good for predicting the number of enrollment, especially when testing the models with the data representing the set_validation. In Fig. 12 we can see the results of the logistic regression algorithm.

```
LogR_=drive.CreateFile({'id':"1qYtqWuW6_Kuti4urKxZzrH4Nx__Gjtz2"
LogR_.GetContentFile('LR.joblib')
loaded_LogR_ = joblib.load("LR.joblib")
print("precision:",accuracy_score(Y_val,loaded_LogR_few))
print("recall:",recall_score(Y_val,loaded_LogR_few))
print("f1:",f1_score(Y_val,loaded_LogR_few))

precision: 0.6
recall: 0.68
f1: 0.6296296296296295
```

Fig. 12. Logistic regression metrics.

The integrative model presented is based on the most accurate classifier models that have allowed predicting the student's status (pass/fail) based on the number of enrollment attempts. Table XI shows the algorithms that have shown the best results for each course.

TABLE XI: INTEGRAL MODEL CLASSIFIERS

Course	Algorithm	Accuracy
Discrete Structures I	XGBoost	0.870
Discrete Structures II	Stacking	0.930
Data Structure and Algorithms	random forest	0.890

V. DISCUSSION

Based on the research carried out 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 seen in the literature consulted.

The reduction of dimensionality through the technique of mutual information and permutation of the random forest algorithm has improved the results, showing that the most determinant variables in this context are gender, college entrance score, age of graduation from high school, time elapsed since graduation from high school until college entrance and age of college entrance among the admission data. The research of [17] addressed the aspect of dimensionality and determined that the most influential variable was the age at which they started their studies, which is a result that is common to this research. However, there are other studies 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 it had other variables to consider such as credits approved in relation to theoretical credits that should have been approved; these changes are due to the fact that I can count on other additional data, compared to the present study, where the score was the most determining variable in the prediction.

Another research with which we can compare dimensionality reduction results was that of [18], which also worked with historical data from a public institution, and determined that the number of failed courses and the father's level of education 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 might do so. There are also those research that create their instruments to collect data at the time of the research, which are data directly related to the purpose pursued, which increases the richness of the results.

In the literature, it has been found that predictions are made by classifying students into pass and fail, dropout and non-dropout, low performance and 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 [17, 19–21] 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 consulted regarding this line of research and educational contexts, show binary classification predictions of an object or event, however, the contribution of the present research work lies in predicting a student's academic performance path based on three sequential courses where students have shown difficulty in passing, where the output of the first model is an input for the next model and so on. First, models have been developed to predict whether the student passes or fails a course and then models have been developed to predict in which enrollment attempt the student has managed to pass the course, which the student can pass the course in a first enrollment attempt or in more than one enrollment attempt. Finally, in this research it has been possible to integrate the models that predict the student's status (pass/fail) and be able to see an early report of how the student's academic performance will be in the three sequential courses that have been mentioned. The classifier algorithms that showed the best results in this research were random forest, XGBoost and logistic regression. The accuracy of the models ranges from 87% to 93%, depending on the algorithm used and the model implemented, where there is an opportunity to improve the accuracy by having more records and more attributes of the students, which leads to retesting the algorithms seen and other classification algorithms.

VI. Conclusions

The results of this research have allowed the construction of a classifier model, which allows to know the academic performance path of students entering three sequential courses which have shown difficulty to be approved. During the process, it has been determined which are the most influential variables depending on the algorithm and the data, which has allowed obtaining a classifier model with greater precision. The selection of the algorithms to implement the models has been through trial and error, seeing the results of the prediction with the test data and the values of the metrics. The tests have allowed to identify the most efficient models, which have been random forest and XGBoost in most cases. The integration of the classification models to graphically show the student's academic performance with respect to the three sequential courses shows a path of academic performance for each student. The quality of the data related to the problem being addressed is of vital importance to obtain a more accurate classifier model, since it has been seen in the literature, research that has used data directly related to the problem and the results have been more conclusive. Another determining aspect in this type of research is that due to the amount of data needed, historical data is used, where there is a possibility that the quality of the data and even more so the small amount of variables may determine an unfavorable result. In this research it has been seen the limitation of records and variables that were available at the university. However, despite this limitation, positive results have been achieved based on the different tests with classification algorithms. Finally, it is concluded that trial and error in using Machine Learning algorithms with the available data, has allowed experimenting and finding the best classifier model, according to the data that have been provided by the university. The present research invites to extend the work by experimenting with more records and variables related to academic performance, which was a limitation in the present research. It is suggested that the new predictors to be used be closely related to the target variable of the prediction.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The author Edwar Abril Saire-Peralta developed the technical part of the research, collected the data, analyzed the data, implemented the models based on the algorithms in Python, did the tests, wrote and revised the article.

The author María del Carmen Córdova-Martínez developed the methodological part of the research, wrote and reviewed the semantics of the article, and translated the article into English.

In the end all the authors have approved the final version.

References

[1] A. Pérez-Luño, J. Ramón Jerónimo, and J. Sánchez Vázquez, "Exploratory analysis of the variables that condition academic performance," Seville, Spain: Pablo de Olavide Univerity, 2000

[2] M. V. Van and N. Roa, "Factors associated with academic performance in medical students," *PSIC. Educación Médica*, 2005.

[3] S. Rodriguez, F. Eva, and T. Mercedes, "Academic performance in the secondary-university transition," *Journal of Educatión*, 2003.

[4] G. M. G. Vargas, "Factors associated with academic performance in university students, a reflection on the quality of public higher education" *Revista Educación*, vol. 31, no. 1, p. 43, 2007. https://doi.org/10.15517/revedu.v31i1.1252

[5] S. Rodríguez Espinar, "Research models of academic achievement: Problems and trends," *Educational Research Magazine*, vol. 3, no. 6, pp. 284–303, 1985.

[6] J. Bell, *Machine Learning: Hands-on for Developers and Technical Professionals*, 2020.

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

[8] P. Chapman, C. Julian, K. Randy, K. Thomas, R. Thomas, and S. Wirth, *CRISP-DM 1.0: Step-by-Step Data Mining Guide*, 1999.

[9] K. Calva, M. Flores and H. Porras, "A prediction model of academic performance for the leveling course ata the National Polytechnic School based on a supervised learning model," *Latin American Journal of Computing*, vol. VIII, no. 1, 2021. https://doi.org/10.5281/zenodo.5770905

[10] J. E. G. Unsihuay and J. W. S. Flores, "Predicting the academic status of undergraduate students using machine learning profiles algorithms," vol. 1, no. 27, pp. 4–10, 2022. https://doi.org/10.47187/perf.v1i27.142
[11] L. C. Bravo, N. Nieves-Pimiento and K. Gonzalez-Guerrero, "Preducting university academic performance using machine learning mechanisms and supervised methods," *Engineering*, vol. 1, pp. 1–25, 2023. https://doi.org/https://doi.org/10.14483/23448393.19514
[12] J. V. Cajahuanca, A. N. Raymundo, A. L. Franco, and J. J. Flores, "Evaluación of different Machine Learning algorithms for its prediction university dropout," *Social Science Magazine*, vol. 28, no. 3, pp. 362–375, 2022. https://doi.org/10.31876/rcs.v28i3.38480
[13] V. D. Gil-Vera and C. Quintero-López, "Predicting student academic performance with artificial neural networks," *Información Technology*, vol. 32, no. 6, pp. 221–228, 2021. https://doi.org/10.4067/s0718-07642021000600221
[14] E. A. Franco, R. E. L. Martínez, and V. H. M. Domínguez, "Predictive models of academic risk in computer science careers with educational data mining," *Journal of Distance Education (RED)*, vol. 21, no. 66, pp. 1–36, 2021. https://doi.org/10.6018/red.463561
[15] O. D. Castrillón, W. Sarache, and S. Ruiz-Herrera, "Predicting academic performance using artificial intelligence techeniques. Predicción," *University Education*, vol. 13, no. 1, pp. 93–102, 2020.
[16] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén, and V. Cornejo-Aparicio, "Analysis of the academic performance of systems engineering students, desertion possibilities and proposals for retention," *Ingeniare*, vol. 28, no. 4, pp. 668–683, 2020. https://doi.org/10.4067/S0718-33052020000400668
[17] A. R. Quijo, "Development of a model to predict the academic performance of EPN studens based on their level of access to ICT and socioeconomic factors," M. S. thesis, Escuela Politecnica Nacional, Quito, 2023.
[18] L. Quiñones and Y. L. Carrasco, "Academic performance usign data mining," *Espacios*, vol. 41, no. 44, pp. 277–285, 2020. https://doi.org/10.48082/espacios-a20v41n44p17
[19] P. M. Zamora, "Mathematical model to predict the degree of student dropout al the Instituto Superior Tecnológico Bolivar," M.S. thesis Technical University of Ambato, Ecuador. In Repositorio Institucional de la Universidad Técnica de Ambato, 2023.
[20] A. J. C. Garcia, "Model for the prediction of undergraduate student dropout, based on data mining techniques," M. S. thesis, Universidad de la Costa - CUC, Colombia, 2020.
[21] H. E. C. Gismondi, "Predictive model based on machine learning as support for academic monitoring of univerity students," M. S. thesis, Universidad Nacional del Santa, Perú, 2021.