A STACKING BASED HYBRID TECHNIQUE TO PREDICT STUDENT DROPOUT AT UNIVERSITIES

[1]ALFREDO DAZA,

[1] School of Systems Engineering, Universidad Cesar Vallejo, PERU.

E-mail: [1]adazave@ucvvirtual.edu.pe

Abstract

University dropout is a very complex problem that affects the Government, Institutions and students and families in the world. The prediction allows to identify the students who are going to desert early, so that the directors of the Universities can establish strategies to mitigate it. Machine Learning methods are the most recent and effective for this problem. However, so far these methods have been applied independently and not in combination. This paper proposes a hybrid model based on decision trees and neural networks, designed following the KDD methodology, to predict with high precision the university student dropout. The proposal was implemented in Rapid Miner Studio 6.4 and applied to a dataset with 1761 student records and 53 variables for training. Through a variable selection procedure that includes 8 algorithms, 27 variables were selected. The results on 100 new records show an accuracy of 87%, 91%, 98% for decision tree models, neural networks and stacking respectively. In addition, the result of sensitivity is 90.6%, 93.3%, 98.7% for decision trees, neural networks and stacking respectively. Regarding specificity, 76%, 84% and 96% have been obtained for decision trees, neural networks and stacking respectively. The results of accuracy, sensitivity and specificity also show that the hybrid model presents better results than the separate models..

Keywords: *Decision trees, University Dropout, Hybrid Model, Prediction, Neural Networks, Stacking.*

1. INTRODUCTION

University dropout is one of the problems that threaten the achievement of a quality education made that is recognized in today's society and has unintended consequences due to its high social, economic and personal costs, this can be understood, as a situation faced by a student when he aspires and fails to finish his professional career. Likewise, an individual is considered as a deserter, that is, that university student who does not exhibit academic activity, which occurs in the early, middle and late stages [1], [2].

According to the National System of Evaluation, Accreditation and Certification of Educational Quality SINEACE, in 2020, more than 174,000 students in Peru have interrupted their university studies. This figure represents 18.27% of the total number of students, which presumes more than 955,000 young people. Now, if you contrast that number with the discontinuation rate it obtained in 2019 (12%), the percentage for 2020 is about 6 percentage points higher [3]. In addition, in the United States, the overall dropout rate is 40% where 30% comes from first-year students [4] and the dropout rate in India is 15.9% [5].

according to [6], the average annual dropout rate in Brazilian higher education during 2020 is 42%, but with a growth trend. The dropout rate in Denmark during the first year was 20%, much higher than in the United States and England, where it is around 3% and 4% [7].

There are many aspects that contribute to the desertion of students in universities this is due to personal, psychological, institutional, economic, academic and social variables.

In the study carried out by [8] points out the reasons and causes that lead to student desertion, it has been determined that both the economic support and educational quality provided by an education center and how they affect the desertion of the aforementioned group of students. This survey shows that the main cause of dropout is due to the dissatisfaction of students with the service they receive from the institution, adding other additional components such as psychological, sociological and economic factors.

The prediction [9], through the use of machine learning techniques that seeks to find which is the most optimal in efficiency in the prediction of abandonment in some short online subjects of the university, being the objective to serve as a basis in future activities and reduce the probabilities of desertion. Therefore, the prediction based on machine learning allows to identify the students who are going to abandon their studies in the Private Universities, so that then the Institutions carry out the strategies in a timely manner to be able to reduce them.

Several research-focused studies have been carried out, based on Data Mining and Machine Learning with different techniques such as Neural Networks [10], Decision Trees [11], SVM [12], Induction Rules [13], which have been applied independently. There are no studies of hybrid models in topics related to Higher Education in university dropouts. However, studies in prediction problems carried out in other areas show in a general way that hybrid models present better results, such as, for example, the study of [14] applied to mining stocks of the New York Stock Exchange, performs a hybrid model based on SVM and neural networks that results in better accuracy, than models applied independently. In another study applied to networks in attack control, the author [15] proposes a hybrid model based on the combinations of SVM and SOM where it obtains better results in accuracy than models applied independently. A study focused on the evaluation of loans to customers in banks, proposed by [16], where it proposes 5 hybrid models having as the base algorithm for their models the MLP and this model combined them with the classification techniques Naïve Bayes + MLP, Logistic Regression + MLP, MLP + MPL, RBF + MLPand C4.5 + MLP, where in each of its hybrid models it obtains better results than when it was applied independently.

In this context, and based on taking actions to reduce dropout, the research proposal is to provide a solution to the problem of deserting many students in educational institutions using Machine Learning and Data Mining, developing a prediction model based on stacking techniques, which helps predict the reason for desertion of students and show us correct data so that the authorities in the Universities make a timely decision and in this way reduce the dropout rate.

The article consists of 5 sections: background and literary review, materials and methods, results and discussions and finalizing the conclusions.

2. BACKGROUND AND REVIEW OF LITERATURE

Research works have been carried out that allow to identify the students who are going to desert in the Universities worldwide, taking into account economic, demographic and academic variables. Table 1 shows some studies that have used classification techniques in the prediction of abandonment in higher institutions.

*Table 1 : **Classification Work to Predict Dropout in Universities.***

Description	Referenc
Work about the prediction of the academic situation in undergraduate students using machine learning algorithms..	(Gamboa and Salinas, 2022) [17
Study about machine learning in university educational environments.	(Henríquez, Salcedo and Sánchez, 2022) [18]
Research about patterns that identify college dropouts applying data mining	(Urbina,Té Cruz, 202
Study about the Experience of predicting with machine learning techniques in a higher institution and the prediction of academic dropout.	(Fernandez, Preciado,Melchor, Rodriguez ,Conejero andSánchez,2021)[20]
Study that talks about the prediction of academic desertion using algorithms such as: SVM and decision tree.	(Ma and Zhou [21]
Work about the prognosis of academic dropout, using machine learning approach..	(Kemper, and Wigger, 2021) [22]
Research referring to the early prognosis of distance dropout in higher education using active learning	(Kostopoulos, Kotsiantis, Ragos, and Grapsa ,2017) [23]
Work about the forecast of students likely to drop out through machine learning techniques.	(González and Peñaloz 2021) [24]
Work about predicting early student dropout using machine learning in online subjects.	(Urteaga, Siri and Garófalo, 2020) [25]
Study about a prediction model to detect students with the possibility of dropping out.	(Rivera, 2021) [26]
Research that identifies the reasons for students dropping out of university courses through the bayes model and fuzzy logic.	(Vázquez, Quintero and Alanís, 2021) [27]

Works mentioned in Table 1 have been reviewed, where the work carried out by [17] in the UNALM with 622 students, being 23.8% the dropout rate, where it was based on the CRISP methodology, where the Boruta algorithm was used to select predictor variables and twelve classification algorithms were applied, after partitioning data groups to train and evaluate . Then, those models with better values of sensitivity, specificity and balanced accuracy were chosen, obtaining as results that the SVM model with linear kernel obtained better specificity with 0.79 and the Random Forest algorithm obtained sensitivity of 0.947.

The authors [18] conducted a study on machine learning in university educational settings: Academic dropout case, where the dropout rate was 46%, after the first semester of their studies. The main objective of the work was to build a prototype based on machine learning to characterize the student permanence in an academic program, proceeding to perform the following phases: pre-processing, selection of algorithms, training, evaluation, execution and deployment of the prediction, where the results obtained show that the Naives bayes algorithm gave as a significant useful result an accuracy of 99%, concluding that machine learning can help us a lot to develop good models to predict data and thus be able to reach solutions to improve the problem of desertion.

The researchers [19], conducted a study in an educational center in Mexico, where it is indicated that the dropout of students is mostly in the 2 years of having entered the students. For the realization of the study was carried out with 10 635 students during the year 2014 to 2019 of 53 bachelor's degree programs, using the weka data mining software. In this study were considered as attributes: age, program, percentage of subjects approved and in relation to accuracy the decision tree algorithm gave better results with 92.12% than the Bayesian and SVM network algorithms.

The researchers [20] conducted a study, which aims to locate each variable relevant to the prediction of desertion, in an engineering school belonging to a Spanish public university with a total of 215,755 data, where 5,426 belong to access records, 194,569 to the degrees of 7 engineering careers and 3 master's degrees in a public institution Spanish University and 15,760 to scholarships, using machine learning algorithms to build the predictive systems, then the prediction models are evaluated according to certain metrics that are considered relevant to measure the proper functioning of the models according to the problem addressed: identify as many students at risk of abandonment as possible, being the best algorithm to predict the SVM with 89.04%.

The authors [21] used dropout and graduation variables were performed with the proposed models: SFS with probit model, SFS as logit model, SFS with decision tree and decision tree assembly. The authors conclude that, in the case of student dropout, the model that provides the best results are decision trees with 86.0% accuracy.

The researchers [22], made a study at the Karlsruhe Institute of Technology (KIT), to carry out the study we worked with 487 students graduated in 2008, being the attributes studied performance problems, financial difficulties, lack of motivation to study, study conditions, failure in the exam, vocational reorientation, family problems and illness, having as results that the decision tree obtained greater accuracy with 94.30%, unlike logistic regression.

The authors [23] conducted a research work, using 344 data from students of the Hellenic Open University, considering pre-university attributes such as: sex, age, number of children, time at work, computer knowledge, computer use at work; and university students: absence or presence in the session of class 1 and 2 (OCSi), performance in the registered tasks 1 and 2 (TESTi), term of study.

For the use of the attributes they considered 5 steps, in the first they only considered pre-university attributes, in the second the OCS1 attribute is added, in the third the TEST1 attribute is included, then in the fourth the OCS2 attribute is added and in the fifth all the attributes are added.
In the experiment for the partitioning of the data they used cross-validation, they also used several models for the experiments such as Minimal Sequential Optimization (SMO), Random Forest (RF), logistic regression (LR), j48, Naive bayes (NB), bayes Networks, neural networks, where five phases were applied and the one that obtained the best results applying all the attributes was Random Forest, with an accuracy

of 85.17% with respect to active learning performance and the accuracy of neural networks is 78.71%. The authors further conclude that this research is a promising start in the implementation of active learning methodologies to detect high-risk students in distance learning courses.

The authors [24] have conducted a study with 904 students of the physics course of the University of virtual education, in the academic periods 2018-1 to 2020-2, based on the KDD methodology, and the attributes they considered are: city, program, gender, social stratum and desertion. The experiment is carried out with Seewtviz, that is, the Python visualization library making use of the Decision Tree, Random Forest and Logistic Regression algorithms, obtaining the Random Forest algorithm a correct classification of accuracy of 59% and possible deserters of 79% of the deserters correctly (true negatives); as the non-deserters (true positives) 26%. Non-deserters, who were classified as deserters (false negatives) corresponds to 74% and deserters, classified as non-deserters (false positives) represent 21%.

The authors [25], conducted a study, taking into account 4 online subjects taught at the SCEU UTN.BA, for which he used data from 654 students during 2018 and 2019, For the study the variables were used: academic, demographic and economic in addition for the experiment the CORELearn and NeuralNet package was used, for training, then the algorithms of classification: KNN, neural networks, decision trees and Random Forest, where results of 31% accuracy were obtained for the KNN algorithm.

The author[26] has conducted a study at the National Intercultural University of the Amazon, using a database of the institution, for the development of the experiment they used the Python program and Google Colab., making use of the classification methods Logistic Regression, Decision Trees, KNN, and Neural Networks, where an accuracy of 89.90% was obtained with the KNN algorithm when classifying the instances correctly, concluding that the application of a prediction model is very advantageous, since it helps higher institutes to use more effective strategies in reducing academic dropout figures.

The researchers [27] conducted a study of the desertion of students using Bayesian models and fuzzy logic, the experiment was carried out with 232 students, using software such as: Weka, SPSS and Excel After the realization of comparisons in machine learning techniques and determining that fuzzy logic behaves positively in situations of classifying those that are difficult with an approximate accuracy of 77%. Concluding that the use of Bayesian models and fuzzy logic are effective tools in reducing the numbers of student dropout in institutions in Mexico, providing optimal data and managing aid to students, either economically or pedagogically.

Considering the disadvantages and comparing them with others, the learning model is the slowest and most rigorous procedure. the precision and algorithms used by each of them, shown in Table II.

Table 2: Dataset, Precision and algorithms used in the revised works.

Referer	Data se	Precision	Algorithms
[17]	622	69.5%	SVM
[18]	336	99%,	Naives bayes
[19]	10, 635	92.12%	Decision trees
[20]	215,75:	89.04%.	SVM
[21]	90	92%	Decision trees
[22]	487	94,30%,	Decision trees
[23]	344	78.71%	Neural networks
[24]	904	59%	Random Forest
[25]	654	31%	KNN
[26]	5803	89.90%	KNN
[27]	232	77%	Fuzzy logic

Table 2 it can be seen that the highest precisions, which has been obtained is 99%, 92% of the algorithms Naives bayes and Arboles de decisión respectively.

With respect to decision trees, it is observed that when the data set is smaller, the accuracy obtained is higher, in the same way in decision trees using the J48 algorithm. Therefore, we could indicate, the smaller the data set, the greater the accuracy.

3. MATERIALS Y METHODS.

3.1 Data Collection

The data grouping belongs to what was obtained from 1861 undergraduate students enrolled in the professional school of systems engineering of a Private University.

The data were collected from the computer science area, in which they gave us an Excel file and the Academic Records area provided us with the data (certificate of studies) in physical that were filled in an Excel file.

The data that have been used in the present study correspond to academic data of secondary and higher education, economic, pre-university data and demographics of students.

3.2 Methodology

For develop the research, the steps shown in the Figure have been followed. 1.

First, the data of the students was integrated, which were obtained from the SQL Server 2000 database of the computer science area and the Academic Records area into an Excel file.

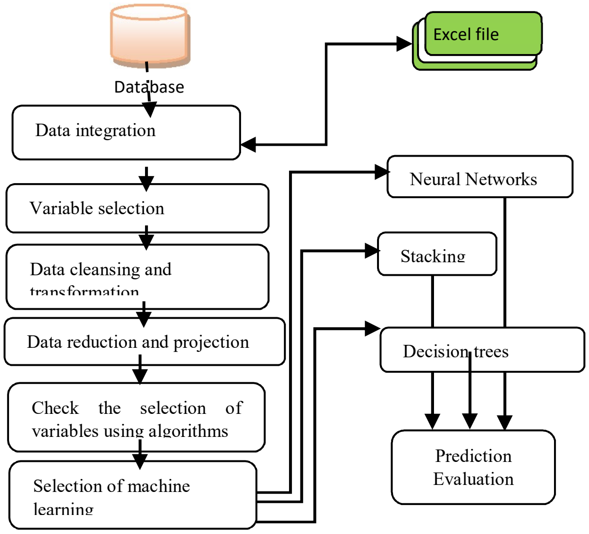
Database
Excel file
Data integration
Variable selection
Data cleansing and transformation
Data reduction and projection
Check the selection of variables using algorithms
Selection of machine learning
Neural Networks
Stacking
Decision trees
Prediction Evaluation

Fig. 1. Process of prediction of desertion in the University through neural networks, decision trees and stacking.

The second step, all the variables that are related to the prediction of the students' dropout were selected through a procedure:

a) Each input variable used in the models of other studies **(A)** was identified.

b) The analysis of the database of a Peruvian Private University was carried out, and then the extraction of the input variables **(B)** was carried out.

c) The proposal of new variables that affect the Higher student dropout rate was made, which have not been used by the models studied. **(C)**

d) The selected input variables are given by the following equation:

$$\mathbf{V.E} = (\mathbf{A} \cap \mathbf{B})\ \mathbf{U}\ (\mathbf{B} \cap \mathbf{C}) = (\mathbf{A}\ \mathbf{U}\ \mathbf{C}) \cap \mathbf{B}.$$

In the third step, the data was cleaned by eliminating those records that have empty values, as well as the ages that were poorly entered into the system.

In the fourth step, the reduction of the variables that were selected in step two was carried out, leaving only the most important for the development of the research, as well as the transformation of some variables such as Sex and the modality of income that had nominal value and was changed to numerical value, so that it can be processed by neural networks because its structure only supports numerical values.

In the fifth step the variables selected in step four were re-verified using 8 variable selection algorithms, with respect to the modeling process three learning machines were selected among which are the decision trees with the J48 algorithm, the neural network with the multilayer perceptrom algorithm and finally the stacking where the decision trees and networks are used neuronal with the aim of discovering the knowledge about the behavior of the variables regarding the desertion of the students in the universities.

To conclude with the phases, an evaluation of the models of decision trees, neural networks and stacking was carried out through precision prediction metrics such as specificity, sensitivity, precision and accuracy.

4. RESULTS AND DISCUSSION.

4.1 Data processing.

In the present study, 1761 data have been considered for training and 100 data for the testing of models.

After performing the cleaning of the data, the transformations of the values of the variables were converted to numerical and the reduction of the variables has been carried out, the variables shown in Table 3 have been considered for the study.

Table 3: Study variables

ID	Variables	Data Type
1	Sex.	Numerical.
2	Age.	Numerical.
3	Prom_col.	Numerical.
4	Curriculum.	Numerical.
5	Cycle.	Numerical.
6	Number of credits approved.	Numerical.
7	Number of credits disapproved.	Numerical.
8	Number of courses approved.	Numerical.
9	Number of disapproved courses.	Numerical.
10	Number of tutorials carried out.	Numerical.
11	Vezcomunica.	Numerical.
12	PromCom.	Numerical.
13	NNotalog.	Numerical.
14	Nnotamatuno.	Numerical.
15	NNotamatdos.	Numerical.
16	NNotamattres.	Numerical.
17	NNotamatTot	Numerical.
18	PromMat	Numerical.
19	TotalCourses	Numerical.
20	Nppa	Numerical.
21	Modality	Numerical.
22	LevelIng	Numerical.
23	AcIngle	Numerical.
24	Com Level	Numerical.
25	AcCompu	Numerical.
26	Category	Numerical.
27	ActESTUDIA	Numerical.

For verify the importance of the variables in the student's desertion at the University, 8 variable selection algorithms have been used [28], as shown in Table 4, where each of them has been applied to the 27 study variables. Where the results obtained, were grouped according to the frequency of each of the variables as shown in Table 5.

Table 4: Variable selection algorithm

N°	Algorithm
1	CfsSubsetEval.
2	CorrelationAttributeEval.
3	GainRatioAttributeEval.
4	InfoGainAttributeEval.
5	OneRAttributeEval.
6	PrincipalComponents.
7	ReliefAttributeEval.
8	SymmetricalUncerAttributeEval.

Table 5 you can see that of the 8 algorithms of selection of variables 7 algorithms have selected the variable PromMat, 7 algorithms have selected the variable NNotamatuno, 8 algorithms have selected the variable amount of tutorials carried, 7 algorithms have selected the variable number of approved courses, 7 algorithms have selected the variable cycle, 8 algorithms have considered the variables amount of credits approved, 7 algorithms have considered the total variable of courses, 7 algorithms have considered the variable NNotamatTot.

4.2 Modeling.

This study used the Rapid Miner Studio tool which is a globally known data science tool and is positioned as a leader in the Gartner Magic Quadrant [29] for data science and data mining platforms in September 2019.

Table 5: Frequency of Variables by Algorithm

Number	Variables	Frequency
1	PromMat	7
2	Nnotamatuno	7
3	Number of tutorials carried out	8
4	Number of courses approved	7
5	cycle	7
6	Number of credits approved	8
7	TotalCourses	7
8	NNotamatTot	7
9	AcIngle	7
10	LevelIng	7
11	PromCom	8
12	NNotamatdos	6
13	NNotalog	7
14	Nppa	7
15	NNotamattres	6
16	vezcommunica	7
17	AcCompu	6
18	Com Level	6
19	Age	7
20	MODALIDAD_NUM	6
21	sexo_num	7
22	Number of courses not approved	7
23	Number of disapproved credits	7
24	Categoria_NUM	7
25	curricula_num	8
26	PROM_COL	7

When is completed all the stages of data processing have been completed, we proceed to the modeling stage, where the modeling of the decision trees has been carried out with the J48 algorithm, neural networks with the multilayer perceptron algorithm and stacking [30] which is a combination of classifiers where in this study the first base will be composed of decision tree classifiers and the second base by the neural networks, in relation to our case study, the randomforest, J48, ADtree algorithms were considered for the decision trees and for the neural networks the multilayer perceptrom algorithms as shown in Figure 8.

The figures at the bottom show the modeling that has been done in each of the cases with the Rapid Miner tool.

Figure 2 shows the modeling with the decision trees where it can be seen that two data operators have been used, one for training with 1761 data and the other with 100 data for the test of the model, to then link it with the operator select the variables and then join it with the validation operator that uses the cross-validation method [31] where it will divide the dataset in k subset of equal size, in this case k=10.

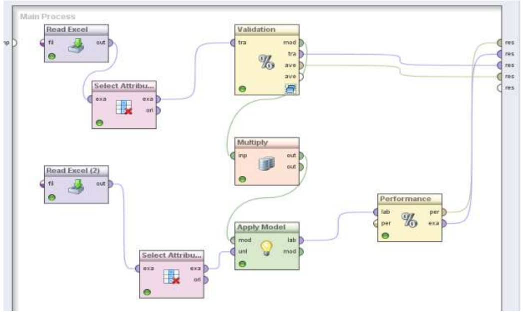

Figure 2: General visualization of the data processing task using decision trees with the J48 algorithm.

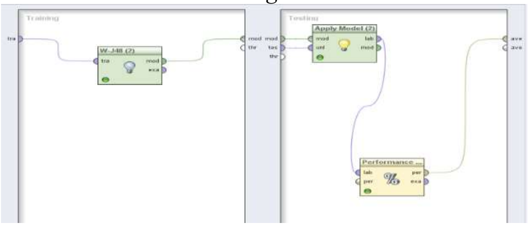

Figure. 3: General visualization of the performance measure using decision trees with the J48 algorithm.

The validation operator shown in Figure 2 is a nested operator containing the J48 operator, the apply operator, and the performance operator shown in Figure 3.

Figure 4 shows the modeling with the neural network, where it can be seen that two data operators have been used, one for training with 1761 data and the other with 100 data for the test of the model, and then link it with the operator select the variables and then join it with the validation operator that uses the cross-validation method where it will divide the data set into k subset of equal size, in this case k=10.

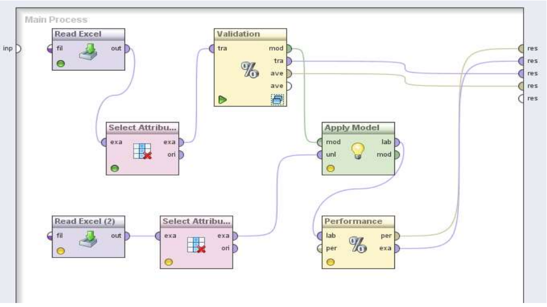

Figure. 4: General visualization of the neural network data processing task with multilayer perceptron.

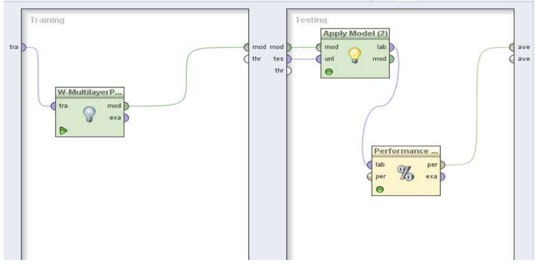

Figure. 5: General visualization of the performance measure using neural networks with multilayer perceptrom.

The validation operator shown in Figure 4 is a nested operator containing the multilayer perceptron operator, the apply operator, and the performance operator shown in Figure 5.

The modeling with the stacking technique is shown in Figure 6, where it can be seen that two data operators have been used, one for training with 1761 data and the other with 100 data for the test of the model, to then link it with the operator select the variables and then join it with the validation operator that uses the cross-validation method where it will divide the data set into k subset of equal size, in this case k=10.

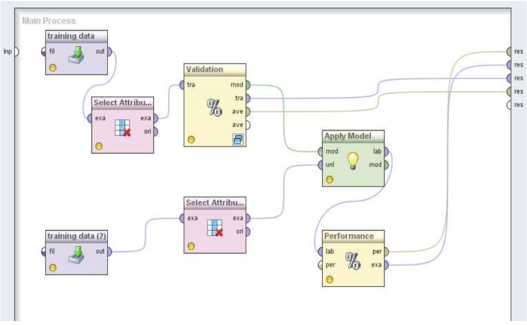

Figure. 6: General visualization of the data processing task with the Stacking technique.

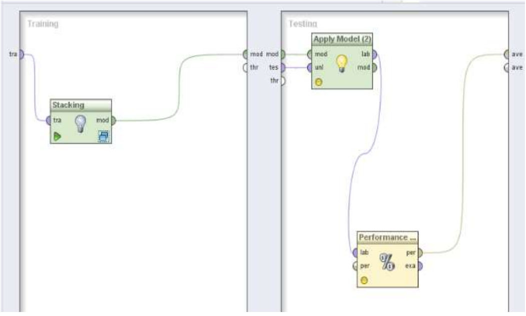

Figure. 7: Overview of the performance measure using stacking.

The validation operator shown in Figure 6 is a nested operator containing the stacking operator, the apply operator, and the performance operator shown in Figure 7.

The stacking operator is a nested operator that is formed by two stages, the first called stage called base 0, where there are several classifiers in this case there is the RandomForest operator, the J48 operator and the ADTree operator, to then enter the metaclassifier or base 1, where the multilayer perceptrom operator is.

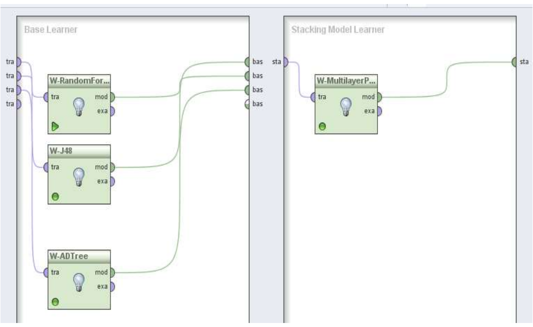

Figure. 8: General visualization of the internal part of the Stacking.

4.3 Interpretation of Results

It is worth mentioning that for the training of the proposed model was made use of 1761 data and the test with 100 new data. The results obtained from the stacking technique (which is a combination of decision trees and neural networks) is 98% accuracy which is very encouraging with respect to the other results obtained by neural networks with 91% accuracy and decision trees with an accuracy of 87%.

The elimination of incomplete data, an adequate handling in training, testing of the data, as well as the use of cross-validation has allowed the obtaining of favorable results

We can also appreciate that the results obtained in each of the labels of the classes are attractive making use of the stacking technique for those who deserted with the label 0 and those who did not desert with the label 1.

The results of the confusion matrix [32], are shown in Table 7, in relation to the stacking technique it can be seen that of 25 students who have deserted 24 have been correctly predicted and incorrectly 1 and of 75 students who continue studying have been predicted correctly 74 and incorrectly 1

Table 6: Confusion Matrix [32], [33]

		Predicted class	
		positive	Negative
Actual Class	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Based on the confusion matrix, accuracy can be calculated using (1), sensitivity using (2) and specificity using (3):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \ldots\ldots(1)$$

$$Sensibility = TP/ (TP + FN) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$Especificity = TN/ (TN + FP) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Table 7: Confusion matrix with Stacking.

Accuracy: 98.00%			
	True 0	True 1	Class precision
pred.0	24	1	96.00%
pred.1	1	74	98.67%
Class Recall	96.00%	98.67%	

The results shown in Table 8 (confusion matrix) in relation to neural networks show that of 25 students who have deserted 21 have been correctly predicted and incorrectly 4 and 75 students who continue studying have been predicted correctly 70 and incorrectly 5.

Table 8: Confusion matrix with Neural Networks

Accuracy: 91.00%			
	True 0	True 1	Class precision
pred.0	21	5	80.77%
pred.1	4	70	94.59%
Class Recall	84.00%	93.33%	

The results shown in Table 9 (confusion matrix) in relation to the decision trees, where it can be seen that of 25 students who have deserted, 19 have been correctly predicted and 6 incorrectly predicted and of 75 students who continue studying have been correctly predicted 68 and incorrectly 7.

Table 9: Confusion matrix with decision trees

Accuracy: 87.00%			
	True 0	True 1	Class precision
pred.0	19	7	73.08%
pred.1	6	68	91.89%
Class Recall	76.00%	90.67%	

Below are the values obtained with each of the models based on Accuracy, sensitivity and specificity

Table 10: Comparison of Precision Metrics

	Arboles de decisión(J48)	Redes neuronales (perceptrom multicapa)	Stacking
Accuracy	87%	91%	98%
Precisión	91.8%	94.6%	98.7%
Sensibilidad	90.6%	93.3%	98.7%
Especificidad	76%	84%	96%

Table 10 shows that accuracy is 91.8% in decision trees, 94.6% in neural networks, and 98.7% in stacking. It can also be observed that decision trees have a sensitivity of 90.6%, neural networks of 93.3% and stacking of 98.7% and, in addition, it is also shown that the specificity in decision trees is 76%, neural networks of 84% and 96% in stacking.

4.4 Discussion

The results obtained from the stacking technique (which is a combination of decision trees and neural networks) is 98% accuracy which is very encouraging with respect to the other results obtained by neural networks with 91% accuracy and decision trees with an accuracy of 87%, while authors such as 18] in their study show that the Naives bayes algorithm gave as a significant useful result an accuracy of 99%, being considered the best algorithm to predict academic desertion and thus be able to arrive at solutions to improve the desertion problem. Similarly, the researchers [19] pointed out that in relation to accuracy the decision tree algorithm was the one that gave the best results with 92.12%; the researchers [22] also showed that the decision tree has an accuracy of 94.30%. Estos resultados son refutados por los investigadores [20] donde realizaron un estudio sobre la deserción y concluyeron que el mejor algoritmo para predecir es el SVM con 89.04%. Finalmente, un estudio similar de [27] hicieron uso de modelos bayesianos y de lógica difusa con una precisión aproximada del 77%. Concluyendo que el uso de modelos bayesianos y lógica difusa son herramientas efectivas para la reducción del índice de deserción escolar en Universidades Púbicas en México.

The results of the confusion matrix [32], are shown in Table 7, in relation to the stacking technique it can be seen that of 25 students who have deserted 24 have been correctly predicted and incorrectly 1 and of 75 students who continue studying have been predicted correctly 74 and incorrectly 1. Other authors such as [24] obtained that the Random Forest algorithm had possible defectors of 79%. of defectors correctly (true negatives); while non-deserters (true positives) have been correctly classified by 26%. Non-deserters, who were classified as deserters (false negatives) corresponds to 74% and deserters, classified as non-deserters (false positives) represent 21%.

After having applied and evaluated the algorithms, it is shown that the accuracy of Decision Trees (J48) is 87%, Neural Networks (multilayer perceptrom) 91% and Stacking 98%. This is contrasted with the study of [26], whose results show that through the four models (Logistic regression, tree decision, KNN, Neural

Network) an accuracy greater than 80% was obtained.

Table 10 shows that the accuracy is 91.8% in decision trees, 94.6% in neural networks and 98.7% in stacking. It can also be observed that decision trees have a sensitivity of 90.6%, neural networks of 93.3% and stacking of 98.7% and, in addition, it is also shown that the specificity in decision trees is 76%, neural networks of 84% and 96% in stacking. However, the authors [17] note that the SVM model with linear kernel obtained better specificity with 0.79 and the Random Forest algorithm obtained sensitivity of 0.947. and concluding that the models chosen for the assembly and the variables related to the first semester were of greater importance than those of the admission exam. Likewise, with the study of [26], it could be concluded that between the four models (Logistic regression, decision trere, KNN, Neural Network) a sensitivity of 80.80% was obtained; 86.90%; 83% and 78.60% Sin embargo, la especificidad fue de 87.30%; 89.50%; 93,10% y 85,10%. Los autores [25] en su estudio determinaron que todos los modelos consiguieron una sensibilidad del 100%, mientras tanto la especificidad dio solo el 18%.

5. CONCLUSIONS

The results of the analysis showed that neural networks with the multilayer perceptrom algorithm have an accuracy of 91%, decision trees with the J48 algorithm have an accuracy of 87%, while Stacking which is a combination of decision trees and neural networks has an accuracy of 98%. From the above, it is concluded that hybrid models based on stacking have a high potential to be able to identify students at risk of leaving the University. Although the results obtained are favorable, it would be appropriate to test the model with a broader data set, to implement the hybrid model based on stacking worldwide.

These results can help teachers, principals and administrators to make timely decisions and can implement appropriate strategies to reduce the dropout rate.

If higher education institutions apply this model, using consistent, up-to-date datasets of students who have dropped out of college, they would benefit the most.

REFERENCES

[1] Miño, M., Factores condicionantes de la deserción universitaria. *Ciencia Latina Revista Científica Multidisciplinar*, 5(4), 5316-5328, 2021.

[2] Cáceres, E., & Alejandra, M. Impacto del COVID-19 en la deserción universitaria de las carreras empresariales. *Revista Científica UNE*, *3*(1), 40-50, 2021.

[3] Miranda Rodriguez, V., & Alarcon Diaz, H. Efectos de los factores de riesgo sobre la interrupción de los estudios en jóvenes universitarios durante la covid-19. *Desde el Sur*, *13*(2), e0021, 2021.

[4] Hanson, M. College Dropout Rates [Internet]. EducationData.org. 2021 [Citado 17 de marzo 2022]. Disponible en: https://educationdata.org/college-dropout-rates

[5] Núñez, A. Deserción y retención: retos en la educación superior. *Revista Científica Retos de la Cienca, 4*(9),15-23, 2020.

[6] Noguera, S. Estudio en Brasil revela que 42% de los alumnos abandonaría universidades privadas por la COVID-19 [Insternet]. Agencia Anadolu. 2020 [Citado 17 de marzo 2022]. Disponible en: https://www.aa.com.tr/es/mundo/estudio-en-brasil-revela-que-42-de-los-alumnos-abandonar%C3%ADa-universidades-privadas-por-la-covid-19/1873890#:~:text=Carrera%20Agencia%20Anadolu-,Estudio%20en%20Brasil%20revela%20que%2042%25%20de%20los%20alumnos%20abandonar%C3%ADa,privadas%20por%20la%20COVID%2D19

[7] Torres Rentería, S., & Escobar Jiménez, C. Determinantes de la deserción y permanencia en la carrera de Medicina: Evidencia del Sistema de Educación Superior ecuatoriano. *Revista Andina de Educación, 5*(1), 1-6, 2022.

[8] Albornoz, N., "Análisis de la deserción estudiantil en institutos de educación superior tecnológicos durante el periodo 2014–2017, región Junín". *Universidad César Vallejo,* Perú, 2019.

[9] Urteaga, I., Siri, L., & Garófalo, G. Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea. *RIED, 23*(2), 147-161, 2020.

[10] Felizzola, H., Jaime, Y., Castillo, A., & Villa, F. Modelo de predicción para la deserción temprana en la Facultad De Ingeniería De La Universidad De La Salle. *Gestión, Calidad y Desarrollo en las Facultades De Ingeniería,* 1-8, 2018.

[11] Daza, A. Un modelo basado en árboles de decisión para predecir la deserción estudiantil en la Educación Superior Privada. *UCV-Scientia, 8*(1), 59-73, 2016.

[12] Alvarado, J. "estudio comparativo del nivel de eficacia en modelos algorítmicos al estimar la deserción de los estudiantes del nivel pregrado en la universidad de huánuco – 2019". *Universidad de Huánuco,* Perú, 2022.

[13] Gulati, H. "Predictive analytics using data mining technique". *In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 713-716). IEEE.*, 2015.

[14] Huillca J and Quispe R, "Sistema inteligente para la predicción del precio diario de las acciones mineras en la Bolsa de Valores de New York usando un modelo híbrido de redes neuronales y máquina de soporte vectorial de regresión", *Universidad Nacional Mayor de San Marcos*, 2019.

[15] V. Deepa and K. Muthamil Sudar and P. Deepalakshmi, "Detection of DDoS Attack on SDN Control plane using Hybrid Machine Learning Techniques", *International Conference on Smart Systems and Inventive Technology (ICSSIT 2018)*, 2018.

[16] Shashi Dahiya and S.S Handa and N.P Singh, "Credit Modelling using Hybrid Machine Learning Technique", *2015 International Conference on Soft Computing Techniques and Implementations- (ICSCTI)*, 2015.

[17] Gamboa, J. y Salinas, J. "Predicción de la situación académica en alumnos de pregrado usando algoritmos de machine learning", Perfiles, Vol. 1, No. 27, 2022, pp.1-7.

[18] Henríquez, C.; Salcedo, D. y Sánchez, G. "El aprendizaje automático en entornos educativos universitarios: Caso deserción académica", Prospectiva, Vol. 20, No. 1, 2022, pp.1-12.

[19] Urbina, A.; Téllez, A. y Cruz, R. "Patrones que identifican a estudiantes universitarios desertores aplicando minería de datos educativa", Revista Electrónica de Investigación Educativa, Vol.23, No.29, 2021, pp.1-15

[20] Fernandez, A.; Preciado, J.; Melchor, F.; Rodriguez, R.; Conejero, J. and Sanchez, F. "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data", IEE Access, Vol.9, No.1, 2021, pp.1-15.

[21] Ma, X. and Zhou, Z., Student pass rates prediction using optimized support vector machine and decision tree. *In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 209-215). IEEE., 2018.

[22] Kemper, L.; Vorhoff, G. and Wigger, B. "Predicting student dropout: A machine learning approach", European Journal of Higher Education, Vol. 10, No.3, 2021, pp.1-20.

[23] Kostopoulos, G., Kotsiantis, S., Ragos, O. and Grapsa, T., Early dropout prediction in distance higher education using active learning. *In 2017 8th International Conference on Information, Intelligence, Systems and Applications (IISA),* pp. 1-6, 2017.

[24] González, J. y Peñaloza, M. "Identificación y predicción de estudiantes en riesgo de deserción académica por medio de modelos basados en machine learning", *Los Libertadores Fundación Universitaria*, 2021, pp.1-16.

[25] Urteaga, I., Siri, L., y Garófalo, G. "Predicción temprana de deserción mediante aprendizaje automático en cursos profesionales en línea", *RIED. Revista Iberoamericana de Educación a Distancia,* Vol. 23, No.2, 2020, pp. 147-167.

[26] Rivera, K. "Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica", *Revista Innovación y Software,* Vol. 2, No.2, 2021, pp.6-13.

[27] Vázquez, P.; Quintero, P. y Alanís, J. "Identificación de causas de deserción en programas de estudio de nivel superior mediante modelos bayesianos y de lógica difusa", *Revista Iztatl Computación*, Vol. 10, No.20, 2021, pp.9-16.

[28] Jalota,C. and Agrawal, R. " Feature Selection Algorithms and Student Academic Performance: A Study", *International Conference on Innovative Computing and Communications, Advances in Intelligent Systems and Computing*, vol.1, 2021, pp. 317–328.

[29] Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms [Internet]. kdnuggets. 2019 [citado 27 abril 2020]. Disponible en: https://www.kdnuggets.com/2019/02/gartner-2019-mq-data-science-machine-learning-changes.html.

[30] Bashir, A. ; Mohammed,M. ; Abedalrazeq , M. and Khalafallah, M., " SVM and Naïve Bayes Stacking Approach for Improving Gene Expression Data Classification Using Logistic Regression", *International Journal of Advances in Soft Computing and its Applications,* Vol.13, 2021, pp. 2074-8523.

[31] Borra, S. and Di Ciaccio, A., “Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data”, *Journal of Rock Mechanics and Geotechnical Engineering*, Vol.13, No.1,2021, pp. 188-201.

[32] Martinez,I.; Viles, E. ; Olaizola,I. “Data Science Methodologies: Current Challenges and Future Approaches”, *Big Data Research,* Vol. 24,No.3,2021, pp.1-18

[33] Laura I. and Santi S., Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications, Switzerland: Springer, 2017.