Original Article

MACHINE LEARNING TO INCREASE APPLICANTS IN THE ADMISSION PROCESS OF A PUBLIC UNIVERSITY IN LIMA-PERÚ

Edward Flores[1*], Justo Solis[1], Juan Grados[2], José Rosales[2], Yeremi Barahona[2], Katherine Llanos[3]

[1]Universidad Nacional Federico Villarreal Grupo de Investigación GISI - EUPG,
[2]Universidad Nacional Federico Villarreal,
[3]Universidad Privada del Norte
Email: eflores@unfv.edu.pe[1]

DOI: 10.47750/pnr.2022.13.S07.503

Abstract

In recent years, admission to public universities by applicants has been a process of increasing competition, because the academic offer has been changing considerably, due to the increase in professional careers in private universities, weakening in In some cases, the study programs of public universities, which is why the present research was proposed, which aims to Implement a predictive model of machine learning to increase the number of applicants to the admission process in a public university, the method used was to use the information from the admissions processes of the years 2018 and 2019 of the public university, before the pandemic, to evaluate the data between seven machine learning classifiers under the conditions of only categorical data, categorical and numerical data and finally data standardized. The results show that the Logistic Regression, Decision tree classification and Random Forest Classification models, in that order, allow the evaluation of the corresponding information, supported by the confusion matrix and the indicator f1-score that allows to properly validate the results for groups that are not homogeneous in the data.

Keywords: Machine Learning, confusion matrix, Logistic Regression, admission process.

Introduction

Admission processes in national universities are increasingly faced with a greater need to compete with private universities due to the growing demand of applicants and the enormous academic offer offered by private universities, which causes a constant concern in public universities who seek to find different market niches to attract applicants and thus be able to cover all the professional careers offered by national universities.

In this regard, starting from the change in the conditions of the courses in higher education institutions from March 2020 to date, due to the contingency caused by the COVID-19 pandemic, has been a challenge for all actors in the process educational, adaptation to the distance education modality, which unexpectedly replaced the traditional face-to-face system [1]. On the other hand, in Ecuador, on access to higher education, a statistical and qualitative analysis of the research methodology was carried out: documentary, statistical and interviews, to elaborate a discussion and recommendations on public policies and compliance with the objectives from 2010 to 2017. The results show how these changes are manifested through the university evaluations and their categorization, the planning of quality management processes in teaching and the relevance of the academic offer to guarantee that the admission system of applicants is meritocratic. Higher education policy has caused a paradoxical situation in this system, both at the national level and within each higher education institution [2].

Higher Education institutions play a transformative role in contemporary society to respond to the social mandate, through policies from governments apply solutions to new challenges, in order to increase the quality, relevance and rigor of access to the students to the University [3]. Public universities in the Mexico City Metropolitan Area accept less than half of their applicants for undergraduate studies. At the Autonomous Metropolitan University (UAM), less than a fifth of the candidates enter. There is little information and attention to the characteristics of applicants to higher education institutions [4]. University admission transcends students from multiple dimensions (organizational, social, personal) that they usually face alone, and their adaptation will largely depend on their abilities to meet and overcome the needs that arise [5].

There is also research that seeks to have an evaluation through a predictive model to evaluate the predictive power that high school academic performance has on academic success at university, as well as some of the implications of its incorporation in the admission process university [6].

Machine Learning (ML) is an artificial intelligence (AI) technique, in which mathematical algorithms are used to learn from data in an effort to formulate an understanding of a particular phenomenon [7].

Regarding this problem encountered, the following question was raised: In what way can a predictive machine learning model be implemented to increase the number of applicants to the admission process in a public university? In the same way, secondary problems were raised: Can the information of applicants from previous years be analyzed to develop a predictive model of machine learning? And also, can a machine learning predictive model be developed to forecast applicants by district for a professional career at a public university?

This approach brought as a general objective: Implement a predictive model of machine learning to increase the number of applicants to the admission process in a public university, and as specific objectives: Analyze the information of applicants from previous years to develop a predictive model of the machine learning and develop a model of machine learning to forecast applicants by district for a professional career in a public university, which promoted this research.

2. Materials and method

All printed material, including text, illustrations, and charts, must be kept within the parameters of the 8 15/16-inch (53.75 picas) column length and 5 15/16-inch (36 picas) column width. Please do not write or print outside of the column parameters. Margins are 1 5/16 of an inch on the sides (8 picas), 7/8 of an inch on the top (5.5 picas), and 1 3/16 of an inch on the bottom (7 picas).

For the present study, it was decided to define what the unit of analysis will be, and in the same way, we proceed to delimit the population to be studied on which the results are to be generalized. Thus, a population is the set of all cases that agree with a series of specifications cited by Lepkowski in 2008 [8]. In accordance with the requirements provided by the university and for the purposes of this study, those applicants to the admission process to a public university who have participated in the admission process in the years 2018 and 2019 will be considered, since due to the examination processes For admission to national universities, the processes are carried out only once a year.

Taking into consideration the aforementioned, those applicants who have taken the admission exam by ordinary modality during the years 2018 and 2019 in the Lima and Callao Regions will be considered, not considering the other types of admission for admission to the university, nor considering the eventual applicants from other regions, with the objective of reducing the bias within the critical mass of data that can cause when processing the largest volume of main information and thus avoid atypical data in the information processing. The university has a wide variety of entry types, under various modalities, however, the largest number of applicants are presented in the ordinary admission exam modality.

The objective of choosing a model selection strategy is to define various procedures for estimating or forecasting data in one or more time series from historical information. These techniques do not try to normalize the data to establish the behavior of one or more variables, but rather perform an analysis for the construction of a conceptual

model in which results are generated through the series. Therefore, for this study the following conceptual models are considered: stationary model, with linear trend and with seasonality [9].

As it is a data analysis investigation, the data-based approach will be used as an instrument, which refers to decisions that are generated based on the analysis and interpretation of specific data rather than on observation.

The information that will be treated for this study corresponds to the ordinary admission processes, for which it will be coordinated with the central admission office of the university. Subsequently, the data will be classified, normalized and standardized to select a training group and a test group, which can be used during the development of the machine learning model to identify the best data model, minimizing the cost of prediction error.

The information will be refined and the data from the ordinary admission processes for the years 2018 and 2019 will be used, considering the regions of Lima and Callao for having a greater number of participants, and in this way, reducing the bias that they may lead to outliers such as prospective applicants in other regions.

The procedure to be carried out for the treatment of the data is indicated below:

To start with the treatment of the information, it is necessary to carry out a preprocessing, which is one of the most important steps in any development and treatment of data for Machine Learning, validating the source data is always the first activity since they are regularly found. incomplete form, or inappropriate formats, which could not be included to be processed by the model.

The next step to be carried out should be the data normalization process, usually they must be adequately standardized so that the process can work in an integral way, not finding data that it cannot validate or classify.

Subsequently, the information must be reviewed to avoid redundancy and in this way it will be necessary to resize the information, in order to avoid collision between variables or information that has been intentionally normalized or added so that it can be processed in the previous stage.

Once the information is in the correct form, a segmentation will be carried out between the total of the information in two groups, which will be a group for training and another group for tests, optimally the most suitable percentage will be sought for perform the division between the total study sample, it is usually recommended that it be a division of 20% for the training and testing group, however, there are different forms of information validation, which will be evaluated depending on the treatment of the model in progress.

Once the training and control groups have been identified, the different algorithms will be compared to determine the best performance and the best approach to the closest probability for the desired forecast. The model will be repeatedly validated to determine the best algorithm that performs the desired forecast.

The following are the algorithms that will be part of the evaluation process for this study:

Logistic Regression, logistic regression models can estimate the probability of different classes. Accident prediction is a binary classification problem in which an event must be predicted as an accident or as an event that is not. Estimate the probability of an accident as a result of the prediction [10].

K-Nearest Neigbors (K-NN), The k-NN algorithm is one of the most recognized and widely used ML algorithms in the area of data classification research. With the advent of big data, the performance and efficiency of the traditional k-NN algorithm is rapidly becoming a critical issue [11].

Support Vector Machine (SVM), The Support Vector Machine (SVM) algorithm is a binary classification model. In two-dimensional space, a straight line makes it the most suitable segmentation line in the middle of the two data classes, and for the high-dimensional data set, it is to establish an optimal decision plane as a reference point of classification [12].

Kernel SVM, kernel-based SVM provides a new horizon for classification that computes the dot product between two feature vectors in a Hilbert space. The function of the kernel is to take data as input and transform it into the required form. Kernels are good when we have no idea about the data, they work fine even with unstructured and semi-structured data. With a suitable kernel function, we can solve any complex problem [13].

Naive Bayes, The Naïve Bayes has proven to be a manageable and efficient method for classification in multivariate analysis. However, the characteristics are often correlated, a fact that violates the Naïve Bayes assumption of conditional independence and can deteriorate the performance of the method. Furthermore, data sets are often characterized by a large number of characteristics, which can complicate the interpretation of the results and slow down the execution of the method [14].

Decision tree classification is the simplest classification algorithm first developed by Quinlan et. Alabama. in 1986. A decision tree is a supervised machine learning algorithm that is used in almost all articles, either as the primary algorithm for software failure prediction or for comparison purposes with other sophisticated approaches [15].

Classification with random forests, the ensemble method we use is called random forest. It is developed by Breiman in 2001 and is used to solve prediction problems [16].

The information provided by the university has been the following types of information:

Table 1. Information provided by the study university

information on the admission process		
Applicant Code	District of Birth	Graduated year
Last name	Country Foreign Nationality	country College
Mother's last name	Place Foreign Nationality	city college
Applicant name	Residence Department	Score
Specialty Code	Province Residence	Condition
Specialty	District Residence	General Order
Entry Modality Code	Address	Faculty Order
Modality Description	Telephone	CP order
Faculty Code	Reference Telephone	Description disability
Faculty Description	Mobile	CONADIS resolution
Sex	Cell Reference	School of Origin
Birth date	E-mail	College Name
Nationality	UNFV link	Native language
Document Type	School of Origin	Native Language Description
Identity Document Number	Name Institution School	Foreign language
Proxy Document Number	College Department	Foreign Language Description
Department	Province College	P1_Number of Applications Desc.
Province Department	District College	P2_Previous_Income

Evaluating Models and Predicting with New Data.

After all the validations of the data group and the test group for each interaction, the aim is to evaluate the error obtained and continue adjusting the model in order to achieve the least bias and the least variance. Within the predictive models of deep learning, an exact prediction model is not found, which is looking for the margin of error bias.

3. Results

As can be seen, various types of information related to the admission process have been reviewed, however, many of these items are not relevant for the present study, such is the case to give some examples such as: telephone (since it is not relevant know their telephone number to know if they are applying and entering the university), country (they are considered only if they are Peruvians and from the Lima and Callao region), language, applicant code (it is a random number generated by the system and not provides any value), name (the admission process does not process data from your applicants when determining admission), among others, therefore, these data will be considered to us to determine the basic criteria necessary to carry out a prediction model.

So what are the data that do add value to the study? In these preliminary results, let's see the following:

Study career, which corresponds to one of the specialties provided by the university and which has places available for applicants to enter the admission contest.

Area to which you are applying, since you want to know where there are more or fewer applicants for the admission process by area of study.

District of residence, since it is important to know the district of origin to determine if there are market niches that allow increasing income to a specific area or career.

School of origin, there are different types of schools, however, for the present study a filtration has been carried out and only the most significant types are allowed: national, private and parochial school.

Sex, which allows determining the trend between the study areas.

Condition, which indicates whether or not you entered the university.

Score obtained, which allows you to identify if you successfully entered the university.

Other elements that could be considered may be the following:

Age, this data is relevant to determine the trend for the number of years the applicant has when selecting a career or area of studies.

Years of graduation, which, similar to the previous one, can determine a trend when applying to university.

However, the latter have not been considered due to the variability found and not being able to define whether they really explain the final result obtained.

The results are described below, previously identifying the information provided by the admission processes of the years 2018 and 2019 of the study university, from which the classification was made, table 2 shows the selected fields:

Table 2. Type of information selected for the study

Admission year	**Type of information**	**Quantity**	**Type of data**
Year 2018 and 2019	Study career	56 specialties	categorical
	Study area	4 areas	categorical
	District of residence	47 districts	categorical
	College of origin	3 types.	categorical
	Sex	2 types	categorical
	Score obtained	0 to 1000	quantitative
	Outcome	2 types	categorical

The data set obtained did not contain any inconsistencies or null values, so it was possible to consider the entire data set without adjustment modifications.

A relevant data for the present study, which is all data are categorical, except for the score obtained when the applicant took the admission exam. In the same way, as it is a classification scenario for the model, the corresponding treatment described in the methodology must be carried out.

It must be taken into account that in the logistic regression models and in the other models to be used, the categorical variables cannot be adequately evaluated, so a conversion must be carried out for each one of them in order to generate variables of the “dummy” type. "Which will allow to develop the model properly. Categorical variables cannot be described on a numerical scale. Since machine learning deals with numbers, transforming each category of these variables into an independent binary variable, also known as a dummy variable, is an alternative solution for introducing categorical variables into machine learning models [17].

Below is an example in Table 3 of what the corresponding conversion would look like.

Table 3. Coding of the area code field in dummy variable

AREA_CODE	**c_1**	**c_2**	**c_3**	**c_4**
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

As can be seen, the area code of the specialties of the admission process is composed of four variables, each variable has only a single combination on the right side for its treatment in the procedures to be developed.

In the same way, to avoid multicollinearity, it is necessary to eliminate a column from those obtained in the conversion process developed with dummy variables, in order to avoid duplication in the identification of variables at the time of model development. Then any column can be removed, leaving only three of them.

Multicollinearity is a phenomenon that occurs when 2 or more independent variables are highly (but not perfectly) correlated. 2 Multicollinearity can inflate (or deflate) the standard errors of the coefficients and, as a result, the coefficients can falsely become insignificant (or significant) [18].

In the case of the other categorical variables, they must follow the same procedure, that is, for example, if the variable degree course has 56 values, this variable, because it is categorical, must be transformed into a dummy variable so that it can be used. will generate 56 additional columns to the initial matrix, one of them will be eliminated, to avoid multicollinearity, as we explained previously. This process will be carried out for each categorical variable.

Therefore, in the end, from the conversion of categorical variables, a matrix will remain with 113 binary columns and a column with a high rank of numerical values that corresponds to the column of the scores obtained in the admission exam, for which, it was identified three different ways to carry out the corresponding evaluation to define a model:

-Evaluate only categorical data.

-Evaluate categorical and non-categorical data.

-Evaluate all standardized data

For which, it was decided to evaluate the models with each of the identified forms. The results for each evaluation are shown below.

Table 4. Evaluation of only categorical data.

Model	**Result**	**2018**					**2019**				
		Precision	**Recall**	**support**	**F1 score**	**accuracy**	**Precision**	**Recall**	**support**	**F1 score**	**accuracy**
Logistic Regression	Accepted	0.67	0.04	424	0.08	0.85	0.62	0.05	283	0.09	0.89
	No accepted	0.86	1.00	2427	0.92		0.89	1.00	2200	0.94	
K-Nearest Neighbors (KNN)	Accepted	0.31	0.08	424	0.13	0.84	0.20	0.05	283	0.07	0.87
	No accepted	0.86	0.97	2427	0.91		0.89	0.98	2200	0.93	
Support Vector Machine (SVM)	Accepted	0.42	0.03	424	0.05	0.85	0.38	0.08	283	0.13	0.88
	No accepted	0.85	0.99	2427	0.92		0.89	0.98	2200	0.94	
Kernel SVM	Accepted	0.43	0.01	424	0.03		0.33	0.04	283	0.07	

	No accepted	0.85	1.00	2427	0.92	0.85	0.89	0.99	2200	0.94	0.88
Naive bayes	Accepted	0.15	0.94	424	0.26	0.22	0.12	0.96	283	0.21	0.16
	No accepted	0.90	0.09	2427	0.17		0.91	0.06	2200	0.11	
Decision tree classification	Accepted	0.23	0.16	424	0.19	0.80	0.23	0.17	283	0.20	0.84
	No accepted	0.86	0.91	2427	0.88		0.90	0.93	2200	0.91	
Random Forests Classification	Accepted	0.25	0.12	424	0.16	0.82	0.24	0.12	283	0.16	0.86
	No accepted	0.86	0.94	2427	0.90		0.89	0.95	2200	0.92	

Table 4 shows us the comparison for each model evaluating the data for the years 2018 and 2019, where the values of the Precision column indicate the proportion that was really correct for those who entered and did not enter the university, it is clearly perceived that For applicants who did not enter, there is a good acceptance of 85% or higher for the year 2018, while for the year 2019 there is an acceptance of 89%, however, it is necessary to evaluate preferably all the values of those applicants who did enter. to the university since they were in smaller numbers compared to the large number of applicants who did not enter.

From the evaluation carried out in the precision field, it can be seen that for the years 2018 and 2019 all the models have low percentages, so it cannot be said that a model has been found that supports the variation between the data.

The results also show us in the accuracy column that the largest number of proposed models have a value greater than or equal to 80%, which is a fairly acceptable accuracy result, (all except for Naive Bayes), however, this value does not save in relation to the specific results, because they are really only indicating that they have been successful in finding applicants who did not enter university, while they cannot predict those students who did.

The recall column shows us the results more clearly, since it is the rate of real hits on the positive values found, where it is clearly seen that the values are very low in all the models for those applicants who managed to enter.

In the same way, the values of the indicator f1-score is very useful to determine the values of unequal classes, which is a very low value within the analysis carried out, practically discarding the results obtained for the group of university entrants.

Table 5. Evaluation of categorical and non-categorical data.

Model	**Result**	**2018**			**2019**

		Precision	**Recall**	**support**	**F1 score**	**accuracy**	**Precision**	**Recall**	**support**	**F1 score**	**accuracy**
Logistic Regression	Accepted	0.96	0.96	424	0.96	0.99	0.91	0.80	283	0.85	0.97
	No accepted	0.99	0.99	2427	0.99		0.97	0.99	2200	0.98	
K-Nearest Neighbors (KNN)	Accepted	0.88	0.89	424	0.88	0.96	0.89	0.69	283	0.78	0.96
	No accepted	0.98	0.98	2427	0.98		0.96	0.99	2200	0.98	
Support Vector Machine (SVM)	Accepted	0.98	0.96	424	0.97	0.99	0.92	0.92	283	0.92	0.98
	No accepted	0.99	1.00	2427	0.99		0.99	0.99	2200	0.99	
Kernel SVM	Accepted	0.77	0.80	424	0.78	0.93	0.65	0.28	283	0.39	0.90
	No accepted	0.96	0.96	2427	0.96		0.91	0.98	2200	0.95	
Naive bayes	Accepted	0.35	0.62	424	0.45	0.77	0.25	0.68	283	0.37	0.73
	No accepted	0.92	0.80	2427	0.86		0.95	0.74	2200	0.83	
Decision tree classification	Accepted	0.99	0.99	424	0.99	1.00	0.98	0.97	283	0.97	0.99
	No accepted	1.00	1.00	2427	1.00		1.00	1.00	2200	1.00	
Random Forests Classification	Accepted	0.97	0.99	424	0.98	0.99	0.95	0.97	283	0.96	0.99
	No accepted	1.00	1.00	2427	1.00		1.00	0.99	2200	0.99	

Table 5 shows us the comparison for each model evaluating the data from the years 2018 and 2019, taking into account that a validation process has been carried out with all the categorical data transformed into dummy variables and the resulting scores have been added to these data. obtained in the admission exam, where the values

of the Precision column indicate the proportion that was really correct for those who entered and did not enter the university, it is perceived that for both types of applicants, entering and non-entering, the values are are above 87%, (less for the Kernel SVM and Naive Bayes models), therefore, it can be said that the precision has increased mainly in the group of applicants who managed to enter the university, since this group represents a much smaller number compared to applicants who did not enter.

The results also show us in the accuracy column that the largest number of proposed models have a value greater than or equal to 95%, which is a fairly acceptable accuracy result, (all except for Naive Bayes). The recall column shows us the results more clearly, since it is the rate of real hits on the positive values found, where it is clearly seen that the values have improved in all the models (except Naive Bayes) for those applicants who managed to enter, which we can think that the prediction models have improved significantly.

The values obtained in this case in the f1-score indicator are much better than those described in the previous table, acceptable values are seen for several models developed in this data analysis to determine which would be the best model to forecast the income of a student to the national university.

Table 6. Evaluation of all standardized data.

Model	**Exam**	**2018**					**2019**				
		Precision	**Recall**	**support**	**F1 score**	**accuracy**	**Precision**	**Recall**	**support**	**F1 score**	**accuracy**
Logistic Regression	Accepted	0.99	0.99	424	0.99	1.00	0.99	0.97	283	0.98	1.00
	No accepted	1.00	1.00	2427	1.00		1.00	1.00	2200	1.00	
K-Nearest Neighbors (KNN)	Accepted	0.69	0.32	424	0.43	0.88	0.51	0.20	283	0.29	0.89
	No accepted	0.89	0.98	2427	0.93		0.91	0.97	2200	0.94	
Support Vector Machine (SVM)	Accepted	0.99	0.97	424	0.98	0.99	0.97	0.98	283	0.97	0.99
	No accepted	1.00	1.00	2427	1.00		1.00	1.00	2200	1.00	
Kernel SVM	Accepted	0.93	0.69	424	0.79	0.95	0.84	0.58	283	0.68	0.94
	No accepted	0.95	0.99	2427	0.97		0.95	0.99	2200	0.97	
Naive bayes	Accepted	0.15	0.97	424	0.27		0.11	0.98	283	0.20	

	No accepted	0.93	0.08	2427	0.15	0.21	0.89	0.03	2200	0.05	0.13
Decision tree classification	Accepted	0.99	0.99	424	0.99	1.00	0.98	0.98	283	0.98	0.99
	No accepted	1.00	1.00	2427	1.00		1.00	1.00	2200	1.00	
Random Forests Classification	Accepted	0.97	0.99	424	0.98	0.99	0.95	0.97	283	0.96	0.99
	No accepted	1.00	1.00	2427	1.00		1.00	0.99	2200	0.99	

Table 6, which corresponds to the evaluation of all standardized data in similar ranges to avoid the dispersion of ranges between columns, shows us the comparison for each model evaluating the data for the years 2018 and 2019, taking into account that it has been carried out a validation process with all the categorical data transformed into dummy variables and to these data the resulting scores obtained in the admission exam have been added, where the values of the Precision column indicate the proportion that was really correct for those who entered and did not enter the university, it is perceived that for both types of applicants, entering and non-entering, the values are above 0.93%, (less for the K-Nearest Neighbors and Naive Bayes models), therefore, it can be say that precision has increased mainly in the group of applicants who managed to enter the university, since this group represents a much smaller number and n comparison with applicants who did not enter.

The results also show us in the accuracy column that the largest number of proposed models have a value greater than or equal to 95%, which is a fairly acceptable accuracy result, (all except for K-Nearest Neighbors and Naive Bayes). The recall column shows us the results with greater clarity, since it is the rate of real hits on the positive values found, where, properly evaluating, there are high values in models that have had low percentages of coincidences, which only affirms that there are few found values.

Making a comparative analysis between the values of table 5 and table 6, which are the best values found for the evaluation of the indicated models, we find that there are three models that best approach the treatment of information, these are Logistic Regression, Decision tree classification and Random Forest Classification, however, to decide which could be the best model for the treatment of categorical data information, we must look at the results of table 3 in these three models, which exclusively classified only categorical data, for which we would only stay with the Logistic Regression model since it better classified the results of the applicants who obtained admission to the university.

The values of the indicator f1-score are in their maximum expression compared to the previous results, (less in K-Nearest Neighbors, Kernel SVM and Naive Bayes), allowing to guarantee a model very close to the real values.

4. Conclusions

According to the analysis carried out, it can be deduced that it is possible to predict whether an applicant can enter the national university of this study, knowing the variables such as: specialty, study area, sex, district of residence, school of origin and the possible score that you expect to obtain in the test, this result may be possible through the predictive models described by priority of Logistic Regression, Decision Tree Classification and Random Forest classification, which allow the handling of categorical variables in a large number of values.

The confusion matrix explained by each model in tables 4, table 5 and table 6, allows us to validate the data coincidences of the proposed models, and that allows us to determine the acceptance values and the rejection values for the coincidences, as well as for false positives and false negatives. In this case, that the data group of the incoming applicants to be evaluated was very small compared to the number of total records, the interpretation of the confusion matrix through the results found in the columns of the indicated tables, allow us to see the approach or distance of the model in front of the data.

For the treatment of categorical data of the study variables, these can work better within the model with quantitative type values to improve the precision of the model and in this way, find the most stable model that is close to the group of values that are desired. predict. According to what has been reviewed in this article, it is necessary to standardize all values in proportional ranges so that the fit values of the model can be optimal.

The value of f1-score, in this type of data analysis, is very necessary because it allows to see the precision and sensitivity in a single metric, quickly visualizing how close or far it is from the real values in the prediction in comparison with the training values of the model. In our case, being two groups with a very unequal distribution of cases, this indicator allows us to identify the differences between the data that are being used to determine whether the model is correct or not.

References

[1] López, T. (2021), La Educación de los jóvenes universitarios en tiempos de la pandemia de COVID-19. el 04 de junio de 2021. Disponible en: https://web.mediasolutions.mx/Notas/?id=20210604003655920 4&temaid=4084.

[2] Latorre-Villacís, V. M. (2020). Reformas universitarias ecuatorianas: el acceso a la educación superior. Panorama, 14(27), 73–88. https://doi.org/10.15765/pnrm.v14i27.1524.

[3] Capote León, Gladys Elena, & Rizo Rabelo, Noemí. (2021). El ingreso a la educación superior en la provincia de Cienfuegos, Cuba. Revista Universidad y Sociedad, 13(2), 283-293. Epub 02 de abril de 2021. Recuperado en 28 de agosto de 2021, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202021000200283&lng=es&tlng=es.

[4] Miller, Dinorah, Garay, Adrián de, & Montoya Zepeda, Iván. (2021). Cruce de desencuentros. Demanda de educación superior y características de los solicitantes de ingreso a la Universidad Autónoma Metropolitana. Revista mexicana de investigación educativa, 26(88), 253-282. Epub 24 de marzo de 2021. Recuperado en 28 de agosto de 2021, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662021000100253&lng=es&tlng=es.

[5] Delgado-García, M., Conde Vélez, S. y Azaustre Lorenzo, M.C. (2021). Validación de un instrumento para detectar necesidades de orientación en alumnado universitario de nuevo ingreso. Revista Española de Orientación y Psicopedagogía, 32(1), 92-115. https://doi.org/10.5944/reop.vol.32.num.1.2021.30742.

[6] Tapasco-Alzate, O.A., Ruiz-Ortega, F. J., Osorio-García, D. y Ramírez-Ramírez, D. (2020). El historial académico de secundaria como factor predictor del rendimiento universitario. Caso de estudio. Revista Colombiana de Educación, 1(81), 147-170. https://doi.org/10.17227/rce.num81-7530.

[7] Asim Suleman A. Alwabel, Xiao-Jun Zeng, (2021). Data-driven modeling of technology acceptance: A machine learning perspective. Doi: https://doi.org/10.1016/j.eswa.2021.115584.

[8] Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. (2014). Metodología de la Investigación - 6ta ed. México D.F., México: Mc Graw Hill Education.

[9] González, C. G. (2017). Tratamiento de datos con R, SPSS y Estadistica. Vigo España: Diaz de Santos.

[10] Wang, J., Song, H., Fu, T., Behan, M., Jie, L., He, Y., & Shangguan, Q. (2021). Crash prediction for freeway work zones in real time: A comparison between Convolutional Neural Network and Binary Logistic Regression model. International Journal of Transportation Science and Technology. https://doi.org/https://doi.org/10.1016/j.ijtst.2021.06.002.

[11] Ali, M., Jung, L. T., Abdel-Aty, A.-H., Abubakar, M. Y., Elhoseny, M., & Ali, I. (2020). Semantic-k-NN algorithm: An enhanced version of traditional k-NN algorithm. Expert Systems with Applications, 151, 113374. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113374.

[12] Zhang, H., Shi, Y., Yang, X., & Zhou, R. (2021). A firefly algorithm modified support vector machine for the credit risk assessment of supply chain finance. Research in International Business and Finance, 58, 101482. https://doi.org/https://doi.org/10.1016/j.ribaf.2021.101482.

[13] Kanwal, A., Mehmood, T., & Butt, M. M. (2021). PLS and kernel SVM based hybrid classifier for discriminating FTIR spectrum data with limited sample size. Chemometrics and Intelligent Laboratory Systems, 215, 104365. https://doi.org/https://doi.org/10.1016/j.chemolab.2021.104365.

[14] Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. Computers & Operations Research, 135, 105456. https://doi.org/https://doi.org/10.1016/j.cor.2021.105456.

[15] Singh, M., & Kumar Chhabra, J. (2021). EGIA: A new node splitting method for decision tree generation: Special application in software fault prediction. Materials Today: Proceedings. https://doi.org/https://doi.org/10.1016/j.matpr.2021.05.325.

[16] Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest based approach for predicting spreads in the primary catastrophe bond market. Insurance: Mathematics and Economics. https://doi.org/https://doi.org/10.1016/j.insmatheco.2021.07.003.

[17] Veiga, R. K., Veloso, A. C., Melo, A. P., & Lamberts, R. (2021). Application of machine learning to estimate building energy use intensities. Energy and Buildings, 249, 111219. https://doi.org/https://doi.org/10.1016/j.enbuild.2021.111219.

[18] Tsagris, M., & Pandis, N. (2021). Multicollinearity. American Journal of Orthodontics and Dentofacial Orthopedics, 159(5), 695–696. https://doi.org/https://doi.org/10.1016/j.ajodo.2021.02.005.