Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 512–517

Procedia Computer Science

www.elsevier.com/locate/procedia

The Second International Workshop of Innovation and Technologies (IWIT 2021)

November 1-4, 2021, Leuven, Belgium

Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test

Jairo Acosta Solano[a], Diana Janeth Lancheros Cuesta[b], Samir F. Umaña Ibáñez[c], Jairo R. Coronado-Hernández[c*]

[a]*Corporación Universitaria Rafael Núñez ,Colombia*
[b]*Universidad Cooperativa de Colombia, Colombia,* [c]*Departamento de Productividad e Innovación*
Universidad de la Costa CUC ,Colombia

Abstract

The purpose of this paper is to evaluate several machine learning models under the CRISP-DM methodology in order to determine, through its metrics, the best model for predicting the performance of high school students in the Colombian Caribbean region in the Saber 11º test, while proposing a new methodology for evaluating the results of the test by regions in order to take into account the socioeconomic particularities of each one of them. The CRISP-DM methodology is taken as a basis due to its maturity, this methodology allows the extraction of business and data knowledge, offers a guide for data preparation, modeling and validation of the models; it is expected that the proposed methodology will be implemented by the Colombian Institute for the Promotion of Higher Education (ICFES), departmental education secretariats and educational institutions. A variety of techniques and tools were used to develop ETL processes to obtain a data set with the most relevant attributes, in order to evaluate four machine learning models developed with the J48 (C4.5), LMT, PART and Multilayer Perceptron algorithms; obtaining that the best data set and the best learning model is obtained using the InfoGain attribute selection method and the LMT decision tree algorithm, respectively. Therefore, this project will facilitate the actors of the National Education System to make decisions for the benefit of students and the quality of education in the country, especially in the Caribbean region.

Keywords: CRISP-DM Methodology, Education, Learning Models, National Education System, Predictive Models.

* Corresponding author. Tel.: +573050899
E-mail address: jcoronad18@cuc.edu.co

1877-0509
10.1016/j.procs.2021.12.278

1. Introduction

The development of a country depends on the actions and decisions taken in accordance with the development of its inhabitants, these aspects are related to the Sustainable Development Goals (SDG) promulgated in the United Nations. This paper seeks to determine a predictive model for one of these objectives, quality education, in the cycle of basic secondary education in the Colombian Caribbean region; However, achieving this SDG will depend on the development and progress achieved in others, such as those related to the end of poverty, zero hunger, health and well-being, decent work, and economic growth [1]. This paper has as its main input the results of the Saber 11th test of the triennium 2017 - 2019 in order to determine the best model to predict student performance, making use of the best attributes of the data set, several of which are closely related to the Sustainable Development Goals. The Colombian state, in order to establish measurement standards of the competences developed by the country's students during their different academic stages, regulates the Saber Test, in charge of the Colombian Institute for the Promotion of Higher Education (ICFES), as an instrument to measure the quality of education in the country. This institute is in charge of evaluating the quality of education at all educational levels, taking as input the databases of tests developed by students in order to support the establishment of policies to improve the educational system [2]. However, most of the educational institutions of the Caribbean Region do not have predictive models that allow taking early actions, according to the characterization of the socioeconomic, institutional and academic attributes of the students of the region; Much of the analysis is carried out based on descriptive analysis of the results of the test, the interested institutions, in general, do not develop machine learning models to determine the incidence of the factors mentioned above, nor do they build machine learning models for predicting student performance and estimating test performance levels. Faced with this situation, there is only one approach made by [3] , where developed a model based on decision trees to classify students above or below the national average in the Saber 11 exams in 2016, this study does not make any comparison of models and does not classify the instances according to the performance measures used by ICFES. It is therefore necessary to understand the general context of the test and the need to measure the quality of education through its performance indicators, understand the different attributes of the data available from the exam in the Caribbean region, carry out the appropriate process of selection of the attributes to be taken into account for the modeling phase, apply supervised learning models J48 (C4.5), LMT, PART and Multilayer Perceptron in order to evaluate which one makes a better prediction through the best socioeconomic attributes identified in this paper.

2. Methodology

This paper is developed using the CRISP-DM (Cross Industry Process Model for Data Mining) methodology that follows a goal-oriented approach, it is a mature approach that continues to be widely accepted in data mining projects through data machine learning mining algorithms [4]. This methodology provides a life cycle approach in applied artificial intelligence projects [5] and it is considered as the ideal methodology for the knowledge discovery process in databases (KDD) [6]. CRISP-DM has several characteristics that render it useful for evidence mining. It provides a generic process model that holds the overarching structure and dimensions of the methodology. The methodology then provides for specialization according to a pre-defined context. To justify the prediction success is specified in Figure 1 the different phases of the methodology and the relationship between them [7] .

Fig.1 Phases of the CRISP-DM Methodology
Source: Authors. Adapted from [5]

3. Contribution Development

3.1. Saber 11° test Overview

The Colombian government, through the Colombian Institute for the Evaluation of Higher Education (ICFES), is in charge of determining the factors that have an impact on the quality of education in all the different grades: primary, secondary, middle and higher. The following table shows the different types of exams that ICFES develops at each educational level in order to monitor the quality of education and educational institutions.

Table 1. Tests applied by ICFES in Colombia

PRIMARY AND SECONDARY	MIDDLE SCHOOL	SUPERIOR
• Saber 3°, 5° y 9° • Avancemos 4°, 6° y 8°	• Presaber • Saber 11° • High school validation	• Saber T y T Saber Pro • Saber T y T Saber Pro-Abroad

Source: Authors with information from the ICFES portal [8]

for this paper the following ítems were considered.

- Obtaining data from the ICFES database system for the years 2017 to 2019.
- Filtering the data by the departments of the Caribbean region: Atlántico, Bolívar, César, Córdoba, La Guajira, Magdalena, San Andrés and Sucre
- Relevant data extraction, transformation and data cleaning.

3.2. Activities for models performance validation in test, Data Mining and Dataset Description

It is necessary to develop the following activities in order to validate the models and find the one that best fits the attributes of the data set:

- Integration of data in a single data set.
- Selection of attributes and instances that will be part of the models.
- Application of decision trees (J48 and LMT), constraints (PART) and neural networks (MLP) machine learning algorithms.
- Choice and application of the model validation method.
- Analysis of the metrics to determine the best model with which the performance in the Saber 11th test can be predicted.

Once the origin of the data has been identified, the data corresponding to the tests results in 2017, 2018 and 2019 were downloaded. The loading process is done successfully in the Power BI Desktop application as shown in figure 3. The data set is divided into six blocks or dimensions, these blocks are: Personal information, Contact information, Socio-economic information, School information and Citation data and results. R language is used [9] in order to obtain an initial data set to work in WEKA and a test set is also obtained in order to demonstrate the predictions developed by the model and their consistency with the evaluation metrics of the examples. Once the CSV file is loaded, we proceed to review what type of data R identifies for each of the variables, all attributes with text-type values are identified as factor type variables, the AGE attribute and PERCENTILE_GLOBAL are correctly identified as numeric data, then makes a review of the levels (Levels) of the categorical variables and reorganizes those (if necessary) that are ordinal type in order to follow the order established by the ICFES in its factors

3.3. Dataset Analysis

In order to show the national average percentiles of the qualifications of each knowledge area, a Power BI dashboard is developed for the results at the national level to compare them with those of the Caribbean region. In figure 2, approaching the nearest integer, the national result is at the 50th percentile where the mean and the median of the scale are located, indicating that half of the students are below the national average and the other part above

this, these results include schools of an official and unofficial (private) nature. Specifically, the results of the Caribbean region are 7.8 percentiles below the national average as shown in Figure 2, being an indication of the efforts that the region must make to achieve the national results, the region has the challenge of improving the socioeconomic conditions and the quality of education with the collaboration of all the actors in the process.

a)

50,34 NATIONAL AVERAGE CRITICAL READING	50,36 NATIONAL AVERAGE MATH	50,40 NATIONAL AVERAGE NATURAL SCIENCES
50,39 NATIONAL AVERAGE SOCIAL SICENCES	50,38 NATIONAL AVERAGE ENGLISH	50,24 - GLOBAL NATIONAL AVERAGE

b)

43,48 NATIONAL AVERAGE CRITICAL READING	42,96 NATIONAL AVERAGE MATH	43,70 NATIONAL AVERAGE NATURAL SCIENCES
43,10 NATIONAL AVERAGE SOCIAL SICENCES	43,65 NATIONAL AVERAGE ENGLISH	42,37 - GLOBAL NATIONAL AVERAGE

Fig. 2. a) National average of the percentiles of the evaluated áreas. b) Average percentiles of the evaluated áreas - Caribbean Region
Source: Authors

3.4. Machine Learning Models

Once a set or series of data sets suitable for the modeling process is available, through the selected machine learning algorithms, it is necessary to know the particularities of each algorithm, establish the validation metrics to use, build the models in WEKA and subsequently evaluate their reliability in predicting student performance in the Saber 11 test. Figure 4 shows different variables as Genre, Birth date, Ethnicityand numer of persons in the family

Fig. 3. Data uploaded to Power BI Desktop
Source: Authors

4. Results

For this research, four machine learning models were developed: two decision tree algorithms, J48 and LMT; a model of decision rules, PART; and a neural network model, Multilayer Perceptron. A model must be built from the application of the selected techniques on the data set, therefore, the confusion matrix is only an alternative to verify the fit of the models [10] The confusion matrix shows through a contingency table the errors and successes made when using a classifier, the best way to understand it is through a model that has two values in the class attribute, as shown in table 2 with the set of the present investigation as an example [11]

Table 2. Confusion Matrix. Source: Authors

Prediction set		Correct set
satisfactory or advanced	Low or insufficient	
True-Positive	False-Negative	satisfactory or advanced
False-Positive	True-Negative	Low or insufficient

Where True positives (VP): these are the instances of the class "Successful or advanced" that are well classified as of that set. True negatives (VN): these are instances correctly classified as of the "Minimal or insufficient" class that actually belong to this set. False negatives (FN): these are instances incorrectly classified as being of the “Minimal or insufficient” class when they actually belong to the “Satisfactory or Advanced” set. False positives

(FP): these are instances incorrectly classified as being of the set "Satisfactory or advanced" when in fact they belong to the class "Minimal or insufficient". In Table 3, the results of the selected data set are extracted using the InfoGain method. Observing the curve, it is evident that the best models are obtained by applying the LMT and MLP algorithms. The accuracy of the LMT model is highlighted with 73.6% of well-classified instances of the validation set; the precision of the model indicates that it makes correct predictions of the instances in a percentage of 72.9%, the highest value of all the evaluated machine learning algorithms, in addition, the model obtained with LMT would have a lower rate of false positives when predicting examples of the class "Satisfactory or advanced" and the class "Minimal or insufficient".

Table 3. Metrics for the IGAE set

Data set	Algorithm	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Training time (seconds)
IGAE	*J48*	0.724	0.735	0.286	0.721	0.735	0.728	0.773	1.80
IGAE	*LMT*	0.736	0.752	0.281	0.729	0.752	0.74	0.813	2,523.82
IGAE	*PART*	0.715	0.725	0.296	0.711	0.725	0.718	0.768	1,237.87
IGAE	MLP	0.729	0.778	0.32	0.709	0.778	0.742	0.808	2,016.83

Source: Authors

For the test process, or small-scale implementation, the model developed for the LMT algorithm will be chosen, applying it to a test set corresponding to the data of the Saber 11th test results developed in the first semester of 2019.

4.1. Development of the predictions in the WEKA application with the test data set

In order to test the performance of the model, it is loaded into the WEKA application, for which the following steps are developed: 1. The WEKA application opens, followed by the Explorer tool. 2. The dataset with which the training and validation of the model was developed is loaded in the Preprocess tab. 3. In the Clasify tab, the model developed with the LMT algorithm is loaded into the Result list work area. 4. In the Test options section, select the Supply test set option and select the test data set. Previously to this data set the values of the class have been assigned the character "?" which WEKA identifies as an empty field. 5. In the same previous section, the other options are displayed by clicking on the More Options button and then on the Output predictions option we choose that the prediction output is in PlainText format. 6. Finally, right-click on the model and choose the Re-evaluate model on current test set option to obtain the corresponding predictions as shown in figure 4.

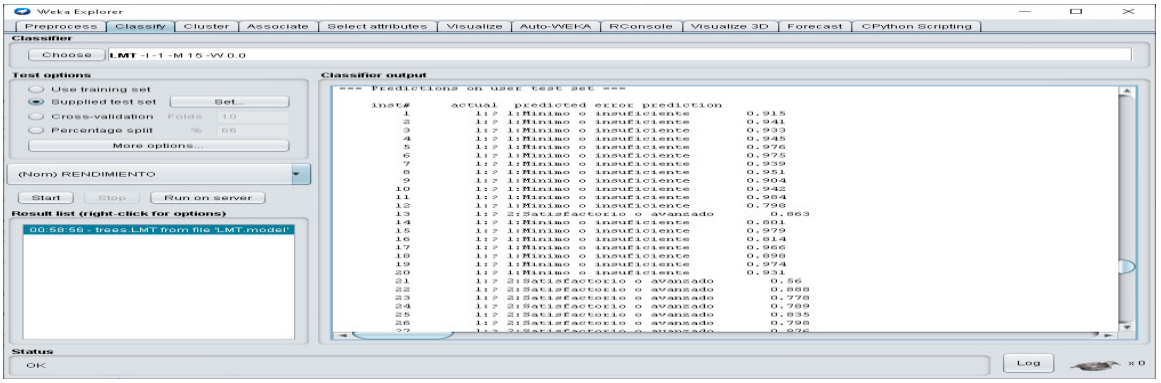

Fig 4. Results of predictions in WEKA
Source: Authors

It can be seen that the output of the prediction is presented in tabular form, the first column is the number or index of the evaluated instance, the second corresponds to the current value of the class label, the third column shows the value of the prediction of the label and the last an estimate of the prediction error.

5. Conclusions

In this research work, different analytical tools have been applied to determine, based on the CRISP-DM methodology, a model and a regional modeling methodology that allows identifying the performance of students in the Colombian Caribbean region in order to predict those students with outstanding performance on the Saber 11th test [4] since it is considered the ideal methodology for the extraction of knowledge from the data [6]. During the preparation phase, it was possible to obtain a dataset that has the best evaluation metrics for the machine learning models trained and validated in the project, as well as a much more manageable dataset that requires less training and validation time. For the modeling and evaluation phases, four machine learning algorithms were trained and validated: J48 (C4.5), LMT, PART and MLP; When performing the validation of these, the Logistic Model Tree (LMT) machine learning model is identified as the one that offers the best metrics among all the models, these models were trained and evaluated with conservative parameters that come by default in WEKA to make the prediction of the performance of students in the Caribbean region, the application of this model will allow the actors of the educational system of the Caribbean region to take intervention actions with those students who have a prediction of minimal or insufficient performance. The main structure of the LMT tree determines that the attributes with the highest information gain are the nature of the school of those evaluated, having an Internet connection, and the school day of these, 25 decision trees are generated in the leaves of the tree of LMT decision to classify the instances of the training set into the two classes or labels of the data [12]. Due to the complexity of the trees generated by the J48 and LMT algorithms, a tree is developed applying the rpart library of the R language [13], where it can be shown, at a higher level of abstraction, which factors have an impact on the performance in the Saber 11th test. Likewise, in order to show that other aspects may have an impact on performance, a series of association rules are developed with the a priori algorithm, with a minimum confidence of 60%, these 20 rules have as a consequence the class of global performance in the Saber 11th test, aspects such as the use of the Internet, the availability of computer equipment, the area of location and nature of the school, and age, largely determine the performance of students.

References

[1] Isis Gómez López, *Desarrollo sostenible*. Elearning, 2020.
[2] ICFES, “ICFES. (2019b). Guía de orientación Saber 11.o 2020-1.” .
[3] R. Ricardo Timarán-Pereira, J. Caicedo-Zambrano, and A. Hidalgo-Troya, “Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°,” *Rev. Investig. Desarro. E Innovación*, vol. 9, no. 2, 2019.
[4] W. Y. Ayele, “Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020.
[5] R. Wirth, “CRISP-DM : Towards a Standard Process Model for Data Mining,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000.
[6] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019.
[7] R. C. Q. Jordi Gironés Roig, Jordi Casas Roma, Julià Minguillón Alfonso, *Minería de datos*. Editorial UOC, 2017.
[8] ICFES, “Icfes Instituto Colombiano para la Evaluación de la Educación - Portal Icfes.”
[9] S. U. Ibáñez and Jairo R. Coronado-Hernández, “Código desarrollado en R para el proceso ETL.”
[10] C. L. Corso, “Aplicación de algoritmos de clasificación supervisada usando Weka,” *Univ. Tecnológica Nac. Fac. Reg. Córdoba*, p. 11, 2009.
[11] J. R. (2021): Umaña, Samir; Coronado-Hernández, “Métricas de evaluación de los modelos. figshare. Dataset.”
[12] J. R. Umaña, Samir; Coronado-Hernández, “Estructura del Logistic Model Tree. figshare. Dataset.”
[13] J. R. (2021): Umaña, Samir; Coronado-Hernández, “Código en R y gráfica del árbol generado por el algoritmo rpart. figshare. Dataset.”