Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

https://doi.org/10.23913/ride.v12i24.1180

Artículos científicos

Algoritmos de aprendizaje automático para la predicción del logro académico

Machine Learning Algorithms for Predicting of Academic Achievement

Algoritmos de aprendizado de máquina para previsão de desempenho acadêmico

Miguel Ángel Morales Hernández
Colegio de Postgraduados, Campus Montecillo, México
morales.miguel@colpos.mx
https://orcid.org/0000-0002-0351-9356

Juan Manuel González Camacho
Colegio de Postgraduados, Campus Montecillo, México
jmgc@colpos.mx
https://orcid.org/0000-0001-5479-7316

Héctor Robles Vásquez
Planeación y Evaluación del Consejo Nacional de Fomento Educativo, México
hroblesvasquez@gmail.com
https://orcid.org/0000-0001-9759-309X

David H. del Valle Paniagua
Colegio de Postgraduados, Campus Montecillo, México
dhvallep@colpos.mx
https://orcid.org/0000-0003-4383-4323

José Rafael Durán Moreno
Consultor independiente, México
rduran1091@gmail.com
https://orcid.org/0000-0002-7886-6408

Revista Iberoamericana para la Investigación y el Desarrollo Educativo
ISSN 2007 - 7467

Resumen

En esta investigación se implementaron dos clasificadores de aprendizaje automático, una red neuronal multicapa (perceptrón multicapa [MLP]) y un modelo de potenciación del gradiente (GB), para predecir el grado de logro académico en las asignaturas de español y matemáticas de alumnos de sexto de primaria (2008) y tercero de secundaria (2011) con base en variables contextuales obtenidas de los Exámenes Nacionales del Logro Académico en Centros Escolares (Enlace) del estado de Tlaxcala, México. Se consideraron 13 variables de entrada y la importancia relativa de éstas, se determinó por medio del algoritmo bosque aleatorio (RF). Los clasificadores MLP y GB se entrenaron y probaron con un conjunto de datos de 11 036 registros de estudiantes que permanecieron en el sistema escolar de 2008 a 2011. Los modelos se entrenaron y probaron en predicción para 2008 y 2011. En español MLP fue superior a GB con una precisión global de clasificación (PG) de 70.1 % en 2008 y 61.1 % en 2011. GB obtuvo mejores resultados en matemáticas con una PG de 68.8 % en 2008 y 63.5 % en 2011. Se observó que el puntaje en español tiene una fuerte asociación con el grado de logro académico en matemáticas. Los puntajes en español y matemáticas tuvieron mayor importancia relativa con respecto a los factores contextuales analizados como: sexo, beca, turno de la escuela. En la población de alumnos analizada se observó que en español y matemáticas la proporción de mujeres es mayor a la proporción de hombres en los grados de logro académico elemental, y bueno o excelente; en contraste, en ambas asignaturas esta proporción se invierte con el grado de logro insuficiente.

Palabras clave: aprendizaje supervisado, árboles de decisión, contexto escolar, redes neuronales artificiales, validación cruzada.

Abstract

In this research, two machine learning classifiers were implemented, a multilayer perceptron (MLP) and a gradient boosting model (GB), to predict the degree of academic achievement in Spanish and mathematics of basic education students in two stages, sixth of primary (2008) and third of secondary (2011), based on contextual variables obtained from the Enlace test of the state of Tlaxcala, Mexico. Thirteen input variables were considered. The relative importance of these was determined by the random forest (RF) classifier. MLP and GB classifiers were trained and tested with a dataset of 11 036 records of students who remained

in the school system from 2008 to 2011. The models were trained and tested in prediction for 2008 and 2011. In Spanish MLP outperformed GB with a global classification accuracy (PG) of 70.1 % in 2008 and 61.1 % in 2011. GB obtained better performance in mathematics with a PG of 68.8 % in 2008 and 63.5 % in 2011. It was observed that the score in Spanish has a strong association with the degree of academic achievement in mathematics. Scores in Spanish and mathematics have greater relative importance with respect to contextual factors considered as sex, scholarship, school shift, and so on. In the population of students analyzed, it is observed that, in Spanish and mathematics, the proportion of women is higher than the proportion of men in achievement levels 1 (elementary) and 2 (good or excellent); in contrast, in both subjects this proportion is reversed at achievement level 0 (insufficient).

Keywords: supervised learning, decision trees, school context, artificial neural networks, cross validation.

Resumo

Nesta pesquisa, dois classificadores de aprendizado de máquina, uma rede neural multicamada (multilayer perceptron [MLP]) e um modelo de potenciação de gradiente (GB), foram implementados para prever o grau de desempenho acadêmico nas disciplinas de espanhol e matemática de alunos do ensino médio. sexta série (2008) e terceira série (2011) com base em variáveis contextuais obtidas nos Exames Nacionais de Desempenho Acadêmico nas Escolas (Enlace) do estado de Tlaxcala, México. 13 variáveis de entrada foram consideradas e sua importância relativa foi determinada usando o algoritmo Random Forest (RF). Os classificadores MLP e GB foram treinados e testados com um conjunto de dados de 11.036 prontuários de alunos que permaneceram na rede escolar de 2008 a 2011. Os modelos foram treinados e testados em previsão para 2008 e 2011. Em espanhol, o MLP foi superior ao GB com uma precisão geral de notas (GP) de 70,1% em 2008 e 61,1% em 2011. GB teve um desempenho melhor em matemática com um GP de 68,8% em 2008 e 63,5% em 2011. A pontuação em espanhol mostrou ter uma forte associação com o grau de desempenho acadêmico em matemática. Os escores em espanhol e matemática tiveram maior importância relativa em relação aos fatores contextuais analisados como: gênero, escolaridade, turno escolar. Na população de alunos analisada, observou-se que em espanhol e matemática a proporção de mulheres é maior do que a proporção de homens no ensino

CC BY

fundamental e nas notas de desempenho acadêmico bom ou excelente; e essa proporção se inverte com o grau de realização insuficiente.

Palavras-chave: aprendizagem supervisionada, árvores de decisão, contexto escolar, redes neurais artificiais, validação cruzada.

Fecha Recepción: Octubre 2021 **Fecha Aceptación:** Abril 2022

Introduction

The evaluation of student learning through large-scale tests (state or national) allows information to be obtained about their degree of academic achievement and the associated contextual variables. The Organization for Economic Cooperation and Development [OECD] (2005) found evidence of how factors such as the school context, school supplies and processes are related to the students' learning process. Mexico began using standardized tests to measure the academic achievement of students in the last two decades. The Ministry of Public Education (SEP) has databases of students who enroll annually at each educational level and the results of the tests that are applied at the national level, such as the Educational Quality and Achievement Examinations (Excale), National Evaluation of Achievement Academic in Schools (Link) or at an international level such as the International Program for Student Assessment (PISA) (National Institute for the Evaluation of Education [INEE], 2019).

The Enlace test was applied on a census basis starting in 2006 to all students from third to sixth grade of primary school and all three years of secondary school. In 2008 it was applied to all three years of high school. The objective of this test was to evaluate the academic achievement of the students in the subjects of Spanish and mathematics; the benchmark for this test is the national curriculum (Martínez, 2015). Link results are measured on a standardized scale that ranges from 200 to 800 points and has four levels of achievement (0 = Poor, 1 = Elementary, 2 = Good, 3 = Excellent). The information has served as support for teachers to compare the results of their school with others with similar characteristics and identify curricular content that students did not acquire in order to take pertinent actions (SEP, 2008). Along with the Enlace test, questionnaires were applied to a sample of students, parents, teachers and directors of the schools that were included in the test to find out personal characteristics, family environment, reading habits, housing characteristics , school infrastructure and teaching methods, and identify limiting factors associated with learning.

Applications of machine learning models to assess school performance are reported in the literature. Đambić, Krajcar and Bele (2016) applied a logistic regression model for the early detection of students with performance problems in a computer course. The model obtained a classification error of 19.0%. Ray et al. (2020) used two classification models (random forest and support vector machine) to predict the school performance of a group of university students based on input variables such as: sex, hours of study, percentage of class attendance, and income family monthly. The random forest model obtained a global classification accuracy of 94.0% and the support vector machine of 79.0%. For their part, Altabrawee, Osama, and Qaisir (2019) applied four machine learning algorithms (neural network, Bayesian decision, decision trees, and logistic regression) to predict student performance in a computer course based on the use of the Internet as a learning medium; Some variables considered were the time spent on social networks, hours of study, sex, education of family members. The neural network achieved the best performance, with an overall accuracy of 77.0% and an area under the AUCROC curve of 0.807.

In Mexico, related work has been carried out to evaluate the effect of factors external and internal to the school on the academic achievement of students. Fernández (2003) used indices of global family capital (educational level of the mother, comfort equipment in the home, availability of books and computers), of the sociocultural context of the school and of the organizational climate; Hierarchical analysis was applied to assess learning in Spanish and mathematics. This author reports that an increase in the global family capital index affects the increase in performance in Spanish and mathematics; however, when housing deprivation is widespread (illiteracy, low income) student outcomes are low.

In this research, the following objectives were proposed, firstly, to implement two supervised machine learning classifiers, namely, a multilayer neural network and a gradient boosting algorithm, to predict the degree of academic achievement (0: insufficient, 1: elementary, 2: good or excellent) in the subjects of Spanish and mathematics of sixth graders (2008) and third graders (2011) in the state of Tlaxcala based on data from the Enlace test; second, to compare the degree of academic achievement in Spanish and mathematics in 2008 and 2011; and third, to determine the relative importance of 13 predictor variables in the classification of academic achievement. The predictor variables were math score, Spanish score, scholarship, school shift, support, type of location, gender, type of school, size of location, marginalization, geographic location (altitude, latitude, and longitude).

CC BY

Materials and methods

Data collection and preparation

The database of academic records used in this study (2008-2011) for the state of Tlaxcala corresponds to a subset of the national database of the Enlace test that was applied from 2006 to 2014 to all students in the third year of primary school. third year of high school, whose purpose was to generate information for parents, teachers, managers and society in general about the academic achievement of students in the educational system in Spanish and mathematics (SEP, 2008). Data records were stored by student, application year, educational level, and grade; with data on the academic achievement of the students (table 1), score obtained in the test (scale from 200 to 800), scholarship holder, shift, type of support and geographic location of the school.

Tabla 1. Intervalos de puntajes para determinar la clase o nivel de logro académico en español y matemáticas

Grado y año	Español		
	0†	1¶	2§
6.º 2008	≤413.85	(413.85, 581.62)	(581.62, 714.01)
7.ºÞ 2009	≤446.31	(446.31, 593.18)	(593.18, 735.69)
8.º¤ 2010	≤445.08	(445.08, 592.41)	(592.41, 735.35)
9.º†† 2011	≤462.94	(462.94, 608.22)	(608.22, 749.18)
	Matemáticas		
6.º 2008	≤412.62	(412.62, 608.13)	(608.14, 735.70)
7.ºÞ 2009	≤507.27	(507.27, 634.85)	(634.85, 737.26)
8.º¤ 2010	≤505.39	(505.39, 634.05)	(364.05, 737.32)
9.º†† 2011	≤525.99	(525.99, 657.03)	(657.03, 762.21)

†0 = Insuficiente; ¶1 = Elemental; §2 = Bueno o excelente. Þ7.° = 1.º de secundaria, ¤8° = 2.º de secundaria, ††9.º = 3.º de secundaria.

Fuente: Prueba Enlace 2008-2013 (SEP, 2008)

From the information available for the state of Tlaxcala, the subset of students who took the Enlace test during four consecutive years (2008 to 2011) from sixth grade to third grade of secondary school was selected; this period marks the beginning of an educational

trajectory (the passage from primary school to third year of secondary school). The selected context variables are: score in Spanish, score in mathematics, scholar, school shift, type of support, type of locality, sex, type of school, size of locality measured by its inhabitants, level of marginalization, location geographic (latitude, longitude and altitude) of the municipality where the school is located; as well as the scores in Spanish and mathematics (Table 2).

Tabla 2. Variables contextuales y localización geográfica seleccionadas del Estado de Tlaxcala, México.

Variable	Descripción	Valores
n_esp	Logro académico en español	0 = Insuficiente; 1 = Elemental; 2 = Bueno o excelente
n_mat	Logro académico en matemáticas	0 = Insuficiente; 1 = Elemental; 2 = Bueno o excelente
p_esp	Puntaje en español	De 200 a 800 puntos
p_mat	Puntaje en matemáticas	De 200 a 800 puntos
becario	Condición de becario	0 = No becario, 1 = Becario
turno	Turno de la escuela	0 = Matutino, 1 = Vespertino
t_sost	Tipo de sostenimiento	0 = Público, 1 = Privado
t_loc	Tipo de la localidad	0 = Urbano, 1 = Rural
sexo	Sexo	0 = Hombre, 1 = Mujer
t_esc	Tipo de escuela	1 = General, 2 = Indígena, 3 = Conafe, 4 = Particular, 5 = Telesecundaria, 6 = Técnica
t_loc	Tamaño de la localidad	1 = menos de 100 habs., 2 = 100 a 249 habs., 3 = 250 a 499 habs., 4 = 500 a 2499

CC BY

		habs., 5 = 2500 a 14 999 habs., 6 = 15 000 ó más habs.
n_mar	Nivel de marginación	1 = Muy alto, 2 = Alto, 3 = Medio, 4 = Bajo, 5 = Muy bajo
lat	Latitud	Decimal
lon	Longitud	Decimal
alt	Altitud	Metros sobre el nivel del mar
pob	Población	En miles de habitantes

Fuente: Elaboración propia con base en los formatos F911 de la prueba Enlace (SEP, 2008) y con datos del Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades (Inegi, 2020)

Table 3 presents the description of the four data sets analyzed by subject and year for the same school population. In order to reduce the imbalance between classes or levels of academic achievement, the good and excellent levels were grouped in level two of academic achievement.

Tabla 3. Descripción de los conjuntos datos analizados por asignatura, año y niveles de logro académico en el estado de Tlaxcala para un total de 11 036 registros de estudiantes

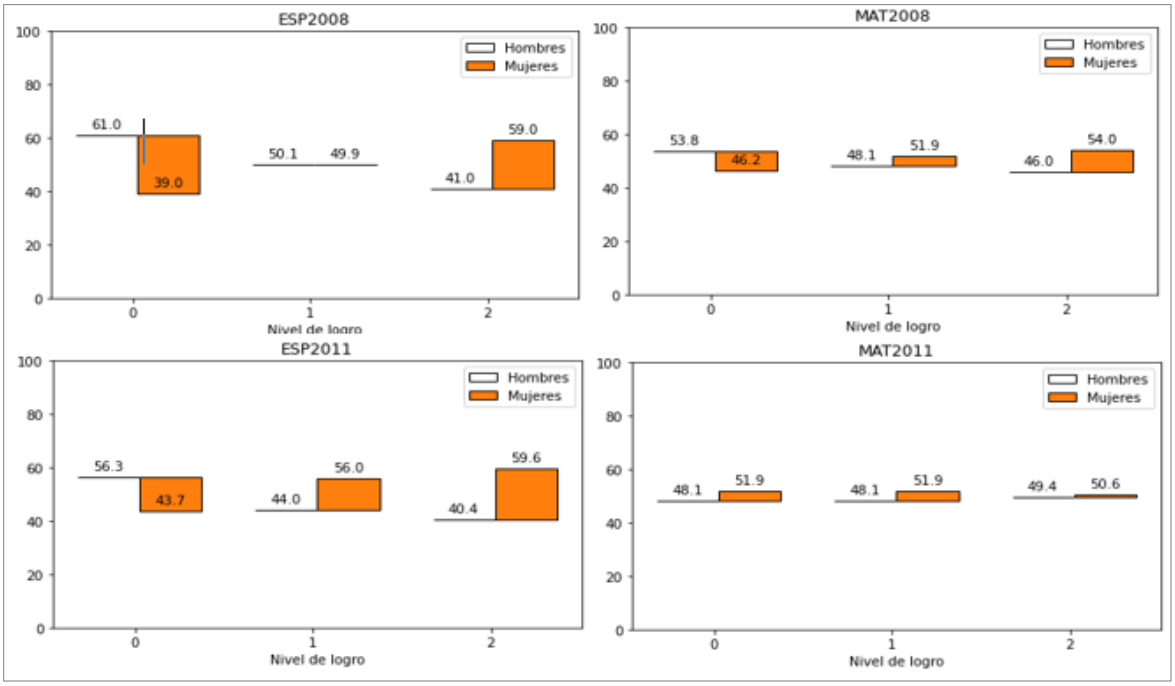
Año y asignatura evaluada	Niveles de logro académico		
	0[†]	1[¶]	2[§]
ESP2008[Þ]	1274	5949	3813
ESP2011[¤]	4194	5362	1480
MAT2008[††]	1544	6039	3453
MAT2011[¶¶]	6254	3608	1174

[†]0 = Insuficiente, [¶]1 = Elemental, [§]2 = Bueno o excelente,
[Þ]Español 2008, [¤]Español 2011, [††]Matemáticas 2008,
[¶¶]Matemáticas 2011.

Fuente: Elaboración propia con base en los datos en Enlace 2008 y 2011 (SEP, 2008)

Figure 1 shows the distribution of achievement levels by gender. In ESP2008 there is a higher proportion of women in class 2 and in MAT2008 they also excel in classes 1 and 2. The same is observed in ESP2011 and in MAT2011, where women excel in both classes.

Figura 1. Niveles de logro académico por sexo para los cuatro conjuntos de datos: ESP2008, ESP2011, MAT2008 y MAT2011



Fuente: Elaboración propia

From the preliminary information of the Enlace test for Tlaxcala, 16 records that had the same information in all the variables were eliminated. The categorical predictor variables t_esc, t_loc and n_mar were transformed into indicator variables for analysis. The four refined data sets consist of 22 predictor variables and 11,036 student records, which represents 44.3% of a total of 24,875 student records evaluated in 2008, in Tlaxcala. For the analysis, four data sets were generated: ESP2008, ESP2011, MAT2008 and MAT2011; In each data set, the class labels correspond to the level of achievement of the students obtained in the corresponding subject: n_esp or n_mat.

Statistical Analysis System (SAS) v. 9.4. To run the machine learning algorithms, Python 3.8 software and the Scikit-learn function library were used to execute the codes and generate results.

CC BY

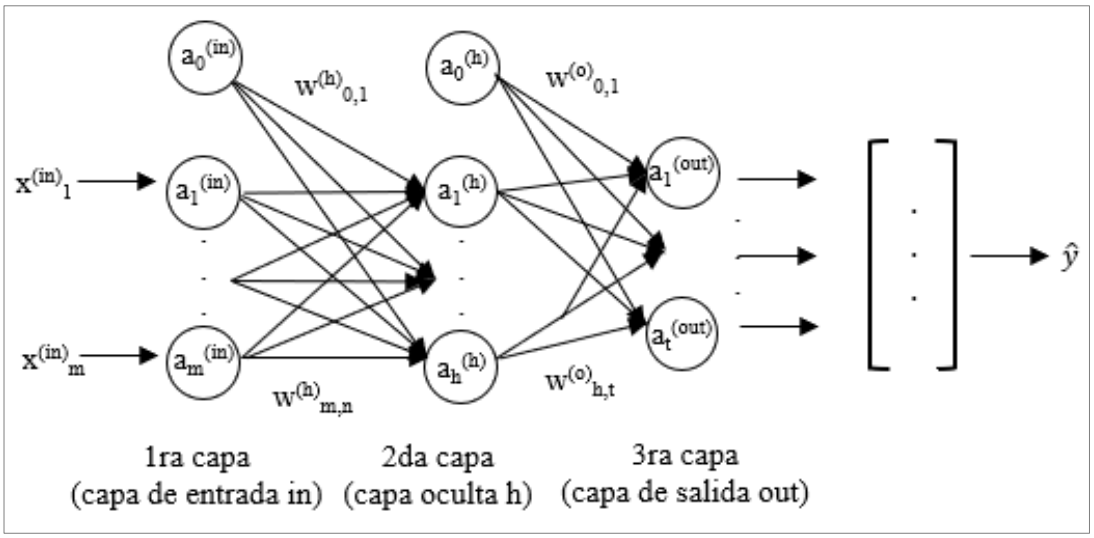
Machine learning classifiers

The purpose of supervised machine learning classifiers is to predict a target class from input variables or features. In this work, three supervised learning classifiers multilayer perceptron [MLP] and gradient boosting [GB] were implemented to predict the level of academic achievement, and random forest [RF]. acronym in English] to determine the relative importance of the predictor variables.

Multilayer perceptron

The MLP classifier is a network of neurons connected by weights or parameters, structured in an input layer (in), one or more hidden layers (h) and an output layer (out). The basic architecture of MLP consists of three layers (figure 2). The more layers the network has, the more complex it is, and the greater the ability to solve complex problems. (Borkar y Rajeswari, 2014).

Figura 2. Arquitectura del clasificador perceptrón multicapa



Fuente: Raschka y Mirjalili (2017)

The operation of the MLP classifier consists of, that given a vector of input data x_i with m predictor variables, M functions are inferred in the hidden layer; then, in the output layer, the predicted response is determined by applying the inferred functions in the hidden layer through a nonlinear transformation (González et al., 2012). All layers are forward connected, and are represented by a weight matrix W, which is initialized with small random

CC BY

values. The hidden layer is activated first $a_1^{(h)}$ by means of an activation function ϕ (for example, a sigmoid function) to the values of Z (matrix of net values), which results from the linear combination of the input variables with the weights W; these values are the inputs of the output layer. The activation of the hidden layer is done with the following expressions:

$$Z^{(h)} = A^{(in)}W^{(h)}, A^{(h)} = \phi\left(Z^{(h)}\right)$$

As $A^{(in)}$ is an array of features or samples $x^{(in)}$; $W^{(h)}$ is the weight matrix, and $\phi(\cdot)$ is the activation function. Similarly, the activation of the output layer is generated:

$$Z^{(out)} = A^{(h)}W^{(out)}, A^{(out)} = \phi\left(Z^{(out)}\right)$$

As $W^{(out)}$ is an array of output weights; and $A^{(out)}$ is a probability matrix with the predicted responses or classes of the network. To determine the classification error, the target class is compared with the predicted class. The back propagation algorithm is used to distribute the errors and partial derivatives are obtained with respect to the network weights to update the model. (Raschka y Mirjalili, 2017).

Gradient Boosting Classifier

The GB classifier consists of a set of individual decision trees that are trained sequentially, in such a way that each tree improves the errors of the previous trees. To predict a new observation, the predictions of all the individual trees in the model are added. GB can use any loss function as long as it is differentiable. A model is fitted, for example, f_1 to predict the response variable, then the errors are calculated $y - f_1(x)$; then, a model f_2 is fitted that tries to predict the errors of the previous model; again a model f_3 is fitted that tries to correct the errors of the previous models and this is repeated m times. To avoid overfitting the model, a regularization parameter is used, which is called the learning rate (λ), that limits the influence of each model in the assembly set.

$$y \approx \lambda \sum_{i=1}^{m} f_i(x)$$

The idea behind boosting is to sequentially tune multiple simple models, where each model uses information from the previous model to "learn from its mistakes" and improve with each iteration; The average of the predictions is taken as the final value. (Rogers y Gunn, 2005).

CC BY

Random forest

RF is an ensemble of decision trees and its goal is to average multiple decision trees to build a more robust model that has better generalization and is less susceptible to overfitting (Raschka and Mirjalili, 2017). To predict the class, the rules of each tree are used and they are assigned by majority vote (Breiman, 2001). The RF algorithm is summarized as follows:

A sample of size n is selected from the set of predictor variables (by random sampling without replacement, bootstrap), the tree grows from an initial sample; for each node d features are selected without replacement; the node is divided with the function that provides the best division according to the information gain (IG) objective function, which is defined by:

$$IG = \left(D_p, f\right) = I\left(D_p\right) - \frac{N_l}{N_p} I(D_l) - \frac{N_r}{N_p} I(D_r)$$

where f is the characteristic to perform the division; N_p is the total number of samples in the parent node; N_l is the number of samples in the left child node; N_r is the number of samples in the right child node; I is the measure of impurity (gini, entropy or classification error); D_p is the dataset in the parent node; D_l is the dataset at the left child node, and D_r is the dataset in the right child node.

RF is also used to determine the importance of a set of variables in the model. The RF algorithm creates classifiers with a random selection of features; this achieves a good exploration of subsets of these, where those variables with greater importance are selected (Rogers y Gunn, 2005).

Performance criteria of prediction models

To evaluate the performance of the MLP, GB and RF classification models, the metrics are obtained from a confusion matrix (MC) which describes the count of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The rows represent the number of samples in the observed class and the columns the number of predictions in each class. The MC diagonal corresponds to the number of samples that the algorithm correctly classifies in each class. If MC only has positive values on the diagonal, it indicates that the classifier correctly classifies all samples. The overall classification

CC BY

accuracy (GP) metric measures the overall proportion of well-classified samples in each class and is calculated as:

$$PG = \frac{VP + VN}{FP + FN + VP + VN}$$

The metrics to measure the performance of the classifier in each class are precision (P), sensitivity (S), specificity (E), and F1 score. They are defined with the following expressions:

$$P = \frac{VP}{VP + FP}$$

$$S = \frac{VP}{FN + VP}$$

$$E = \frac{VN}{VN + FP}$$

$$F1 = 2\frac{P \times S}{P + S}$$

In this case, the value of F1 summarizes P and S in a single metric, is an appropriate estimator in unbalanced classes, and varies between zero and one. The receiver operating characteristics (ROC) curve is a curve that relates values of S versus 1-E. The different points on the curve correspond to the cut-off points used to determine if the test results are positive. The AUCROC value (area under the ROC curve) is interpreted as the probability that in two samples, one positive and one negative, the test assigns a higher probability to the positive sample, correct classification (Mandrekar, 2010). Its value ranges between zero and one; the higher the AUCROC, the better the classification, a value close to 0.50 indicates a poor classification. The P-S curve is the result of plotting P versus S. This allows us to observe from which S there is a degradation of P and vice versa. The ideal result is a curve that approaches the upper right corner (high P and S), which generates an area under the AUCP-S curve, the closer to one, the better the model. (Saito y Rehmsmeier, 2015).

Model training and validation

To train the models, each data set was divided into two random partitions, 80% for training and 20% for testing. For each dataset (ESP2008, MAT2008, ESP2011 and MAT2011) the RF classifier was applied to assess the relative importance of the input

CC BY

features. In the model test, 20% of the data was considered and the case of unbalanced classes at the extremes is presented (table 4).

Tabla 4. Número de observaciones de los conjuntos de prueba por clase objetivo o nivel de logro académico, en español y matemáticas 2008 y 2011

Clase	ESP2008	ESP2011	MAT2008	MAT2011
0†	255	839	309	1251
1¶	1190	1073	1208	722
2§	763	296	691	235

†0: insuficiente; ¶1: elemental; §2: bueno o excelente.

Fuente: Elaboración propia, con base en datos de SEP (2008)

Selection of optimal hyperparameters

In machine learning models, optimal hyperparameters are tuned in training for best performance. The selection of the optimal hyperparameters consists of finding the combination of values of the hyperparameters that maximizes the performance of the classifier based on a metric, in this study PG was used. The selection of hyperparameters for each classifier was performed through cross validation (CV). Training was performed with a random sample of 80% of the total data set. The CV method consists of randomly subdividing the training set into k disjoint subsets of the same size. Then, for each combination of hyperparameter values (table 5), the model is executed k times. In each iteration k, one of the disjoint subsets is used as a validation set and the rest as a training set (80% training and 20% validation) and a value of the PG performance metric is obtained. After evaluating different combinations of hyperparameter values, the combination of hyperparameter values that maximizes the average PG obtained from VC is selected. (k = 5).

Tabla 5. Valores para la búsqueda y selección de hiperparámetros de los clasificadores perceptrón multicapa (MLP) y potenciación del gradiente (GB)

Modelo	Hiperparámetros	Valores considerados
MLP	*nco*	50, 100, 200, 250
	fa	*tanh, relu, logistic*
	op	*sgd, adam, lbfgs*
	re	0.0001, 0.001, 0.1
	ta	*constant, adaptative*
	mi	200
GB	*mi*	90, 100, 110
	pa	4, 5, 6
	ha	3, 4

nco: neuronas en capa oculta; *fa:* función de activación; *op:* optimizador de pesos; *re:* regularizador; *ta:* tasa de aprendizaje; *mi:* máximo de iteraciones; *pa:* profundidad de árbol; *ha:* hojas por árbol.

Fuente: Elaboración propia

After the selection stage of the optimal hyperparameters of the MLP and GB classifiers, the final evaluation of their performance was carried out. To determine the PG of each model, the complete data set of each input scenario (Table 3), the optimal combination of hyperparameters selected and the VC procedure (k = 5) were considered. In each iteration, a PG value of the model is obtained; At the end of the k iterations, the average of the PG values and their standard deviation, and the other metrics proposed in the study, were calculated.

Relative importance of input characteristics

RF was used to determine the relative importance of the input features or variables in predicting the target class. RF builds a large number of classifiers based on randomly selected subsets of variables. At each RF node an input variable is selected that is used to partition the node and maximize the information gain (performance metric). Variable importance measures are used to determine the performance of the machine learning model (Rogers and Gunn, 2005).

CC BY

To calculate the relative importance of the 13 predictor variables with the RF model, the VC procedure (k = 5) was applied to select the optimal hyperparameters. Subsequently, the feature importance option was applied to the RF model using the Python Scikit-learn library. This stage was carried out for the four data sets: ESP2008, ESP2011, MAT2008 and MAT2011.

Results

Optimal hyperparameters

The optimal hyperparameters selected with cross validation and grid search of the MLP and GB classifiers for each dataset are illustrated in Table 6. The optimal hyperparameters of each classifier, in general, depend on the analyzed dataset.

Tabla 6. Hiperparámetros óptimos de los clasificadores perceptrón multicapa (MLP) y potenciación del gradiente (GB) para cada escenario de entrada de datos

Modelo	Hiperparámetros	ESP2008	ESP2011	MAT2008	MAT2011
MLP	*Nco*	100	200	50	30
	fa	*logistic*	*tanh*	*relu*	*logistic*
	op	*adam*	*sgd*	*sgd*	*adam*
	re	0.1	0.1	0.1	0.0001
	ta	*constant*	*adaptative*	*constant*	*constant*
GB	*mi*	90	90	90	110
	pa	4	4	4	4
	ha	3	4	3	4

nco: neuronas en capa oculta; *fa:* función de activación; *op:* optimizador de pesos; *re:* regularizador; *ta:* tasa de aprendizaje; mi: máximo de iteraciones; *pa:* profundidad de árbol; *ha:* hojas por árbol.

Fuente: Elaboración propia

CC BY

Classifier performance

The average performance and its standard deviation in prediction, of the MLP and GB models for the four scenarios analyzed in the Enlace test, in terms of global classification accuracy, MLP was superior to GB with the ESP2008 scenario and GB was superior to MLP with ESP2008. the MAT2008 scenario (table 7).

Tabla 7. Precisión global (*PG*) en predicción promedio (+/- desviación estándar) de los clasificadores perceptrón multicapa (MLP) y potenciación del gradiente (GB) en español y matemáticas

Modelo	ESP2008	ESP2011	MAT2008	MAT2011
MLP	0.701(+/- 0.033)	0.611(+/- 0.019)	0.672(+/- 0.034)	0.631(+/- 0.008)
GB	0.695(+/- 0.024)	0.610(+/- 0.016)	0.688(+/- 0.008)	0.635(+/- 0.006)

Fuente: Elaboración propia

The MLP classifier obtained better overall performance than GB to predict the degree of academic achievement of students in Spanish in 2008 and 2011, with a PG of 70.1% and 61.1%, respectively; however, the performance of both classifiers is very similar in both subjects in 2008 and 2011 (table 7).

In the subject of Spanish, MLP and GB obtained better performance in 2008 to classify classes 1 and 2 than class 0 (F1 of 77.0% and 75.0% versus 33.0%, respectively). On the other hand, in 2011, MLP and GB obtained better performance to classify classes 0 and 1, than class 2. This is reinforced when observing the results of $AUC_{P\text{-}S}$ with values of 0.70 and 0.60 for classes 0 and 1 (table 8, figure 3).

CC BY

Tabla 8. Desempeño en predicción de los clasificadores perceptrón multicapa (MLP) y potenciación del gradiente (GB) por clase de logro académico en español 2008 y 2011

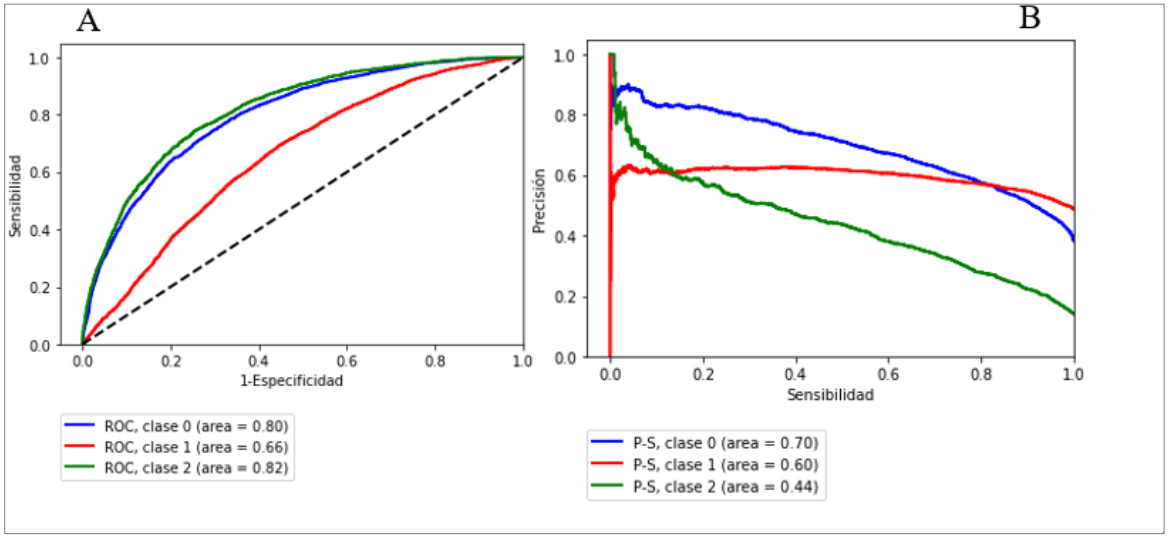
Modelo		ESP2008			ESP2011		
	Clase	*F1*	AUC_{ROC}	$AUC_{P\text{-}S}$	*F1*	AUC_{ROC}	$AUC_{P\text{-}S}$
MLP	0	0.32	0.86	0.46	0.65	0.80	0.70
	1	0.77	0.74	0.72	0.67	0.66	0.60
	2	0.75	0.87	0.75	0.41	0.82	0.44
GB	0	0.33	0.86	0.40	0.65	0.80	0.69
	1	0.77	0.71	0.68	0.67	0.64	0.57
	2	0.75	0.87	0.77	0.42	0.81	0.41

F1: f1-score, AUC_{ROC}: área bajo la curva ROC, $AUC_{P\text{-}S}$: área bajo la curva *P-S*, clases 0: insuficiente, 1: elemental, 2: bueno o excelente

Fuente: Elaboración propia

MLP and GB obtained low performance to classify class 2 with F1 = 0.41. This is confirmed by observing the P-S curves with a low value of AUCP-S for class 2 (figure 3).

Figura 3. Curvas de desempeño en predicción del clasificador perceptrón multicapa (MLP) en español 2011 por clase (0: insuficiente, 1: elemental, 2: bueno o excelente) A: curvas ROC, B: curvas *P-S*



Fuente: Elaboración propia

For the 2008 math dataset, GB outperformed MLP, with an average PG of 68.8% (table 3). Classes 1 and 2 (F1 = 77.0% and 74.0%) are better classified than class 0 (F1 =

43%). MLP obtained AUCP-S of 0.68 and 0.73 for classes 1 and 2, respectively, and 0.50 for class 0. GB obtained a performance very close to MLP.

Using the 2011 math dataset, GB was slightly better than MLP (GB had an average PG of 63.5%). GB outperformed MLP to classify class 0 with F1 of 77%. Both models had low performance to classify classes 1 and 2. From the results of AUCP-S, it is verified that both models had a good performance to classify class 0, and low for classes 1 and 2 (0.37) (table 9, figure 4).

Tabla 9. Desempeño en predicción de los clasificadores perceptrón multicapa (MLP) y potenciación del gradiente (GB) por clase de logro académico en matemáticas 2008 y 2011

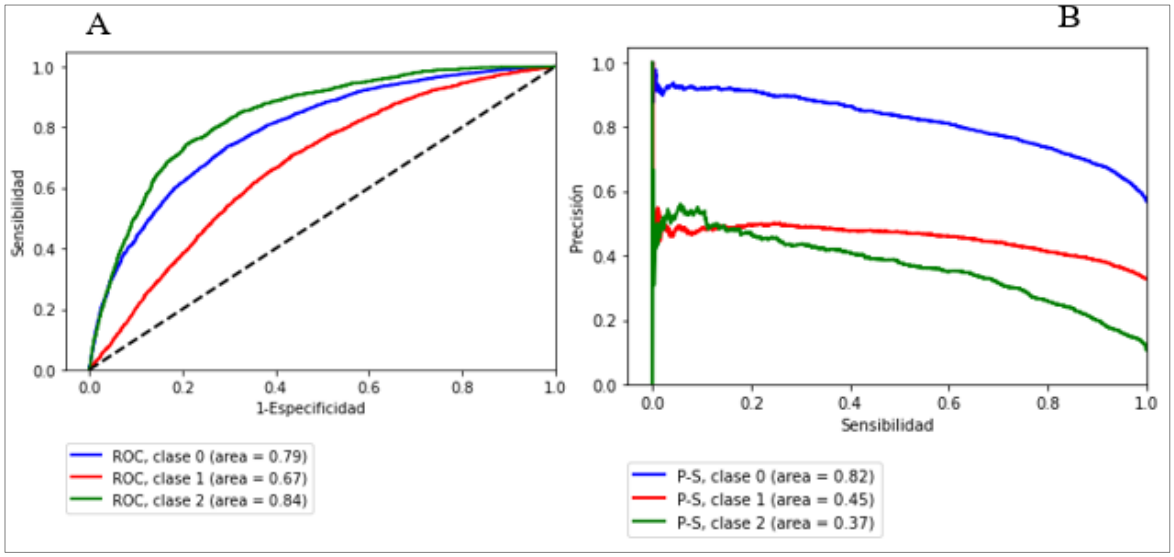
Modelo		MAT2008			MAT2011		
	Clase	*F1*	AUC_{ROC}	$AUC_{P\text{-}S}$	*F1*	AUC_{ROC}	$AUC_{P\text{-}S}$
MLP	0	0.43	0.86	0.50	0.76	0.79	0.82
	1	0.77	0.70	0.68	0.44	0.68	0.47
	2	0.74	0.86	0.73	0.30	0.83	0.38
GB	0	0.39	0.86	0.48	0.77	0.79	0.82
	1	0.78	0.70	0.69	0.46	0.67	0.45
	2	0.73	0.87	0.76	0.39	0.84	0.35

P: precisión; *F1*: f1-score, AUC_{ROC}: área bajo la curva ROC, $AUC_{P\text{-}S}$: área bajo la curva *P-S*, clases 0: insuficiente, 1: elemental, 2: bueno o excelente

Fuente: Elaboración propia

CC BY

Figura 4. Curvas de desempeño en predicción del clasificador potenciación del gradiente (GB) en matemáticas 2011 por clase (0: insuficiente, 1: elemental, 2: bueno o excelente), A: curvas ROC, B: curvas *P-S*

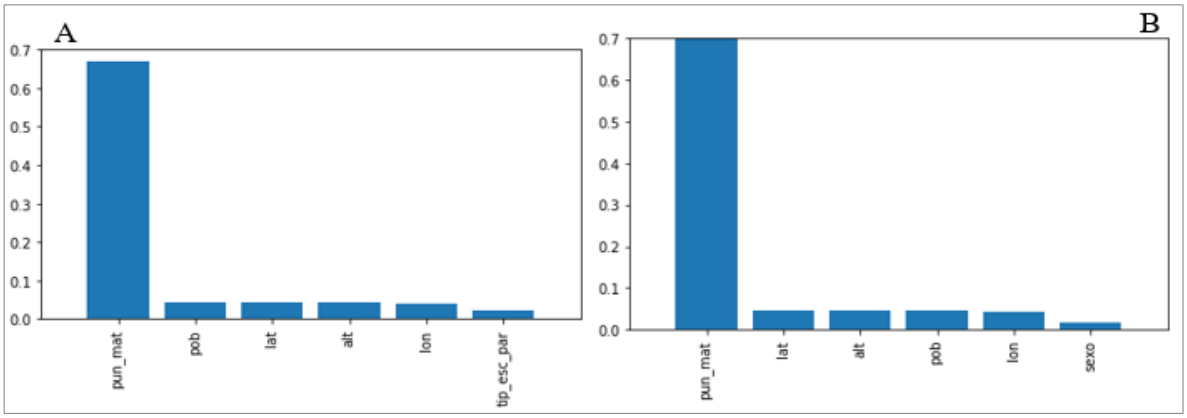


Fuente: Elaboración propia

Relative importance of features

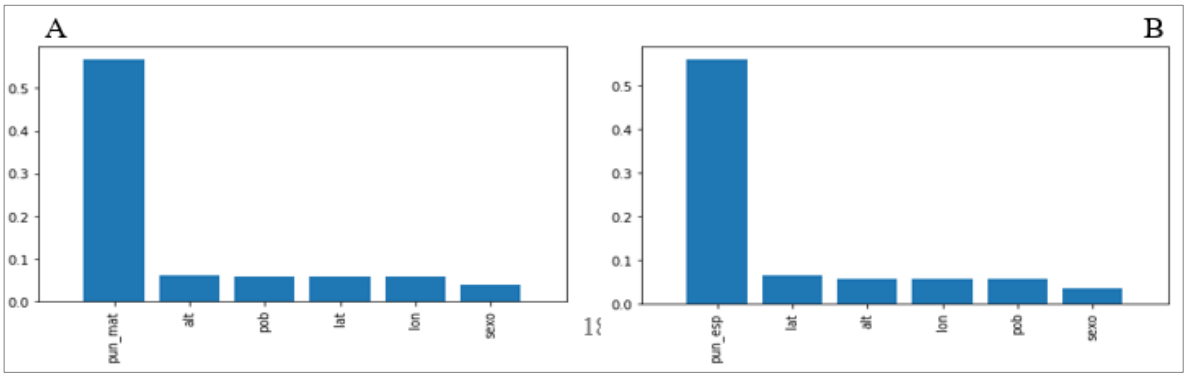
Among the results of relative importance of the input variables, determined with the RF classifier for the four data sets (Spanish and mathematics, 2008 and 2011), it was found that the score variables pun_esp and pun_mat are the most important to predict the academic achievement (figures 5 and 6). That is, there is a strong association (0.65) between the score obtained by a student in mathematics and his degree of academic achievement in Spanish (see sections a) of figures 5 and 6). Similarly, there is a strong association (0.7) between the score obtained by a student in Spanish and the degree of academic achievement in mathematics (see items b) of figures 5 and 6). This is preserved at the beginning (2008) and at the end of the analyzed period (2011).

Figura 5. Importancia relativa de variables de entrada obtenida con el clasificador bosque aleatorio (RF) A: español 2008, B: matemáticas 2008



Fuente: Elaboración propia

Figura 6. Importancia relativa de variables de entrada, obtenida con el clasificador bosque aleatorio (RF) A: español 2011, B: matemáticas 2011



Fuente: Elaboración propia

In addition to the variables corresponding to test scores, the variables of sex, particular t_esc, pop, lat, lon, and alt are important for RF models. That is, they are variables that influence the classification of students in some of the levels of achievement analyzed.

Discussion

The results obtained show the association of the 13 input variables with the three classes or levels of academic achievement (insufficient, elementary, and good or excellent) obtained by the students of the state of Tlaxcala. One of the most important variables was the score obtained in mathematics to predict the degree or level of academic achievement in Spanish and the score obtained in Spanish to predict the degree of academic achievement in

mathematics. In order of importance, although to a lesser extent, the geographical location of the school, the population of the locality where the school is located and the sex of the student are also listed. The observed influence of the pun_esp and pun_mat variables on the results obtained for the classification of students in achievement levels is highlighted.

Fernández (2003) points out that the learning of the Spanish language results from a process of accumulation of pedagogical experiences that the student has during his stay at school and the learning of mathematics as a constructive process that is related to the formulation or understanding of concepts with problem solving. It is important to observe the grades of the students in Spanish in order to predict the classification of achievement levels in mathematics and vice versa, that is, it is convenient to know what happened in the other subject as a general measure of the students' ability.

With the different metrics selected to compare the two models, the area under the AUCP-S curve provided more information to assess the performance of the classifiers in discriminating the percentage of correctly classified samples, and the results are similarly reflected in the confusion matrix.

The two machine learning algorithms (MLP and GB) obtained an overall performance of correct classification (PG) greater than 60.0%). A limitation of the work to increase the performance of the classifiers was the presence of unbalanced target classes, the classifier tends to give greater importance to the majority classes. To improve the work, additional context variables can be considered and see if they improve the classification. Alvarez et al. (2007) used variables associated with students referring to socioeconomic indicators, characteristics of the school and institutional aspects (state pedagogy, union influence, etc.) to determine which factors influence school performance in mathematics, science and reading of PISA, likewise, Hussain and Qasim (2021) used historical grade data to predict student grades using machine learning algorithms.

Conclusions

Machine learning multilayer perceptron (MLP) and gradient boosting (GB) classifiers obtained comparable performances in terms of overall classification accuracy (GP) in predicting levels of academic achievement (0: insufficient, 1: elementary, and 2: good or excellent) of elementary and middle school students in the state of Tlaxcala, based on

CC BY

contextual variables extracted from the Enlace test (National Assessment of Academic Achievement in Schools). In math, GB had a PG of 68.8% in 2008, and 63.5% in 2011; likewise, in 2008, MLP and GB performed better in classifying classes 1 and 2 than class 0 (insufficient). In contrast, in 2011, in both subjects, MLP and GB performed better at classifying classes 0 and 1 than class 2.

The contextual variables used in this study showed an association with the levels of academic achievement; in particular, the variables intern, sex and school shift. The score in Spanish obtained by a student influences the level of academic achievement in mathematics and vice versa. These results show the importance of machine learning algorithms to identify relevant factors that affect the school performance of students from the analysis of massive data of existing school information in the Ministry of Public Education.

The GB and MLP classifiers represent an alternative approach to identify the variables or contextual factors that favor or limit the academic achievement of students and constitute a decision-making support tool to identify low-performing students and propose focused solutions to structural problems. such as school dropout.

Future lines of research

To complement the work, it is important to consider other contextual variables such as the education of the parents, characteristics of the family, characteristics of the school, etc., and that can be obtained by crossing information on school performance from the Link test with the questionnaires of context that rose alongside the test.

These new variables can be analyzed with machine learning models to assess their influence on school performance. In addition, it is possible to test other classification algorithms such as support vector machines (SVM) or k-nearest neighbor (KNN). Likewise, with the approach applied in this study, it is possible to select school information from other entities or regions of the country to evaluate school performance in other socioeconomic contexts.

Acknowledgment

To the National Council of Science and Technology (Conacyt) for the financial support provided to carry out this research.

CC BY

References

Altabrawee, H., Osama, A. J. and Qaisir, A. S. (2019). Predicting Students' Performance Using Machine Learning Techniques. *Journal of University of Babylon for Pure and Applied Sciences*, *27*(1), 194-205. Retrieved from https://doi.org/10.29196/jubpas.v27i1.2108.

Álvarez, J., García, M. V. and Patrinos, H. A. (2007). *Institutional Effects as Determinants of Learning Outcomes. Exploring State Variations in Mexico*. Washington, United States: The World Bank.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. Retrieved from https://doi.org/10.1023/A:1010933404324.

Borkar, S. and Rajeswari, K. (2014). Attributes Selection for Predicting Students' Academic Performance Using Education Data Mining and Artificial Neural Network. *International Journal of Computer Applications*, *86*(10), 25-29. Retrieved from https://doi.org/10.5120/15022-3310.

Đambić, G., Krajcar, M. and Bele, D. (2016). Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. *International Journal of Digital Technology & Economy*, *1*(1), 1-11.

Fernández, T. (2003). *Determinantes sociales, organizacionales, e institucionales de los aprendizajes en la educación primaria en México: un análisis de tres niveles (2001)*. México: Instituto Nacional para la Evaluación de la Educación. Recuperado de https://www.inee.edu.mx/wp-content/uploads/2019/01/P1C126.pdf.

González, J. M., de los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., Babu, R. and Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *125*(4), 759-771. Retrieved from https://doi.org/10.1007/s00122-012-1868-9.

Hussain, S., and Qasim, K. M. (2021). Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. *Annals of Data Science.* Retrieved from https://doi.org/10.1007/s40745-021-00341-0.

Instituto Nacional de Estadística y Geografía [Inegi]. (2020). Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades. Recuperado de https://www.inegi.org.mx/app/ageeml/#.

Instituto Nacional para la Evaluación de la Educación [INEE]. (2019). Evaluaciones. Bases de datos. Recuperado de https://www.inee.edu.mx/evaluaciones/bases-de-datos/.

Mandrekar, J. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology, 5*(9), 1315-1316. Retrieved from https://doi.org/10.1097/JTO.0b013e3181ec173d.

Martínez, F. (coord.) (2015). *Las pruebas Enlace y Excale. Un estudio de validación*. Ciudad de México, México: Instituto Nacional para la Evaluación de la Educación.

Organisation for Economic Co-operation and Development [OCDE]. (2005). School Factors Related to Quality and Equity: Results from PISA 2000. Paris, France: Organisation for Economic Co-operation and Development. Retrieved from https://read.oecd-ilibrary.org/education/school-factors-related-to-quality-and-equity_9789264008199-en#page1.

Rai, S., Shastry, K., Pratap, S., Kishore, S., Mishara, P. and Sanjaty, H. (2020). Machine Learning Approach for Student Academic Performance Prediction. In Bhateja, V., Grobelnik, M., Peng, S., Sataphaty, S. and Zhang, Y. (eds.), *Evolution in Computational Intelligence. Frontiers in Intelligent Computing: Theory and Applications* (pp. 611-618). Singapore: Springer.

Raschka, S. and Mirjalili, V. (2017). *Python Machine Learning. Machine Learning and Deep Learning with Python-scikit-learn, and TensorFlow*. Birmingham, England: Packt Publishing.

Rogers, J. and Gunn, S. (2005). Identifying Feature Relevance Using a Random Forest. In Saunders, C., Grobelnik, M., Gunn, S. and Shawe-Taylor, J. (eds.), *Subspace, Latent, Structure and Feature Selection* (pp. 173-184). Bohinj, Slovenia: Springer.

Saito, T. and Rehmsmeier, M. (2015). The Precision-Recall Plot is More Informative than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE, 10*(3). 1-21. Retrieved from https://doi.org/10.1371/journal.pone.0118432.

Secretaría de Educación Pública [SEP]. (2008). *Enlace. Educación básica. Manual técnico 2008*. México: Secretaría de Educación Pública. Recuperado de http://enlace.sep.gob.mx/ba/manuales_tecnicos/.SEP-IEIA.

CC BY

Rol de Contribución	Autor (es)
Conceptualización	Juan Manuel González-Camacho (principal), Miguel Angel Morales Hernández (que apoya)
Metodología	Miguel Ángel Morales (igual) -Hernández, Juan Manuel González-Camacho (igual)
Software	Miguel Ángel Morales -Hernández (principal), Juan Manuel González-Camacho (que apoya)
Validación	Miguel Ángel Morales -Hernández (principal), Juan Manuel González-Camacho (que apoya)
Análisis Formal	Miguel Ángel Morales -Hernández (igual), Juan Manuel González-Camacho (igual)
Investigación	Miguel Ángel Morales -Hernández (principal), Juan Manuel González Camacho (que apoya)
Recursos	Miguel Ángel Morales -Hernández (igual), Juan Manuel González-Camacho (igual)
Curación de datos	Miguel Ángel Morales Hernández
Escritura - Preparación del borrador original	Miguel Ángel Morales -Hernández (igual), Juan Manuel González-Camacho (igual)
Escritura - Revisión y edición	Miguel Ángel Morales -Hernández (igual), Juan Manuel González-Camacho (igual), David H. Del Valle Paniagua (que apoya), Héctor Robles (que apoya), José Rafael Durán Moreno (que apoya)
Visualización	Miguel Angel Morales Hernández
Supervisión	Juan Manuel González-Camacho
Administración de Proyectos	Juan Manuel González-Camacho
Adquisición de fondos	Miguel Ángel Morales -Hernández (igual), Juan Manuel González-Camacho (igual)

CC BY