International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Determining Academic Performance in Mining Engineering Students Incorporating Socioeconomic Factors through a Supervised Neural Network

Marco Cotrina*[1], Jairo Marquina[2], Eduardo Noriega[3], Jose Mamani[4], Eusebio Antonio[5], Solio Arango[6], Hans Portilla[7]

Submitted: 29/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: In recent times, artificial intelligence (AI) has become an indispensable pillar in the educational sector. One of its key applications is the prediction of students' academic performance based on personal variables such as their socioeconomic context, residence address, among others. This study introduces and develops a model based on a supervised artificial neural network designed to analyze academic performance considering socioeconomic factors. To calibrate the model, information was collected from 40 mining engineering students in the VIII and X cycles at the National University of Trujillo, Huamachuco Campus through virtual surveys, evaluating aspects such as housing type, living conditions, and food consumption patterns (including red meat, fish, fruits, and vegetables). The neural network architecture consisted of an input layer with 6 neurons, four hidden layers composed of 10, 8, 5, and 3 neurons respectively with a ReLU activation function, and an output layer with a single neuron with a sigmoid activation function. The neural network achieved an accuracy of 75.0%, and when comparing these results with other models such as Random Forest with an accuracy of 50.0% and SVM with an accuracy of 62.5%, the neural network obtained the highest accuracy compared to the other models using the same data.

Keywords: *Socioeconomic factors, supervised neural networks, academic performance, Random Forest, SVM*

1. Introduction

Artificial Intelligence (AI) has redefined various industries, and the education sector is no exception. The incorporation of advanced technologies, such as deep learning [1], automation [2] and natural language processing, brings benefits that enrich the educational field. These tools not only facilitate a more detailed understanding of students' learning patterns, but also enhance the improvement of their academic outcomes [3].

Low academic performance represents one of the most significant challenges in higher education globally. This problem acquires complexity due to the multiplicity of factors that influence student performance [1]. According to Tejedor and García-Valcárcel [4], performance is affected by psychological, academic, pedagogical, and socio-familial elements. The repercussions of low academic performance are diverse, with university dropout standing out as one of the most serious. Academic performance also functions as a barometer of institutional quality, making its monitoring an essential aspect for educational entities. In this sense, SINEACE [5] in Peru points out that continuous evaluation and early intervention in the face of academic deficiencies are fundamental pillars in the quality criteria for university programs, essential for their accreditation. The recent COVID-19 health crisis prompted the transition to virtual educational modalities, introducing additional challenges in the adaptation process for both students and teachers. In this emerging hybrid learning environment, particular obstacles have presented themselves. Liao and Wu [6] argue that while hybrid learning has opened doors to new professional development opportunities for students, it has also introduced challenges as students now find themselves more susceptible to varied distractions during their training.

In their study, Hellas [7] explored the necessary features for making predictions, identified algorithms that could enhance these predictions, and quantified aspects of student academic performance. They posed the question:

[1] Associate Professor, Department of Mining Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0003-3801-0370
[2] Assistant Professor, Department of Mining Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0002-5880-8227
[3] Associate Professor , Department of Mining Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0001-7674-7125
[4] Associate Professor , Department of Chemistry, Universidad Nacional del Altiplano de Puno, PUNO.
ORCID ID : 0000-0001-7803-7936
[5] Associate Professor , Department of Mining Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0001-5072-5411
[6] Associate Professor , Department of Mining Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0003-3594-0329
[7] Associate Professor , Department of Metallurgical Engineering, National University of Trujillo, PERU
ORCID ID : 0000-0001-6014-9243
** Corresponding Author Email: mcotrinat@unitru.edu.pe*

What is the current state of the art in predicting student performance? Their findings indicated that the features used to predict student academic performance can be categorized into five groups: demographic (age, gender), personality (self-efficacy, self-regulation), academic (high school performance, course performance), behavioral (registration data), and institutional (quality of high school, teaching methods). The methodologies employed can be divided into classification (supervised learning, e.g., Naive Bayes, decision trees), clustering (unsupervised learning, e.g., data partitioning), statistical (e.g., correlation, regression), and other methods. It was observed that regression models (linear) and classification methods are the most used tools, with the former typically serving as a prediction method, while the latter are often compared with classification algorithms, resulting in multiple prediction outcomes.

In their article, Namoun & Alshanqiti [8] analyzed a decade of research work from 2010 to November 2020 to present a fundamental understanding of the intelligent techniques used for predicting student performance. The methodology involved searching electronic bibliographic databases, including ACM, IEEE Xplore, Google Scholar, Science Direct, Scopus, Springer, and Web of Science, analyzing a total of 62 relevant articles. The focus was on three perspectives: 1) the predictive analysis models developed for forecasting student learning outcomes, and 2) the dominant factors impacting student results. They applied PICO and PRISMA practices to synthesize and report the main findings. The results indicated that the intelligent models suggested for predicting learning outcomes were mainly statistical analysis (45.16%), supervised machine learning (40.32%), data processing (8.06%), both supervised and unsupervised learning (4.83%), and unsupervised machine learning (1.61%). Predictive algorithms consisted of correlation and regression models (51.61%), neural networks (14.51%), decision trees (14.51%), Bayesian-based models (8.06%), support vector machines (3.22%), instance-based models (1.62%), and other models (6.45%). The top five high-performance prediction models were hybrid random forests (99.25-99.98%), 3-L feedforward neural networks (98.81%), random forests (98%), naive Bayes (96.87%), and artificial neural networks (95.16-97.30%). The five lowest-performing prediction models were linear regression (50%), bagging (48-55%), mixed-effects logistic regression (69%), discriminant function analysis (64-73%), and logistic regression (76.2%). Dominant factors affecting student learning outcomes included access time to educational resources, site engagement, time and number of online sessions, evaluation during the semester, assignments, exam scores, and exam grades. Another significant factor was students' interest and enthusiasm, intrinsic motivations, and the teacher-student relationship.

In their research, Helal [9] developed various classification models to predict student performance. The data included student enrollment details as well as activity data generated by the Learning Management System (LMS). Enrollment data contained information about students, such as sociodemographic characteristics, university admission basis, and type of attendance. The findings indicated that students with lower participation in quiz activities or less frequent viewing of book or file resources mostly do not succeed. Furthermore, students with poor academic backgrounds (ATAR, admission basis), belonging to a lower social status, or studying part-time often have limited time and, as such, fail to reach their academic potential.

In their study, Muhammad [10] analyzed significant socioeconomic factors affecting a student's performance in Khyber Pakhtunkhwa, Pakistan. The methodology involved a dataset collected from 100 different schools in Pakistan, comprising over 5550 students who were surveyed using an appropriate questionnaire. To select the most prominent features from the dataset, two different feature selectors (FCBF and Relief) were used, and their performance was measured alongside ML models. The accuracy of the utilized classifier models was as follows: decision tree at 73.10%, multilayer perceptron at 74.11%, KNN at 81.13%, random forest at 79.00%, and ANN (N=3) at 80.00%.

In the realm of academic performance prediction, a variety of approaches have been explored, leveraging both conventional and innovative data sources to enhance prediction accuracy. Benablo [11] forecasted student performance using data from social media platforms such as Facebook, Twitter, Instagram, and YouTube, employing the Support Vector Machine (SVM) model. In a test of 100 instances, this model reported a precision of 100%, recall of 96.8%, and an F1 score of 98.4%, positioning it as the most effective model reported. Similarly, Amazona & Hernandez [12] suggested the utilization of a deep learning neural network model for prediction purposes. The outcomes from this model demonstrated a precision of 98%, an F1 score of 97%, and a recall of 98%. Rodríguez [13] tested a systematic procedure to implement artificial neural networks for predicting academic performance in higher education and analyzed the significance of various well-known predictors of academic performance. The sample encompassed 162,030 students of both genders from public and private universities in Colombia. The findings indicated the feasibility of systematically implementing artificial neural networks to classify student academic performance as high (82% precision) or low (71% precision). Yağcı [14] proposed a new model based on machine learning algorithms to predict undergraduate students' final exam grades, sourcing data from midterm exam grades. The performance of random forests, nearest neighbor, support vector machines, logistic regression, Naive Bayes, and k-nearest neighbor algorithms were calculated and compared for predicting students' final exam grades. The dataset consisted of academic performance grades of 1,854 students who took the Turkish Language I course at a state university in Turkey. The proposed model achieved a classification accuracy of 70% to 75%, with predictions made using only three types of parameters: midterm exam grades, department data, and faculty data.

In aligning with the investigation currently being conducted, the three-pillar framework proposed by Cuccurullo [15] offers a comprehensive analysis of the social, intellectual, and conceptual structures within the field. The thematic mapping reveals central themes including “academic performance,” “education,” “adolescence,” “socioeconomic factors,” and “socioeconomic status,” with “student” and “performance” serving as pivotal themes. Through content analysis, we delineated pivotal trends in the literature concerning socioeconomic factors and academic achievement in students. The integration of bibliometric and content analyses proves indispensable for pinpointing literature gaps and formulating prospective research inquiries [16]. Hence, we advocate for broadening the research scope to encompass additional determinants like health, economic, social, and educational factors influencing students' academic performance.

Several researchers have explored algorithmic techniques to predict student academic performance. Gil and Quintero [17] used a multilayer perceptron-type artificial neural network, achieving an accuracy of 73%. Vargas [18], on the other hand, applied supervised learning algorithms on first semester students of Systems and Computer Engineering, registering an accuracy of 81.97% using the K-NN model. Cabana [19] also employed neural networks, obtaining an error of 5% during training and a mean square error of 6.2% in validation. Saire [20] used classification algorithms, with an accuracy ranging from 87% to 93%. Zevallos [21] used a multilayer perceptron with resilient backpropagation learning algorithm and hyperbolic activation function, achieving 84% accuracy. For their part, Blanco [22] opted for deep neural networks, achieving 78% accuracy. Rincón-Flores [23] diversified their methods, employing both K-nearest neighbors (KNN) and Random Forest, and recording 80% accuracy.

The motivation for this investigation is rooted in the pressing need to understand and address the multifaceted influences on academic performance within the field of mining engineering education. Despite the wealth of research on academic performance predictors, a significant gap exists in comprehensively incorporating socioeconomic factors into predictive models. Traditional approaches often emphasize academic and psychological factors, overlooking the nuanced ways in which a student's socioeconomic background can impact their educational outcomes. This oversight presents a critical limitation, as socioeconomic status is increasingly recognized as a determinant of educational access, engagement, and success. Furthermore, the advent of artificial intelligence (AI) in educational settings offers unprecedented opportunities to explore complex datasets and uncover insights that were previously inaccessible. However, the application of AI in predicting academic performance has predominantly focused on academic and behavioral data, with less attention given to integrating and analyzing socioeconomic variables. This research gap underscores a missed opportunity to leverage AI for more holistic and equitable educational assessments. This investigation aims to bridge this gap by developing a supervised artificial neural network model that explicitly incorporates socioeconomic factors to predict academic performance among mining engineering students. By doing so, it seeks to provide a more nuanced understanding of how socioeconomic contexts influence academic success, offering a basis for targeted interventions and support mechanisms for students at risk due to socioeconomic constraints. This work focuses on the formulation and development of a predictive model for the academic performance of mining engineering students at the National University of Trujillo, Huamachuco campus. Using socioeconomic factors and based on deep neural networks, we seek to address the persistent challenge of low academic performance. These predictions will allow institutions to proactively identify at-risk students and take supportive measures, ensuring the improvement of their academic performance. This, indirectly, strengthens institutional reputation.

2. Material and Methods

In this section, we detail the architecture of the neural network designed to determine the academic performance of mining engineering students at the National University of Trujillo, Huamachuco.

2.1. Architecture

Fig. 1 illustrates the architecture used in the proposed model.

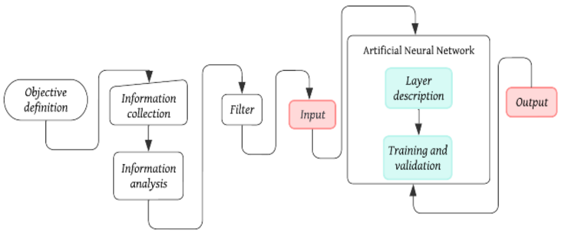

Fig. 1. Architecture used in the proposed model.

2.1.1. Definition of the objective

The main objective of this research was to determine the academic performance of Mining Engineering students of cycles VIII and X of the National University of Trujillo, Huamachuco, using family socioeconomic variables. This projection sought to provide tools to teachers and institutions to identify those students with a higher risk of academic failure and, consequently, implement appropriate support interventions.

2.1.2. Collection of information

To carry out this research, a set of data obtained through virtual surveys formulated in Google Forms, specifically directed to Mining Engineering students, was used. The collected data were later consolidated in a .csv file format.

The obtained data set contains information on 40 students, 6 socioeconomic factors (house material, house condition, consumption of red meat, consumption of fish, consumption of fruits and vegetables) and academic performance indicators (average grade of the last semester and learning level).

2.1.3. Information analysis

After collecting the information, it is essential to subject it to a thorough analysis. This evaluation allows us to segment and manage the data more effectively. The data set obtained is detailed below, where each item represents a column within the database:

- Enrollment code: reflects the enrollment code assigned by the university.

- Department of birth: documents the student's place of origin or birth.

- Age: records the corresponding age of each student.

- Sex: indicates the gender of the students, being "M" for male and "F" for female.

- Main housing material captures information about the main building material of the student's home. Options include: brick or cement block, adobe or tapia, wood, stone or ashlar and others.

- Housing condition: reflects the current condition of the dwelling, with options such as: excellent, very good, good, fair, and poor.

- Consumption of red meat: documents the frequency with which the student consumes red meat. The options range from "never" to "always".

- Fish consumption: as with the previous item, it records the regularity of fish consumption by the student.

- Fruit consumption: indicates how often the student consumes fruit.

- Vegetable consumption: measures the regularity with which the student incorporates vegetables in his/her diet.

- Average rating: indicates the student’s actual average rating, with a range from 0 to 20.

- Achievement level: indicates the level of achievement obtained in the last semester, with a rating of 0 (initial level), 1 (achieved level).

The table containing the data comprises a total of twelve columns, of which four contain personal information, specifically: enrollment code, department of birth, age, and gender. This results in only eight attributes of the database being relevant to our analysis. The dependent variable, labeled "average grade and level of achievement", is presented in a specific column of the table, and serves to determine the students' grade point average, thus reflecting their academic performance.

2.1.4. Filters and preprocessing

As illustrated in Fig. 2, the correlation between variables is examined using Spearman's method, while Fig. 3 explores these same correlations using Pearson's method. Upon detecting a significant interaction between vegetable consumption and fish consumption across both proxies, we proceeded to apply the feature_importances algorithm of the sklearn library. This approach was used to identify which of these variables exerts a greater influence on the performance of both conventional machine learning models and neural networks. By using the feature_importances function, it was determined that the consumption of vegetables has a significant influence on the results obtained by the models analyzed.

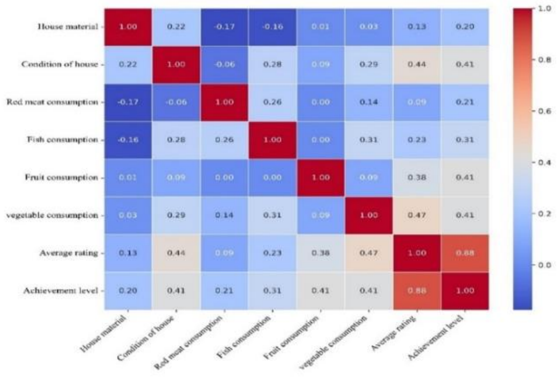

Fig. 2. Spearman’s correlation coefficient.

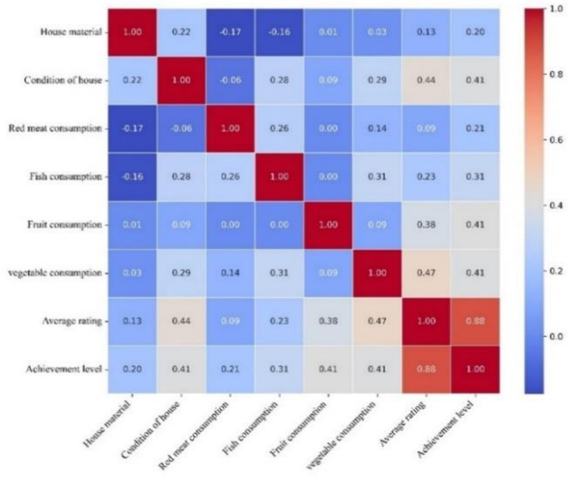

Fig. 3. Pearson's correlation coefficient.

2.1.5. Input

The variables selected to predict the academic performance of mining engineering students of the VIII and X cycle of the National University of Trujillo, Huamachuco campus, are:

- House material: weighted value assigned based on the predominant material of the house.

- Condition of the house: weighted value assigned based on the condition of the house.

- Red meat consumption: weighted value related to the frequency of red meat consumption.

- Fish consumption: weighted value related to the frequency of fish consumption.

- Fruit consumption: weighted value related to the frequency of fruit consumption.

- Vegetable consumption: weighted value based on the frequency of vegetable consumption.

2.1.6. Output

Teachers will use these results to effectively identify those students with a high probability of obtaining an unsatisfactory weighted average (academic performance) using socioeconomic factors and determine their achievement level to verify student placement. For the prediction of academic performance, the following results are available:

Average grade: Reflects the average obtained on a scale of 1 to 20.

Learning level: Reflects the learning level of the students, with the objective of classifying academic performance. Initial level or low academic performance (0-13) and achieved level or high academic performance (14-20).

The proposed artificial intelligence will use the learning level variable to indicate the results.

2.1.7. Artificial neural networks

The architecture of the designed neural network, illustrated in Fig. 4, is articulated in six stratified layers for data processing. The configuration is as follows:

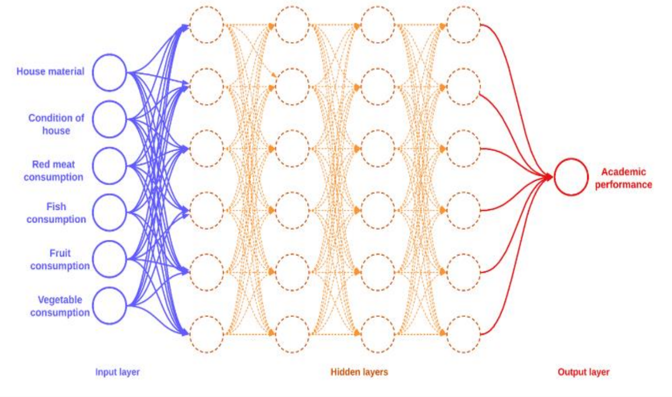

Fig. 4. Neural network architecture chart.

- Input Layer: Composed of six neurons, this layer serves as the initial gateway for the dataset variables. The "ReLU" activation function was chosen due to its computational efficiency, evidenced by the speed of computational execution and gradient derivation compared to other activation functions.

- Hidden Layers: Four intermediate layers are structured whose number of neurons was determined based on a rule of thumb suggested by Heaton [24], which postulates that the number of hidden neurons should be less than twice the size of the input layer. Following this guideline, layers with ten, eight, five and three neurons, respectively, were established, all implementing the "ReLU" activation function to maintain consistency in nonlinear processing throughout the network [25].

- Output Layer: Consisting of a single neuron, this layer symbolizes the endpoint of the network, whose function is to output the model prediction. The choice of the "sigmoid" activation function is congruent with the binary nature of the output, which translates into a spectrum of values between 0 and 1, representing respectively low and high academic performance.

The selection of an artificial neural network (ANN) for our study, despite the modest size of the data set, was driven by the superior ability of ANN to model the inherent complex and nonlinear relationships between socioeconomic factors and academic performance. ANNs excel at detecting subtle patterns within data, a critical advantage over traditional models such as SVMs or decision trees, which may not capture the intricate dynamics of our study's approach. Furthermore, ANNs offer scalability and adaptability for future research expansions, making them a strategic option to achieve high predictive accuracy in this domain. Recognizing the potential risks of overfitting associated with smaller data sets, we implemented specific techniques such as dropout and early stopping to ensure the robustness and reliability of our model, thus justifying the use of ANN as a deliberate and informed choice for our objectives. research.

The following hyperparameters were used for training:

Activation function ReLU

$$max\ (0, x) \tag{1}$$

Activation function Sigmoid

$$S(x) = \frac{1}{1+e^{-x}} \tag{2}$$

Optimizer Adam

$$m_t = \beta m_t + (1-\beta)\left[\frac{\partial L}{\partial w_t}\right] \tag{3}$$

Binary_CrossEntropy

$$H_P(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i\ log\big(p(y_i)\big) + (1-y_i) log\ (1 - p(y_i)) \tag{4}$$

Metrics Accuracy

$$Acurracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5}$$

Metrics Loss

$$Log\ loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log\ (p_{ij}) \tag{6}$$

EarlyStopping

$$\int_P\ (x) = \int_Y\ y\ dp\ (y|x), x \in X \tag{7}$$

Table 1. Hyperparameters for training of the artificial neural network

Hyperparameter	Value
Epochs	100
Activation function	ReLU & Sigmoid
Optimizer	Adam
Loss function	Binary_CrossEntopy
Metrics	Accuracy & Loss
Overfitting	EarlyStopping
Dropout	0.5

To counteract overfitting in our neural network, we have implemented a robust strategy that involves partitioning the dataset into 80% for training and 20% for validation, complemented by a limit of 200 training epochs where the backpropagation algorithm is applied. Essentially, this approach is designed to tune the weights and biases by optimizing the loss function. Key to our regularization tactic is the integration of Dropout layers in all hidden layers, a proven technique to avoid over-reliance of the network on individual features and encourage generalization. This carefully calibrated training protocol provides a balance between effective learning and preventing excessive memorization of training data, positioning the model as a reliable predictive tool for academic performance evaluation.

2.1.8. Random forest

To ascertain the optimal hyperparameter configuration that maximizes the performance of our model, the GridSearchCV method from the sklearn library was utilized. This systematic technique probes an array of predefined hyperparameter values and selects the set that yields the best outcomes in accordance with the training data. It is crucial to note that GridSearchCV was employed to optimize various machine learning models, not merely a single type. The hyperparameters considered in the exhaustive search for the Random Forest model included:

- n_estimators: An array of values [100, 300, 500, 700], to define the number of trees in the forest.

- max_features: Options of [“auto”, “sqrt”, “log2”], specifying the number of features to consider when looking for the best split.

- max_depth: A range of [None, 5, 10, 15, 20], to set the maximum depth of the trees.

- min_samples_split: An array of [2, 5, 10], determining the minimum number of samples required to split a node.

- min_samples_leaf: Values of [1, 2, 4], for the minimum number of samples required at a leaf node.

Following the execution of GridSearchCV, the hyperparameters selected for the Random Forest model were:

- n_estimators= 100,
- max_features= sqrt,
- max_depth= None,
- min_samples_split= 2,
- min_samples_leaf= 4

It is noteworthy to mention that the feature_importances function was utilized within this model to identify the most salient features relative to the dataset. As a result, “condition of house” emerged as the most significant characteristic, as depicted in Fig. 5.

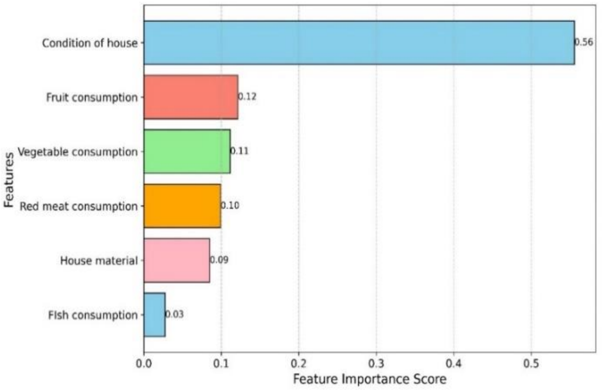

Fig. 5. Feature importance score with random forest.

2.1.9. SVM

In the case of the Support Vector Machine (SVM) model, the aforementioned GridSearchCV function was deployed, utilizing the following input parameters:

- C: A set of values [0.1, 1, 10, 100], which controls the penalty of the error term.

- gamma: An array [1, 0.1, 0.01, 0.001], that defines the influence of a single training example.

- kernel: A selection from [“rbf”, “poly”, “sigmoid”], determining the type of hyperplane used to separate the data.

The hyperparameters that were ultimately chosen for the SVM model included:

- C= 10,
- gamma= 0.001,
- kernel= “rbf”.

3. Results

This section presents the results of the proposed process, as explained in the contribution section. It is worth recalling that the proposed algorithm is an artificial neural network, which predicts academic performance based on socioeconomic factors among students in the eighth cycle of Mining Engineering at a National University in Trujillo.

Below is the confusion matrix (Fig. 6) of the artificial neural network model after training. It illustrates that of a total of 24 students with high academic performance or who have achieved an average grade between 14-20 and are located at the achieved level, where 24 were correctly predicted, while of 16 students with low academic performance or who placed at the initial level of academic achievement with average scores of 0-30, 14 were accurately predicted.

Furthermore, with the random forest model, Fig. 7 shows the confusion matrix, indicating that of a total of 24 students with high academic performance, 16 were correctly predicted, while 8 were misclassified. Similarly, of 16 students with poor academic performance, 11 were accurately predicted and 5 were misclassified.

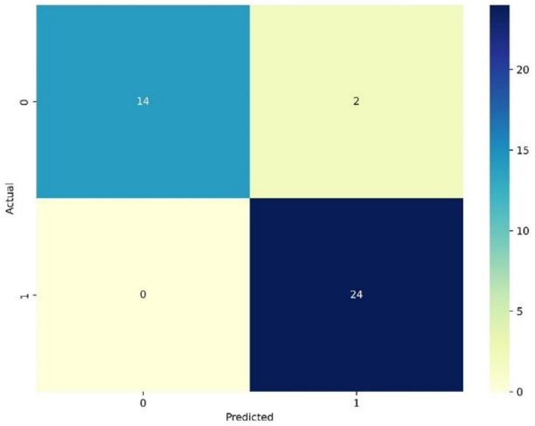

Fig. 6. Artificial Neural Network confusion matrix.

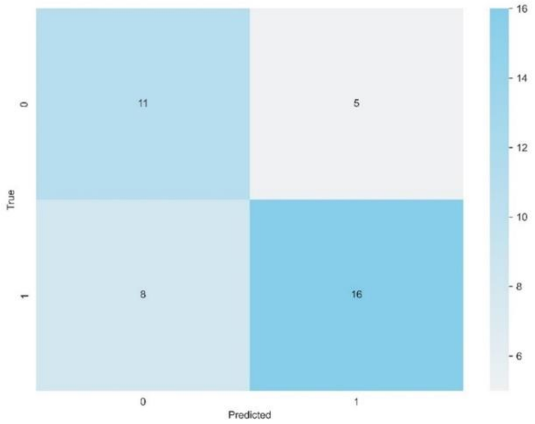

Fig. 7. Random forest confusion matrix

The confusion matrix of the SVM model shown in Fig. 8 indicates that out of a total of 24 students with high academic performance, or those who have achieved a score between 14-20 and are categorized in the "achieved" level, only 16 they did it correctly. predicted, while 8 were misclassified. Similarly, of 16 students with low academic achievement, or those categorized at the "initial" academic achievement level with scores ranging between (0 and 13), only 11 were correctly predicted and 5 were misclassified.

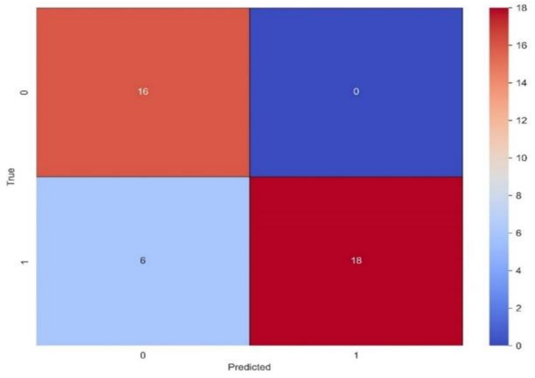

Fig. 8. SVM confusion matrix

Table 2 presents the results of the confusion matrices and accuracy scores of the proposed and developed models. The compared models were trained and validated using the same dataset. The neural network model achieved the highest accuracy of 75%, followed by SVM with 62.5%, and finally, random forest with 50.0%.

Table 2. Results of the confusion matrices

Models	Value positives	False positive	True negatives	False negatives	Accuracy
ANN	24	0	14	2	75.0%
RF	16	8	11	5	50.0%
SVM	18	6	16	0	62.5%

To track the evaluation of the training and validation process of the artificial neural network, precision and loss metrics were utilized for each epoch. Fig. 9 illustrates the recorded precision and loss for each epoch, representing the evolution of a neural network model over 100 epochs. It is observed that the loss (blue line) starts approximately at 0.9 and exhibits a decreasing trend, stabilizing around 0.6, indicating an effective learning process. The precision (orange line) begins around 0.4 and rises, reaching a plateau near 0.6, demonstrating an improvement in the model's predictions. The validation loss (green line) and validation precision (red line) fluctuate but generally maintain stability, indicating consistent performance of the model on unseen data.

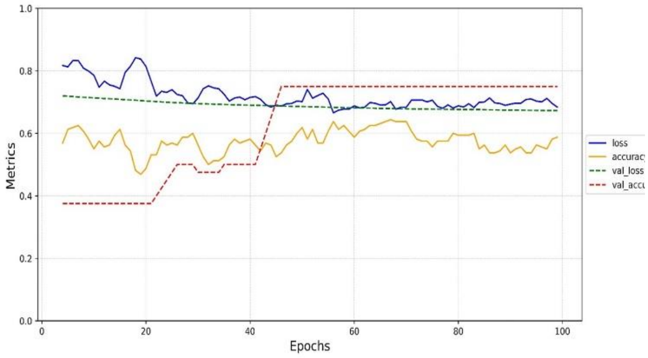

Fig. 9. Precision and loss of artificial neural network.

4. Discussions

This section compares the performance of the proposed model with those models that have shown the best performance according to the literature. We implemented two machine learning models, which were trained using the same dataset as the neural network to facilitate comparison. Benablo [11] evaluated SVM and KNN techniques, noting that the SVM model achieves a 98% accuracy due to the type of model used to predict performance based on input data. Furthermore, it details three additional models, highlighting variations in dataset instances. The larger the number of instances, the higher the accuracy, occasionally reaching 100%. Therefore, we chose to implement our own SVM model, training it with our dataset for comparative purposes. The SVM results were lower than those of the artificial neural network. Following the comparison with Amazona and Hernández's [12] proposal, three different models were utilized: Naive Bayes, deep learning, and decision tree, with accuracies of 96%, 98%, and 93%, respectively.

In their research, Rodríguez-Hernández [13] implemented an artificial neural network to predict the academic performance of final-year students. The study indicates that data from a sample of 162,030 students were used, and the trained model provided an accuracy of 82%. However, personal information about the students (e.g., socioeconomic status, household characteristics, personal background, types of schools attended, and students' earnings from work) was also utilized to train the model. The models applied included random forest, neural network, SVM, logistic regression, Naive Bayes, and KNN, with the neural network achieving an accuracy of 86.30%. Yagci [14] did not use students' personal data, relying solely on information such as midterm and final exam grades, career path, and academic department. The 86% accuracy rate achieved in this research study is the highest recorded accuracy rate. Thus, a proprietary Random Forest model was implemented, trained with the same dataset as our proposed model. The Random Forest model's outcome was the least favorable, with an accuracy of 50.0%.

Zevallos [21] and Capuñay [26] explored the implementation of neural networks in the prediction of academic performance. Despite focusing on different educational levels, both studies show the important potential of neural networks in this area. In our research, we developed a supervised neural network composed of an input layer with six neurons, representing various socioeconomic factors. The network configuration, which includes four hidden layers, was designed to maximize predictive accuracy. Using the "adam" optimizer, recognized for its effectiveness in complex optimization problems, we train the network with 60% of the data set and validate it with the remaining 40%. When comparing these values with our implemented models, in this research the "adam" optimizer was used and the data set was divided into 80/20, where 80% of the data were for training and 20% for validation. These findings agree with the research of Véliz [27], which highlighted the relevance of socioeconomic variables in predicting academic performance.

The findings of Zevallos [21] and Capuñay [25] already pointed out the need for a model that quickly fits the data, a feature evident in our loss curve graph. This rapid initial decline confirms the model's ability to learn efficiently. An essential aspect to determine the accuracy of a regression model is the residual histogram. Observing that the normal distribution (Gaussian bell) is centered at 0.00 confirms that the model does not have systematic biases in its estimates, which is an empirical validation of its adequate performance. The presented metrics shed light on the effectiveness of the neural network: with a mean absolute error (MAE) of 4.75%, a slight deviation of the model predictions from the actual values is recognized. However, this data, combined with a loss metric of 0.38% and accuracy of 95.25%, gives us an overall picture of a robust and reliable model. When comparing these results with our trained models, we can affirm that our neural network model obtained the highest precision with 75.0%, followed by SVM with 62.5% and finally Random Forest with 50.0%.

Comparing our model with the literature, we observe that personal data significantly impact model outcomes, regardless of the model type used. For this reason, and due to the increasing vulnerability of personal data highlighted by breaches and data leaks, we propose an academic performance prediction tool for students that does not expose confidential student information and yields results comparable and close to those observed in the literature, which can be further enhanced in future work.

5. Conclusion

In this research, a model of artificial neural network was proposed to predict the academic performance of students in the eighth and tenth cycles of the Mining Engineering career at the National University of Trujillo using socioeconomic factors, achieving an accuracy of 75%, without exposing their personal data.

The issue of low student performance impacts both educational institutions and the students themselves. In response to this, we have created a model that goes beyond existing proposals in the literature and can anticipate the weighted average of an academic cycle. This tool represents a valuable source of information for educational authorities, allowing them to take proactive measures to combat low performance and improve educational quality.

5.1. Model limitations.

Despite the significant advancements this study represents in predicting academic performance using socioeconomic factors through a supervised artificial neural network, it is crucial to acknowledge certain inherent limitations to our approach. Firstly, the selection of socioeconomic variables, though meticulously chosen, does not exhaustively cover all possible factors that may influence academic performance. Variables such as family support, access to educational resources outside the home, and mental health, among others, could further enrich our model. Secondly, the sample size of 40 students, while providing valuable initial insights, limits the generalization of our findings to broader populations or different educational contexts. The implementation of Dropout and other strategies to counter overfitting, though effective, also reflects the ongoing challenge of balancing the model's predictive capacity with the complexity of real data. Lastly, the research focused exclusively on mining engineering students from the VIII and X cycles at the National University of Trujillo, Huamachuco, implying a need to validate the model's applicability in other academic programs and educational levels. These limitations underscore significant opportunities for future research, including the expansion of the dataset, the incorporation of additional variables, and the extension of the model to various academic disciplines and educational contexts.

5.2. Future research.

Future research should broaden the range of variables considered. The analysis could be deepened and enriched by incorporating psychological dimensions, economic factors, and consideration of the places where the students attended primary and secondary school.

Author contributions

Marco Cotrina: Methodology, Writing-Original draft preparation, Field study **Jairo Marquina:** Software, Investigation, Validation., Field study **Eduardo Noriega:** Visualization, Software., Field study **Jose Mamani**: Investigation., Field study **Eusebio Antonio:** Writing-Reviewing and Editing., Field study **Solio Arango:** Data curation., Field study **Hans Portilla:** Conceptualization.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] H. Chavez, B. Chavez-Arias, S. Contreras-Rosas, J. Alvarez-Rodríguez y C. Raymundo, “Artificial neural network model to predict student performance using nonpersonal information,” *Frontiers in Education*, 2023, 10.3389/feduc.2023.1106679.

[2] A. Singh, V. Bhadauria y G. Mangalaraj, “A Teaching Module Illustrating ERP Item Value Automation,” *Journal of Information Systems Education*, vol. 34, nº 1, pp. 16-31, 2023.

[3] K. Schürholt, D. Kostadinov y D. Borth, “Hyper-Representations: Self-Supervised Representation Learning on Neural Network Weights for Model Characteristic Prediction,” *35th Conference on Neural Information Processing Systems (NeurIPS*

2021), 2022.

[4] F. Tejedor y A. García-Valcárcel, "Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos) Propuestas de mejora en el marco del EEES," *Revista de Educación*, nº 342, pp. 443-473, 2007.

[5] SINEACE, *Modelo de Acreditación para Programas de Estudios de Educación Superior Universitaria*, Lima-Perú, 2016.

[6] S. Liao, D. Zingaro, K. Thai, C. Alvarado, W. Griswold y L. Porter, "robust machine learning technique to predict low-performing students," *ACM Trans. Comput. Educ.*, nº 19, pp. 1-19, 2019, 10.1145/3277569.

[7] A. Hellas, P. Ihantola, A. Peterson, V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom y S. Nam, "Predicting academic performance: a systematic literature review," *ITiCSE 2018 Companion: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pp. 175-199, 2018, 10.1145/3293881.3295783.

[8] A. Namoun y A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Applied Sciences*, vol. 11, nº 1, p. 237, 2021, 10.3390/app11010237.

[9] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. Murray y Q. Long, "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Systems*, vol. 161, nº 1, pp. 134-146, 2018, 10.1016/j.knosys.2018.07.042.

[10] Y. Muhammad, M. Abul, S. Almotairi, K. Farooq, F. Granelli y L. Strážovská, "The Role of Socioeconomic Factors in Improving the Performance of Students Based on Intelligent Computational Approaches," *Electronics*, vol. 12, nº 9, p. 1982, 2023, 10.3390/electronics12091982.

[11] C. Benablo, E. Sarte, J. Dormido y T. Palaog, "Higher education Student´s academic performance analysis through predictive analytics," *Proceedings of the 2018 7th International Conference on Software and Computer Applications-ICSCA 2018*, pp. 238-242, 2018, 10.1145/3185089.3185102.

[12] M. Amazona y A. Hernandez, "Modelling student performance using data mining techniques," *in Proceedings of the 2019 5th International Conference on Computing and Data Engineering—ICCDE' 19*, pp. 36-40, 2019, 10.1145/3330530.3330544.

[13] H. Rodríguez, M. Musso, E. Kyndt y E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Computers and Education: Artificial Intelligence*, vol. 2, 2021, 10.1016/j.caeai.2021.100018.

[14] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ*, vol. 9, nº 11, 2022, 10.1186/s40561-022-00192-z.

[15] C. Cuccurullo, M. Aria y F. Sarto, "Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains," *Scientometrics*, vol. 108, pp. 595-611, 2016, 10.1007/s11192-016-1948-8.

[16] S. Ahmed, M. Alshater, A. El Ammari y H. Hammami, "Artificial intelligence and machine learning in finance: A bibliometric review," *Research in International Business and Finance*, vol. 61, 2022, 10.1016/j.ribaf.2022.101646.

[17] V. Gil-Vera y C. Quintero-López, "Predicción del rendimiento académico estudiantil con redes neuronales artificiales," *Información Tecnológica*, vol. 32, nº 6, pp. 221-228, 2021, 10.4067/S0718-07642021000600221.

[18] A. Vargas, "*Predicción del rendimiento académico empleando algoritmos de aprendizaje supervisado en estudiantes del primer semestre de la carrera profesional de ingeniería de sistemas e informática de la UNAMAD, 2020*," Puerto Maldonado, 2022.

[19] S. Cabana, "*Análisis predictivo del rendimiento académico en los alumnos de la escuela profesional de ingeniería en informática y sistemas de la UBJBG, utilizando redes neuronales, semestre 2017-I*," Tacna, 2018

[20] E. Saire, "*Predicción de la ruta de rendimiento académico con algoritmos de clasificación*," Arequipa, 2023.

[21] R. Zevallos, "*Predicción del rendimiento académico mediante redes neuronales*," Callao-Lima, 2017.

[22] J. Blanco, S. Lovelle, R. Fernandez y E. Perez, "Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales," *Ingeniare. Revista chilena de ingeniería,* vol. 24, pp. 715-727, 2016, 10.4067/S0718-33052016000400015.

[23] E. Rincon-Flores, E. López-Camacho, J. Mena y O. López, "Predicting academic performance with artificial intelligence (AI), a new tool for teachers and students," *in 2020 IEEE Global Engineering*

Education Conference (EDUCON), pp. 1049-1054, 2020, 10.1109/EDUCON45650.2020.9125141.

[24] J. Heaton, "Introduction to Neural Networks with Java," *hesterfield, MO: Heaton Research*, p. 129, 2009.

[25] Y. Bai, "RELU-Function and Derived Function Review," *SHS Web of Conferences* 144, 2022, 10.1051/shsconf/202214402006.

[26] D. Capuñay, "*Modelo basado en redes neuronales para proyectar el rendimiento académico de segundo grado de secundaria en la Institución Educativa N°16093-Jaén*," Chiclayo, 2021.

[27] H. Véliz, "*Aplicaciones de redes neuronales para evaluar el desempeño académico de los maestristas de la unidad de posgrado de ingeniería de sistemas*," Huancayo, 2022.