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

https://doi.org/10.23913/ride.v16i32.2841

Scientific articles

Predicción del rendimiento académico con un modelo de clasificación: una comparación entre estudiantes rurales y urbanos de educación media superior

Prediction Using Two Classification Models: A Comparison between Rural and Urban Upper Secondary Students

Previsão do desempenho acadêmico com um modelo de classificação: uma comparação entre estudantes do ensino médio de áreas rurais e urbanas

Yolanda Moyao Martínez
Benemérita Universidad Autónoma de Puebla, México
yolanda.moyao@correo.buap.mx
https://orcid.org/0000-0002-7259-3525

Carmen Cerón Garnica
Benemérita Universidad Autónoma de Puebla, México
carmen.ceron@correo.buap.mx
https://orcid.org/0000-0001-6480-6810

Resumen

Este trabajo de investigación tuvo como propósito predecir el aprovechamiento académico por debajo de la meta esperada, definida a partir del promedio mínimo aprobatorio establecido por la institución, en estudiantes de nivel medio superior en Puebla a través del uso de modelos predictivos de aprendizaje automático supervisado, basados en técnicas de clasificación y minería de datos. En este estudio, el término *aprovechamiento académico* se emplea como indicador del *rendimiento académico* de los estudiantes. Para ello, se procesaron dos conjuntos de información provenientes de una preparatoria rural y una urbana de la Benemérita Universidad Autónoma de Puebla (BUAP). Se empleó el algoritmo de clasificación de bosques aleatorios *(Random Forest*) propuesto por Breiman (2001), el cual fue entrenado y evaluado de manera independiente en ambos conjuntos de datos, utilizando

como métricas de desempeño la precisión y la sensibilidad (r*ecall*). En ambos conjuntos de datos se obtuvo una precisión del 0.72, lo que indica un desempeño comparable del modelo en los contextos rural y urbano. El modelo permitió identificar elementos clave relacionados con el aprovechamiento académico, tales como, la asistencia a clases y el aprovechamiento anticipado en algunas materias, los cuales mostraron una correspondencia descriptiva con el nivel de escolaridad. Asimismo, los hallazgos revelaron algunas diferencias entre ambos contextos, lo que sugiere la necesidad de implementar propuestas de mejora adaptadas a cada contexto, pero que sean muy particulares a cada entorno. A partir de estos resultados, se derivan implicaciones directas para las organizaciones educativas y el sector público, particularmente en el diseño de estrategias orientadas a la prevención del rezago académico y con ello, la deserción escolar en ambos contextos.

Palabras clave: bosques aleatorios, contexto geográfico, deserción escolar, minería de datos, modelos de clasificación.

Abstract

This work aimed to predict academic achievement below the expected target, defined according to the minimum passing grade established by the institution, among upper-secondary education students in Puebla through the use of predictive models based on supervised machine learning, employing classification techniques from data mining. In this study, the term *academic achievement* is used as indicator of *academic performance* of the students. For this purpose, two datasets obtained from a rural and urban upper secondary school of the Benemérita Universidad Autónoma de Puebla (BUAP) were processed. The Random Forest classification algorithm, proposed by Breiman (2001), was employed and trained and evaluated independently on both datasets, using accuracy and sensitivity (recall) as performance metrics. An accuracy of 0.72 was obtained for both datasets, indicating comparable model performance in rural and urban contexts. The model made it possible to identify key factors related to academic performance, such as class attendance and prior achievement in specific subjects, which showed a descriptive association with the level of educational attainment. Furthermore, the findings revealed differences between both contexts, suggesting the need for improvement strategies tailored to the specific characteristics of each environment. Based on these results, relevant implications emerge for educational institutions and the public sector, particularly for guiding the design of strategies

aimed at preventing academic underachievement and, consequently, reducing school dropout in both rural and urban contexts.

Keywords: random forests, geographical context, school dropout, data mining, classification models.

Resumo

Esta pesquisa teve como objetivo predizer o desempenho acadêmico abaixo da meta esperada, definida pela média mínima de aprovação estabelecida pela instituição, em alunos do ensino médio de Puebla, utilizando modelos preditivos de aprendizado de máquina supervisionado baseados em técnicas de classificação e mineração de dados. Neste estudo, o termo "desempenho acadêmico" é utilizado como indicador do desempenho acadêmico dos alunos. Dois conjuntos de dados foram processados, um de uma escola de ensino médio rural e outro de uma escola de ensino médio urbana da Benemérita Universidad Autónoma de Puebla (BUAP). O algoritmo de classificação Random Forest, proposto por Breiman (2001), foi utilizado, sendo treinado e avaliado independentemente em ambos os conjuntos de dados, utilizando acurácia e recall como métricas de desempenho. Uma acurácia de 0,72 foi obtida em ambos os conjuntos de dados, indicando desempenho comparável do modelo nos contextos rural e urbano. O modelo permitiu a identificação de elementos-chave relacionados ao desempenho acadêmico, como frequência às aulas e aproveitamento precoce em algumas disciplinas, que apresentaram correlação descritiva com o nível de escolaridade. Da mesma forma, os resultados revelaram algumas diferenças entre os dois contextos, sugerindo a necessidade de implementar propostas de melhoria adaptadas a cada contexto, mas altamente específicas para cada ambiente. A partir desses resultados, surgem implicações diretas para organizações educacionais e o setor público, particularmente no desenvolvimento de estratégias voltadas para a prevenção do baixo rendimento acadêmico e, consequentemente, da evasão escolar em ambos os contextos.

Palavras-chave: florestas aleatórias, contexto geográfico, evasão escolar, mineração de dados, modelos de classificação.

Date Received: August 2025 **Date Accepted:** February 2026

Introduction

In educational institutions in the state of Puebla, the issue of academic achievement below the state average or the minimum standards established by the Ministry of Public Education (SEP) is frequently addressed, and school dropout rates are a possible consequence. This study focuses particularly on the upper secondary level; however, this situation continues to pose a significant problem for both upper secondary educational institutions and state education authorities.

In practice, it is common that, even though educational institutions possess school datasets, they lack sufficient technological tools to process and analyze this information in order to develop educational strategies that contribute to the prevention and mitigation of this problem. Therefore, two datasets differentiated by geographic area were analyzed: one with information on urban students and the other with information on rural students, both at the upper secondary level in Puebla (see the Methodology section for details on sample size and data collection).

Although there are some international research studies that address this problem through the use of data mining strategies, such as the studies by (Cortez & Silva, 2008; Romero & Ventura, 2020) who apply different data mining models to academic datasets of *secondary school students school* (equivalent to upper secondary education in the Mexican context). The main objective of this research is to identify key factors that impact academic achievement and to predict, with a certain degree of accuracy, whether students will succeed or face difficulties during their schooling. Based on this background, the present study adopts a data mining approach to analyze its applicability in the upper secondary education context in Puebla.

Regarding performance measures, these studies do not use metrics such as precision or sensitivity ; instead, they use global accuracy (PCC) to assess classification tasks and root mean square error (RMSE) for regression tasks. According to the authors, the best prediction results are achieved when variables (G1 and G2) representing the grades from the first two school terms are included (Cortez & Silva, 2008). Therefore, they conclude that predictive models can be used as valuable tools to anticipate academic achievement with a high level of accuracy (Cortez & Silva, 2008) .

Although international studies have demonstrated the effectiveness of predictive models in education, their application has been primarily in contexts where socioeconomic and educational conditions are more equitable for students. In contrast, the state of Puebla

exhibits significant gaps between rural and urban areas that considerably influence academic continuity and achievement, consequently increasing the risk of school dropout. Recent studies have shown that students from rural areas face greater disadvantages such as inadequate educational infrastructure, limited access to technological resources, and socioeconomic challenges (INEGI, 2022; CONEVAL, 2023). Therefore, it is essential to adapt these models to the local context.

In Mexico, and particularly in the context of Puebla, there is little research that allows for comparing results and evaluating the efficiency of these tools and models in local contexts, which limits educational institutions in making evidence-based decisions.

In the state of Puebla, high school dropout rates remain a critical challenge requiring immediate attention, as Puebla exhibits the most pronounced urban-rural gap in the country in terms of school abandonment. According to the Ministry of Public Education (SEP, 2023), rural communities show the highest rates of absenteeism and academic underachievement. Nationally, high school dropout rates reached 13.5% in rural areas, compared to 8.2% in urban areas (CONEVAL, 2023), highlighting the significant impact of students' backgrounds on academic performance.

On the other hand, in Puebla, the school dropout rate reached 11.3% during the 2021–2022 school year. For the 2022–2023 and 2023–2024 school years, it decreased to 9.7% and 9.5%, respectively. Even so, this remains the highest among the different educational levels in the state (SEP, 2023).

Similarly, various studies have shown that factors such as distance to school and the availability of public transportation make a difference in rural areas. According to the Ministry of Public Education (General Directorate of Educational Planning, Programming, and Statistics, 2024), students have to travel 40 to 60 minutes daily to get to school, which impacts attendance and, consequently, academic performance. Conversely, in urban areas, the variables with the greatest impact are school absences, school choice based on proximity, and certain extracurricular activities, reflecting trends that this study's model seeks to analyze.

Although slight progress has been observed, the issue remains a challenge for educational institutions. Empirical evidence shows that the difference between students in rural and urban contexts is not only due to individual factors, but also to structural differences inherent to each context. Therefore, there is a clear need to develop comparative and predictive analysis tools for academic achievement in order to promptly identify factors that contribute to school

dropout and thus plan differentiated prevention strategies according to the context. This not only allows for practical application in the review and improvement of educational policies, but also contributes to understanding how structural and contextual conditions can affect academic achievement in the state of Puebla.

In this context, the following research question arises: How effective are predictive models based on data mining in preventing low academic performance among high school students in Puebla? The objective of this study is to design and evaluate the use of data mining techniques to better understand and anticipate the academic performance of high school students in Puebla. Therefore, two datasets differentiated by geographic area were analyzed: one rural and one urban. These datasets exhibit educational differences associated with structural factors of the territory (Pérez Pérez et al., 2025). Statistical metrics were also used to evaluate the model's performance.

This article is organized as follows: first, the theoretical framework is presented based on educational, psychological, sociological, and computational theories, along with a review of the state of the art related to predictive models; next, the methodology used in the development of the predictive model is described; then, the results obtained are presented and analyzed; and finally, the conclusions and recommendations directed to entities in the educational and public sectors are presented.

Theoretical framework

To gain a clearer understanding of the process of academic achievement and the factors associated with it, it is enough to review the conceptual support provided by some of the educational, psychological, social, and computational theories, as well as their integration into predictive models.

Data mining allows for the development of predictive models of academic achievement and risk, facilitating the early detection of factors that may influence academic performance. These algorithms can identify sociodemographic, psychological, and economic factors, among others, with the aim of implementing educational support strategies for at-risk students.

Therefore , in order to design effective predictive models, it is important to review some of the theories that explain the various dimensions that influence learning, such as the theory of meaningful learning, the theory of motivation and the theory of cultural capital (Bourdieu & Passeron, 1990), in order to understand how these interact with academic achievement,

laying the foundations that will allow the identification of the key factors in predictive algorithms.

According to Ausubel (1968), the theory of meaningful learning emphasizes the importance of prior knowledge as a foundation for effective learning, allowing for the anticipation of success in specific subjects. Piaget (1972) highlights the role of cognitive development in knowledge construction, indicating that intellectual maturity influences learning capacity, which in turn supports the prediction of academic achievement.

For his part, Vygotsky (1978) emphasizes that the social context and the zone of proximal development (ZPD) significantly influence cognitive development, providing key variables for predicting academic achievement. Likewise, these concepts are reflected in the differentiated academic management required in rural and marginalized urban areas (Pérez Pérez et al., 2025).

In the psychological context, Bandura (2001, pp. 7-10) highlights that those with high self-efficacy act and think about the belief in their own capabilities and skills, in such a way that this action translates into the motivation, confidence and initiative of the students, allowing them to achieve better academic performance.

Similarly, Ryan & Deci 's motivation theory (2000, pp. 1-3), applied in different fields, including education, is seen as a broad model that explains why people feel motivated to perform certain activities and how this factor positively influences academic achievement.

Self-Determination Theory identifies three basic psychological needs:

i. Autonomy: perception of personal control over one's actions and decisions;
ii. Competence: the ability to perform assigned tasks and obtain satisfactory results; and
iii. Relationship: inclusion in work groups in order to receive social support.

Meeting these three needs allows students to develop skills and knowledge in a meaningful and lasting way, promoting meaningful learning and academic success.

On the other hand, Bourdieu & Passeron (1990, pp. 31-33) suggest that academic achievement may depend largely on cultural capital [1]transmitted from generation to generation, in addition to individual academic achievement and available resources. It can be inferred that these elements enable students with greater cultural capital to have more

[1]Bourdieu & Passeron (1990) state that habitus includes cultural capital, enduring habits, and is composed of knowledge, skills, abilities, habits, and cultural content.

opportunities to achieve academic success, unlike those with less cultural capital, since this inequality acts as a barrier that limits the development of their academic skills.

Buschini (2023, pp. 448-452) revisits Luhmann's theory to highlight that the organizational structure of the educational system is a significant factor shaping the educational process and conditioning teaching and learning processes. The school, as a social and organizational system, is not only based on a hierarchical structure that integrates different functions, but also possesses a degree of autonomy that determines how learning is organized and developed within the system, significantly impacting this process.

Finally, from a computational perspective, all these elements can be integrated into algorithms capable of predicting academic achievement based on large educational datasets. To achieve this, the *Cross Industry Standard Process data mining model is adopted. The Critical-Related-Process Data Mining* (CRISP-DM; Chapman et al., 2000) model is recognized as a systematic and standard method for developing data mining projects and predictive models. This model consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, and allows for the orderly organization of the construction and validation of the models applied in this study.

It can be concluded that their theories lay the groundwork for understanding how, through the integration of different factors and taking into account the educational and psychological approach, it can be inferred that relevant learning, motivation, and self-efficacy impact academic achievement. Finally, it can be understood that computational theories create the conditions for integrating these elements into predictive models.

These theoretical approaches have been extensively addressed in the literature. Several studies (Peña-Ayala, 2014; Romero & Ventura, 2020; Taylor *et al* ., 2016) indicate that educational data analysis tools have proven effective and are useful for uncovering significant trends that impact academic achievement. Furthermore, they facilitate evidence-based decision-making aimed at improving the teaching and learning process through the planning of appropriate strategies to reduce educational gaps.

Similarly, studies by Contreras-Bravo *et al* . (2021), Gil-Vera & Quintero (2021), Castrillón *et al* . (2020), and Díaz-Martínez *et al* . (2021) indicate that educational data mining employs various algorithms, such as random forests, artificial neural networks (ANNs), and logistic regression. These algorithms are used to analyze large volumes of academic data, detect relevant patterns, and predict the risk of low academic achievement, in

order to to identify students at risk, design appropriate interventions, and optimize teaching and learning processes.

educational data mining (EDM) strategies *has proven to be a useful resource for identifying key elements in academic achievement, with the aim of preventing school dropout risks and designing improvement actions. Recent studies have revealed how data* analysis facilitates the development of predictive models that aid in well-informed decision-making in educational settings, contributing to improved organizational effectiveness (Romero & Ventura, 2020; Merceron & Tato, 2023; Ordoñez-Avila *et al* ., 2023).

Main algorithms used in educational prediction models

Predictive models have evolved into useful tools for identifying factors that impact students' academic performance, supporting evidence-based decision-making by teachers and educational administrators. The most commonly used algorithms include random forests, artificial neural networks, logistic regression, and clustering *algorithms* .

According to Aguilar-Reyes (2025), the analysis of academic achievement using machine learning models was approached from two predictive perspectives: multinomial logistic regression and classification trees, which allowed for the identification of significant elements associated with student performance. In his results, the author reports that the logistic regression model achieved 100% accuracy in classification, while the model based on classification trees obtained 70.83% accuracy, under the experimental conditions and validation scheme described in his study.

In other research studies, the analysis has been conducted with higher education students using a predictive model based on artificial neural networks. Based on these results, the author concludes that variables such as positive family relationships, the number of absences, and previous grades significantly influence academic performance, and reports that the model achieved an 87% [2]predictive accuracy (Choque-Aguilar, 2024).

Similarly, Gil-Vera & Quintero (2021) present the development of a predictive model based on artificial neural networks (ANNs) to identify the risk of low academic achievement in upper secondary school students. Their study considered variables such as time spent studying, absences, and social media use, among other relevant variables. The model

[2]Choque-Aguilar (2024). Key variables such as positive family relationship, number of absences, and previous grades stand out.

achieved 73% accuracy [3]in classifying students, and the most significant predictive variables were study time, absences, and time spent using social media.

Specifically in the case of Mexico, Rico-Páez (2022) conducted research using three algorithms— Naive Bayes, k-NN (k nearest neighbors), and C4.5 decision tree—to predict the academic performance of students at a public university. The results show that, based on the first three assessable activities of the course, it is possible to achieve 70% accuracy in predicting whether a student will pass the course, demonstrating the feasibility of early assessment and timely student guidance.

On the other hand, Acosta-Gonzaga and Ramírez-Arellano (2020), researchers at the National Polytechnic Institute, conducted a study predicting academic achievement and showed that data mining techniques, support vector machines (*SVMs)* , and traditional statistical methods, particularly linear regression, are all effective. However, SVMs stand out for their superior predictive performance, while linear regression is better at identifying and analyzing the most significant factors associated with academic achievement.

Methodology

This section describes the methodological strategy of the study, including the research design, how the problem was addressed, the type of data required, the collection instruments, and the analysis methods that will be used to answer the research question.

Methodological approach

The methodological approach used was quantitative, as it allowed for the analysis of school datasets through the use of statistical and computational tools specific to data mining, such as classification algorithms and predictive achievement metrics. This approach was fundamental for anticipating students' academic performance by identifying patterns in the data. To support the choice of a non-experimental design and the quantitative approach, the contributions of the general methodology proposed by Hernández *et al* . (2014) were incorporated.

Furthermore, the evaluation of these predictive models was performed using the *hold-out method* , partitioning the dataset into two subsets: 80% allocated to training the random forest

[3]Gil-Vera & Quintero (2021). They highlight the most significant variables as time spent studying, absences from classes and the use of social networks.

model and 20% reserved for its evaluation. Statistical metrics, such as accuracy and sensitivity, as well as those reported in Tables 1 and 2, were calculated on the test set to estimate the model's predictive capacity on unseen data. This explains why the support shown in these tables is less than the total sample size and helps avoid overfitting bias, ensuring a neutral evaluation of its performance.

Furthermore, a comparative analysis of the results was conducted to propose strategies that contribute to improving decision-making processes in educational organizations. This was all done using a quantitative approach, based on statistically sufficient sample sizes to ensure the reliability of the results. Finally, this approach ensured methodological consistency when comparing the performance of the models in both analytical contexts.

This study employed a non-experimental, descriptive-correlational research design. Furthermore, the study is cross-sectional, as the analysis was conducted using data collected at a single point in time. This approach is appropriate due to the absence of variable manipulation and the observational nature of the analysis, given that it utilized existing datasets of upper secondary school students. Specifically, data from students residing in Tehuacán and Puebla City in the state of Puebla were analyzed to identify patterns, relationships, and key elements associated with academic achievement.

This descriptive approach characterizes the study group based on sociodemographic and school variables, while the correlational approach focuses on the relationship between these variables and academic achievement.

Sample characteristics

The sample size is 198 participants for the urban context and 384 for the rural context, providing a sufficient number of observations to support the reliability of the results, considering a 95% confidence level and an acceptable margin of error in educational studies. These sample sizes promote data representativeness and the reliability of the results, providing an appropriate basis for quantitative analysis and the validation of predictive models.

The inclusion criteria considered were:

a) students enrolled in the upper secondary level of BUAP;
b) to have complete information on the academic and sociodemographic variables used in the study;
c) belong to one of the defined contexts (urban or rural).

Records with incomplete data or inconsistencies in academic information were excluded.

A non-probabilistic convenience sampling method was used, considering the data sets available in rural and urban contexts.

The data collection instrument was a structured questionnaire, as shown in Appendix A, which was designed to gather key information for the study. Specifically, it allowed for the collection of data on the following factors: a) academic; b) personal and demographic; c) family and social; and d) contextual, with the aim of obtaining a comprehensive view of the elements that influence the academic performance of upper secondary school students in Puebla.

Furthermore, the questionnaire was designed based on instruments used in previous studies and is currently undergoing reliability assessment using Cronbach's alpha coefficient, in order to guarantee the internal consistency of the data considered.

Study variables and operationalization

The study variable was academic achievement, categorized into two levels: “High” and “Medium,” based on students’ grade point averages and following institutional evaluation criteria. The range [8.5, 10.0] was considered “High,” and the range [7.0, 8.49] was considered “Medium.” This categorization allowed the problem to be addressed as a supervised classification task for training the predictive model.

Results

Random forest model analysis

For the rural dataset, the Gini impurity criterion was used *, and* for the urban dataset, the entropy criterion was used *in constructing* the decision trees for the random forest model. The maximum tree depth was set to 10 levels as a regularization strategy. Two hundred trees were trained to improve prediction stability. Additionally, parameters were configured with a minimum leaf size of one sample and a minimum of two samples for node splitting, allowing for data partitioning.

Tables 1 and 2 show the results obtained regarding the overall performance of the model for the rural and urban approaches, respectively, which allows us to observe the following:

- The model's accuracy was slightly higher in the urban approach than in the rural one, suggesting a better overall classification capacity in this context.

- The macro accuracy obtained was 0.72 for the rural dataset and 0.73 for the urban dataset, reflecting a similar performance of the model in the two environments in terms of correct classification for the two classes.
- In both contexts, the macro sensitivity reached a value of 0.71, indicating an equivalent capacity of the model to correctly identify the real cases of each class in both rural and urban areas.
- The *F1-score* macro was 0.71 in both datasets, which shows a good balance between accuracy and sensitivity of the model in both environments analyzed.

Classification report

Tables 1 and 2 show the results of the evaluation metrics achieved by the model on the test dataset for each of the academic achievement classes “High” and “Medium” in the rural and urban contexts, respectively.

Table 1. Performance indicators in the rural context

Class	Precision	Sensitivity	*F1-score*	Support
High	0.70	0.76	0.73	21
Half	0.74	0.67	0.70	21
Accuracy			0.71	42
Macro average *(Macro avg)*	0.72	0.71	0.71	42
Weighted average *avg)*	0.72	0.71	0.71	42

Source: Prepared by the author based on analysis using Python

- "Upper" Class:
  - A sensitivity was recorded in the rural dataset of 0.76, while in the urban context it was 0.85, indicating an adequate ability to identify students with high academic achievement.
  - The accuracy was moderate, with a value of 0.70 in the rural context and 0.69 in the urban context, which shows that some of the instances classified as "High" actually corresponded to the "Medium" class.
  - The *F1-score* was 0.73 in the rural dataset and 0.76 in the urban dataset, suggesting a good balance between accuracy and sensitivity, and therefore an overall performance suitable for this category.

Table 2. Performance indicators in the urban context

Class	Precision	Sensitivity	*F1-score*	Medium
High	0.69	0.85	0.76	13
Half	0.78	0.58	0.67	12
Accuracy			0.72	25
Macro average	0.73	0.71	0.71	25
Weighted average	0.73	0.72	0.71	25

Source: Prepared by the author based on analysis using Python

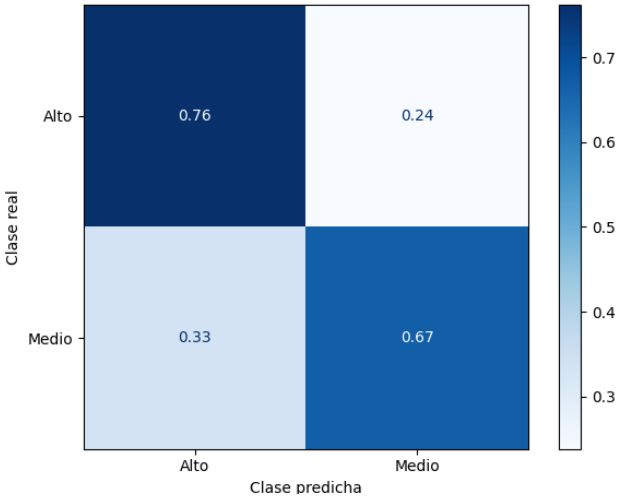
- Middle Class:
  - The accuracy was 0.74 in the rural dataset and 0.78 in the urban dataset, indicating that most of the predictions made for this category were correct.
  - Regarding sensitivity, a value of 0.67 was obtained for the rural context and 0.58 for the urban context, indicating that a proportion of the students belonging to this class was classified as "High".
  - The *F1-score* was 0.70 in the rural dataset and 0.67 in the urban dataset, reflecting acceptable performance when considering both accuracy and sensitivity for this category.

Normalized confusion matrix

For the rural and urban datasets, Figures 1 and 2 show the confusion matrix, whose entries represent the proportion of normalized cases for each combination of actual and predicted classes. It can be seen that, when the actual class is “High,” the model correctly classifies 0.76 of the cases in the rural context and 0.85 in the urban context, while it makes errors in 0.24 and 0.15 of the cases, respectively, by assigning them to the “Medium” class.

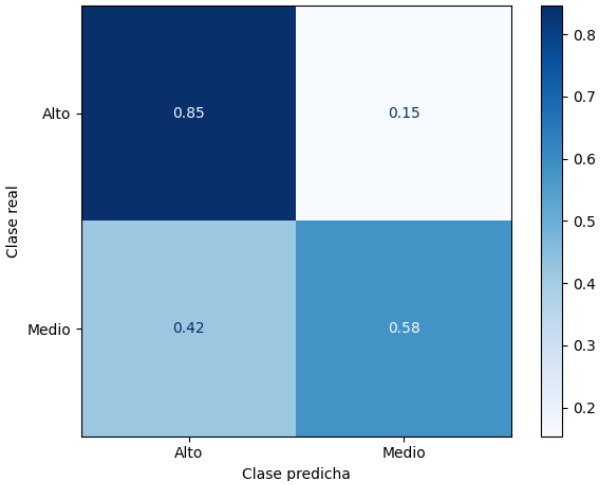
Likewise, when the actual class is “Medium”, the model correctly classifies 0.67 of the cases in the rural context and 0.58 in the urban context, while incorrectly classifying 0.33 and 0.42 of the cases, respectively, by assigning them to the “High” class, which suggests a tendency to overclassify the “High” category.

Figure 1. Row-normalized confusion matrix for the rural context



Source: Predictive model implemented in Python

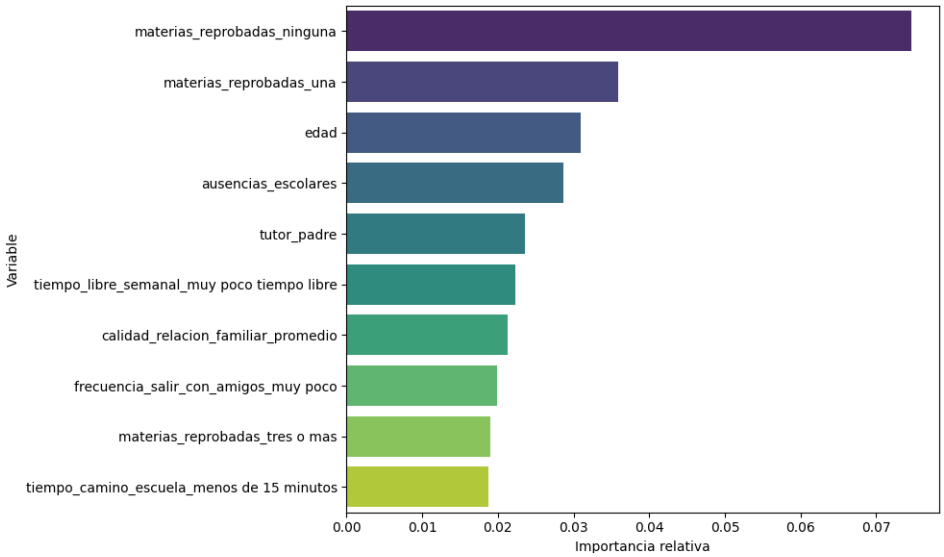
Figure 2. Row-normalized confusion matrix for the urban context



Source: Predictive model implemented in Python

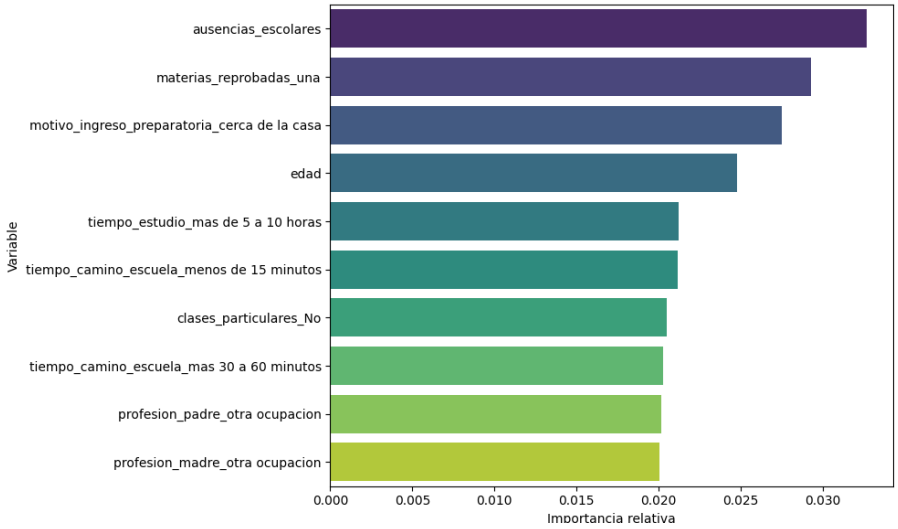
For the rural and urban datasets, the bar charts in Figures 3 and 4, respectively, show the key variables used by the model to predict student academic achievement. The level of importance of each variable was assessed based on its proportional contribution to the decrease in impurity in the model trees, measured using the Gini coefficient (*mean decrease in impurity*).

Figure 3. Ten most significant variables for the rural context.



Source: Predictive model implemented in Python

Figure 4. Ten most significant variables for the urban context.



Source: Predictive model implemented in Python

Inferential analysis by rural and urban context

To complement the results of the predictive model and analyze in detail the structural differences between the two contexts, a comparative inferential analysis was conducted between students from rural and urban areas. This analysis included categorical and

numerical variables to determine whether the observed differences between the two groups were statistically significant.

Categorical variables

The results obtained show that several student variables present statistically significant differences between rural and urban contexts. Among them is the variable type of high school, which showed a statistically significant difference (χ^2 (1) = 575.55, $p < 0.001$).

Likewise, important differences were identified in variables associated with the family environment, such as living with the parents (χ^2(1) = 4.79, $p = 0.029$), the educational level of the mother ($\chi^2(3) = 22.45$, $p < 0.001$) and the educational level of the father ($\chi^2(3) = 34.96$, $p < 0.001$).

Significant differences were also identified in the occupations of the parents: mother's profession (χ^2 (4) = 22.86, $p < 0.001$), father's profession (χ^2 (4) = 27.97, $p < 0.001$), as well as in variables associated with travel times and academic habits, such as the time to get to school (χ^2 (3) = 10.09, $p = 0.018$) and the time spent studying (χ^2 (3) = 10.36, $p = 0.016$).

Likewise, variables such as the frequency with which students go out with friends (χ^2(4) = 14.19, $p = 0.007$) and alcohol consumption during the week ($\chi^2(4) = 23.70$, $p < 0.001$) and during the weekend ($\chi^2(5) = 19.37$, $p = 0.002$) showed evident differences between contexts.

In contrast, variables such as sex, school support, family support, private lessons, intention to study for a bachelor's degree, health, and extracurricular activities did not show significant differences between the two groups.

Numerical variables

For the numerical variables, the results show statistically significant differences only in age ($U = 17{,}884.00$, $p < 0.001$, $r = 0.296$), with slightly higher values in students from rural areas. The effect size corresponds to Rosenthal's coefficient, indicating a moderate effect size.

Furthermore, neither the number of school absences ($U = 16{,}086.50$, $p = 0.923$) nor the grade of the first semester ($U = 21{,}696.00$, $p = 0.262$) showed statistically significant differences between both contexts.

General interpretation

Overall, the inferential analysis reveals statistically significant differences between rural and urban contexts in family, educational, and behavioral variables. However, variables directly associated with initial academic performance, such as absences and semester grades, did not show statistically significant differences.

Discussion

The following section presents a comparative analysis with previous studies that address the prediction of academic achievement. It also highlights the specific contribution of this study within the educational context of the state of Puebla.

The results obtained through the random forest model show an approximate accuracy of 0.72 to 0.73 in both the rural and urban contexts, which is consistent with the results reported in previous research (Cortez & Silva, 2008; Gil-Vera & Quintero, 2021) . These findings suggest the viability of using this type of model as a support tool in predicting academic achievement in educational contexts.

However, this study provides an analysis that allows for the comparison of students from rural and urban contexts in Puebla and the identification of variables that have a greater influence in each context. In the rural context, attendance, family support, and the number of failed subjects show a more significant relevance, while in the urban context, school absences, the number of failed subjects, and the reason for entering high school appear as factors with greater weight, according to the results presented in figures 3 and 4.

Furthermore, the model's performance was not evaluated solely through accuracy, but also incorporates complementary metrics such as sensitivity and *the F1-score* , in order to analyze the balance between the ability to identify at-risk students and the precision of the assessments. In particular, the macro *F1-score* (≈ 0.71 for both contexts) indicates a suitable balance between sensitivity and precision, which facilitates the timely identification of students at potential academic risk.

Although difficulties were identified in classifying between “Medium” and “High” levels, particularly in urban contexts where there was less sensitivity to the “Medium” category, these results allow for the identification of areas for improvement in the model. In particular, the confusion between the two categories suggests the need to refine the criteria for

separating performance levels and provides a basis for designing individualized interventions aimed at reducing school dropout rates and educational inequality.

In the context of rural high schools, the results show that students with no failing grades demonstrate better academic performance. Likewise, absences appear to have a negative effect on their academic achievement. Another relevant factor is the time available for studying and the commute; when students have little time or must travel long distances, their academic performance tends to be lower. These conditions could influence their energy levels, punctuality, and ability to complete schoolwork.

Furthermore, the family environment has a significant influence on students, particularly in relation to their primary tutor. When the father acts as the tutor, greater stability and family support are observed in the student's education. Given the importance of family relationships, this interaction could be linked to a better emotional state among students, which would benefit their academic performance. Additionally, another relevant factor is peer interaction; the results suggest that less social interaction or a lack of interpersonal skills is associated with lower academic achievement.

On the other hand, in the urban context, students' time constraints are again evident, as many report a location close to the high school and a short commute. This indicates that time is a very valuable resource for students and likely for their families. When time is invested in studying, a positive influence on academic performance is observed. Furthermore, students' age is also related to their academic performance; it is hypothesized that those with above-average ages may experience difficulties integrating into the group, which could affect their academic achievement.

Furthermore, it is again observed that failing subjects and school absences negatively impact academic performance. However, unlike in rural contexts, social interaction does not appear to be a relevant factor, while the parents' occupations play a significant role. Finally, private tutoring emerges as an important aspect; this phenomenon could be explained by the fact that urban classrooms tend to have a larger number of students, which could limit individualized attention from the teacher. In contrast, in rural high schools, the use of private tutors tends to be less frequent.

As shown in Figures 3 and 4, these variables most relevant by context were significantly associated with academic achievement ($X^2, p < 0.005$), highlighting their relevance in both rural and urban areas.

Analysis of the local context of Puebla

The results obtained in this study identify some key points in the socio-educational landscape of Puebla, which differ from those found in previous international studies. In particular, variables such as class attendance, family support, the number of failed subjects, and student age were detected, which could be influenced by structural and cultural aspects specific to this region.

In rural areas, given the significant importance of attendance and family support, it can be inferred that school attendance depends on several factors specific to the region. Many small communities in Puebla face deficiencies related to infrastructure and access to educational resources. Furthermore, transportation difficulties significantly impact punctuality and regularity of school attendance.

On the other hand, based on the model's results, which show that paternal support and a supportive family environment play an important role for students in rural areas, it is suggested that these conditions can help mitigate structural inequalities. Consequently, strategies are needed that integrate social support programs and community support specifically for rural areas.

In contrast, in the urban context, it can be observed that the variables that play an important role are the following: school absences; choice of high school due to its proximity; hours dedicated to studying; and age, reflecting a trend more linked to personal and structural habits and decisions.

It is suggested that age could be related to situations of repeating courses or academic lag, due to economic and social conditions present in urban contexts of Puebla, which cause students to interrupt their academic development.

Similarly, the lack of private tutoring and long commute times could be related to urban variables, such as geographic distribution and the limited availability of additional support materials. These conditions require specialized attention, both to promote educational equity and, potentially, to improve academic achievement.

The results obtained show that, unlike what some international studies usually consider, in Puebla it is necessary and fundamental to take into account the diversity present in the different rural and urban regions, with the aim of evaluating academic achievement more accurately.

The influence of socio-familial variables in rural environments, compared to personal variables in urban contexts, suggests that educational plans and intervention strategies should

be tailored to each context in order to address the specific socio-economic and cultural situation of the state of Puebla. This is based on findings that show that, in rural areas, factors such as school attendance and family support are key determinants of academic achievement, while in urban contexts, variables related to study habits, study time, and students' age predominate.

Taken together, these findings suggest that this local-level analysis not only serves as a complementary tool, but also demonstrates the validity and importance of predictive models for making well-founded and supported educational decisions.

Recommendations

Based on the results presented, it is recommended to implement actions that promote school retention and strengthen teacher support in rural and urban contexts through specialized training and adequate resources (Hernández et al., 2014; Cedillo-Arce et al., 2024). It is also suggested that educational staff be trained in the use of predictive models that allow them to identify not only at-risk students but also high-achieving students, with the aim of providing them with timely support. To this end, the constant monitoring of the most significant variables in each context is vital for the detection and implementation of effective educational strategies.

Specifically, for rural high schools, it is recommended to implement community initiatives that help shorten travel times to school, such as school bus routes or carpooling, to reduce the effort required to travel and improve school attendance. It is also suggested that families be encouraged to participate through workshops and guidance programs for parents and guardians, with the aim of strengthening family support and promoting students' academic success.

For upper secondary education in urban areas, it is recommended to implement individualized tutoring for at-risk students, especially those without the resources for private tutoring, to monitor academic progress and support their success. Additionally, it is suggested to promote independent study habits through the use of online platforms and educational applications, allowing students to reinforce their knowledge at their convenience, thus increasing learning effectiveness and retention of course content.

Specific limitations of the study in the context of Puebla

Although this study provides significant findings on predicting academic achievement in high school students in Puebla, it is important to point out some specific limitations that may affect the interpretation of the results and restrict its exploration to the entire student population of the state.

First, due to limited administrative resources, the study was confined to two geographic areas within the state: the city of Puebla for the urban context and the Tehuacán region for the rural context, using non-probability convenience sampling. This limitation means that the results are not directly generalizable to other regions of the state with socioeconomic and cultural characteristics different from those considered in this analysis.

Similarly, since the information was obtained through a self-administered questionnaire, the possibility that some students answered in a socially desirable manner, particularly regarding family support, coexistence, and study habits, cannot be ruled out, which could introduce bias into the responses. Furthermore, self-selection bias may exist, as students with greater academic interest were likely more willing to participate in the study.

Therefore, these limitations should be considered when interpreting the results, as socially desirable responses and self-selection bias could influence the model's performance, possibly inflating metrics such as accuracy.

Another significant aspect is the disproportion between the sizes of the rural and urban samples, which could impact the model's performance and the comparison between the two contexts. Furthermore, some variables, such as academic achievement levels, are determined by institutional criteria that may vary between schools, affecting the compatibility of the results and limiting their external validity.

On the other hand, it should be mentioned that certain factors, such as technological resources available to households and schools; the quality of teaching; and detailed information on socioeconomic conditions, were not considered within the analyzed data set, which could be fundamental to strengthening the prediction model of academic achievement, particularly in rural areas, where these elements tend to play a determining role in access to educational opportunities and in reducing inequalities.

Finally, although the random forest model achieved stable performance, the lower accuracy observed in classifying the “Medium” and “High” classes, particularly due to confounding between these two categories, suggests that the model can be improved through methodological adjustments. These could include incorporating variables not considered in

this study, adjusting hyperparameters , balancing classes, or implementing a longitudinal analysis approach, with the aim of increasing the effectiveness in identifying students at academic risk.

In conclusion, these delimitations can be interpreted as lines of future research that consider all geographical areas of the entity, integrate multiple and varied variables, and incorporate longitudinal monitoring methods, as well as complementary qualitative approaches, to analyze in greater depth the evolution of academic achievement.

Conclusions

The results suggest that the random forest model achieves balanced and consistent academic performance in both contexts when analyzing sensitivity by class. In the rural context, the model correctly identifies students in the "High" class at a rate of 0.76, while in the urban context this rate reaches 0.85. In contrast, the correct identification of the "Medium" class is lower, with values of 0.67 in the rural context and 0.58 in the urban context.

These results indicate that the model is more effective when it recognizes students with high academic achievement, although it has greater difficulties in differentiating those belonging to the "Medium" class in both contexts, which suggests the need to strengthen the discrimination between these intermediate performance categories.

The balance between the metrics suggests that the model is suitable in educational contexts where early identification of both types of achievement is a priority.

In the rural context, according to the normalized confusion matrix, there is a tendency to confuse the "Medium" class with the "High" class, with a proportion of 0.33 of the cases, and to a lesser extent to confuse the "High" class with the "Medium" class, with a value of 0.24.

In the urban context, this confusion is more pronounced for the “Middle” class, which is classified as “High” in 0.42 of the cases, while the “High” class is confused as “Middle” in a proportion of 0.15.

In general, the random forests model It shows adequate academic performance, although it can be improved, especially in the recognition of the "Medium" grade level in both contexts, where sensitivity reaches values of 0.67 in the rural context and 0.58 in the urban context. For the rural context, the model identifies some key variables for predicting academic performance, such as history of failing grades, age, absences, and family

environment. Likewise, variables related to study time, social interaction, and the commute to school also play an important role.

This information constitutes a useful tool for designing and implementing educational prevention programs tailored to respond to the cultural and socioeconomic differences of rural and urban contexts in Puebla (Pérez Pérez et al., (2025), with the purpose of improving attention to their specific demands and promoting the academic achievement of students.

For students in rural areas, support strategies focused on addressing transportation needs and access to technological tools and educational materials are recommended. In urban areas, however, individual support for students classified as "Medium" risk, according to the predictive model, becomes more relevant.

Similarly, the importance of integrating predictive models based on data mining into local educational institutions is recognized as a technological tool to support planning and monitoring processes. These models also contribute to the effective development of educational plans and prevention strategies, strengthening evidence-based decision-making.

These predictive tools aim to support the implementation of actions designed to reduce high school dropout rates, particularly in rural areas where marked socioeconomic disparities exist. This predictive model helps to highlight educational gaps and justify the implementation of both general and differentiated actions, since rural students have little or no access to technological resources, school transportation, extracurricular activities, and support sessions to address their learning gaps.

Educational and governmental bodies can benefit from using data mining to support the development of targeted support initiatives, channel resources equitably, and adopt measures for the distribution of educational resources among regions.

This research contributes to the early identification of factors associated with low academic performance in the first semesters, a crucial stage for continued education. It also provides an objective basis for guiding tutoring, academic support, and resource allocation through the use of data mining strategies.

Future lines of research

The results obtained in this study highlight the need for a more thorough analysis of aspects such as variable selection, classification algorithms, and model validation processes, with the aim of optimizing prediction models and designing more effective actions to prevent school dropout. In future work, to delve deeper into the factors with the greatest impact, it will be important to review feature selection techniques . *selection*) that allow for optimizing their identification and measuring their influence on academic achievement more accurately.

Later, it is proposed to implement longitudinal data models to periodically monitor the evolution of academic achievement and to assess in advance the risk of school dropout.

Likewise, the aim is to measure the impact of the proposed strategies, directed at the variables that the model identifies as most relevant, in order to assess their effectiveness in improving academic performance.

To this end, the development of a digital platform is suggested that provides educational institutions with a tool for the operational implementation of this model, with the aim of promptly locating students at high risk of dropping out of school.

Finally, the study aims to expand to more geographical areas of Mexico in order to evaluate the external validity of the model and determine whether the behaviors observed in high school students in Puebla are consistently manifested in other contexts, or whether there are particularities associated with the geographical area that limit its generalizability.

References

Acosta-Gonzaga, Elizabeth, & Ramirez-Arellano, Aldo. (2020). Estudio comparativo de técnicas de analítica del aprendizaje para predecir el rendimiento académico de los estudiantes de educación superior. *CienciaUAT*, 15(1), 63-74. Epub 22 de diciembre de 2020. https://doi.org/10.29059/cienciauat.v15i1.1392

Aguilar-Reyes, J. E., Mejía-Peñafiel, E. F., Morocho-Barrionuevo, T. P., & Velasco Castelo, G.-M. (2025). Estudio del rendimiento académico mediante la comparación de modelos de regresión y árboles de clasificación. *Telos*: *Revista de estudios Interdisciplinarios en ciencias sociales*, *27*(1), 94-115. https://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1317-05702025000100094

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart and Winston. https://archive.org/details/in.ernet.dli.2015.112045/page/n3/mode/2up?utm_source=chatgpt.com

Bandura, A. (2001). *Social cognitive theory: An agentic perspective*. *Annual Review of Psychology*, *52*, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture*. SAGE Publications. https://archive.org/details/reproductionined0000bour

Buschini, J. D. (2023). Niklas Luhmann y la teoría general de los sistemas sociales. En A. A. M. Camou (Coord.), *Cuestiones de teoría social contemporánea* (pp. 443–471). La Plata: Universidad Nacional de La Plata; EDULP. https://www.memoria.fahce.unlp.edu.ar/libros/pm.5846/pm.5846.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Castrillón, O. D., Sarache, W., & Ruiz-Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria*, *13*(1), 93–102. https://doi.org/10.4067/S0718-50062020000100093

Cedillo-Arce, J. M., Beltrán-Abreo, H. M., Saltos-Arce, M. I., & Soriano-Barzola, F. R. (2024). Explorando la minería de datos en la gestión educativa superior: desafíos y

oportunidades en la era digital. *Reincisol, 3*(5), 1368–1385. https://doi.org/10.59282/reincisol.V3(5)1367-1385

Choque-Aguilar, M. R. (2024). Red neuronal para predecir el rendimiento académico. *Revista Simón Rodríguez, 4*(8), 22–35. https://doi.org/10.62319/simonrodriguez.v.4i8.31

CONEVAL. (2023). Informe de pobreza multidimensional 2022: Resultados nacionales y por entidad federativa. Consejo Nacional de Evaluación de la Política de Desarrollo Social. https://www.coneval.org.mx

Contreras-Bravo, L. E., Fuentes-López, H. J., & Rivas-Trujillo, E. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. *Boletín Redipe*, *10*(13) 171–190. https://doi.org/10.36260/rbr.v10i13.1737

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. En A. Brito & J.Teixeira (Eds.), *Proceedings of the 5th Annual Future Business Technology Conference* (pp. 5-12). EUROSIS. https://doi.org/10.24432/C5TG7T

Díaz-Martínez, M. A., Ahumada-Cervantes, M. de los Ángeles, & Melo-Morín, J. P. (2021). Árboles de decisión como metodología para determinar el rendimiento académico en educación superior. *Revista Lasallista de Investigación, 18*(2), 94–104. https://revistas.unilasallista.edu.co/index.php/rldi/article/view/2724

Gil-Vera, V. D., & Quintero-López, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información Tecnológica*, *32*(6), 221–228. https://doi.org/10.4067/s0718-07642021000600221

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6.ª ed.). McGraw-Hill. https://uniclanet.unicla.edu.mx/assets/contenidos/254857_DOC_2023-03-01_18:46:18.pdf

INEGI. (2022). Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH 2022). Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/programas/dutih/

Merceron, A., & Tato, A. (2023). Introduction to neural networks and uses in educational data mining. En M. Feng, T. Käser, & P. Talukdar (Eds.), *Proceedings of the 16th*

International Conference on Educational Data Mining (pp. 578–581). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115774

Ordoñez-Avila, R., Salgado Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon, 9*(3), e13939. https://doi.org/10.1016/j.heliyon.2023.e13939

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Pérez Pérez, A. M., Custodio Valenzuela, M., Cerón Garnica, C., Mila Avendaño, V. M., & Moyao Martínez, Y. (2025). La gestión académica en centros educativos urbanos marginales y zonas rurales orientada a la preparación del docente para la alfabetización inicial. *Edutec*. *Revista Electrónica de Tecnología Educativa, 33*(1), 1–12. https://doi.org/10.58299/edutec.v33i1.335

Piaget, J. (1972). *Psychology and epistemology: Towards a theory of knowledge.* Penguin. https://books.google.com.mx/books/about/Psychology_and_Epistemology.html?id=7DkdAQAAMAAJ&redir_esc=y

Rico-Páez, Andrés. (2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(24), e044. https://doi.org/10.23913/ride.v12i24.1196

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Secretaría de Educación Pública. (2023). Estadística e indicadores educativos: Puebla, ciclo escolar 2022–2023. Gobierno de México. https://planeacion.sep.gob.mx

Secretaría de Educación Pública. Dirección General de Planeación, Programación y Estadística Educativa. (2024). Estadística educativa. Puebla. Ciclo escolar 2023–2024 [Informe]. https://planeacion.sep.gob.mx/Doc/estadistica_e_indicadores/EstIndEntFed2023/21_PUE.pdf

Taylor, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. En A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, methodologies, and applications* (pp. 379–396). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch16

Macías-Ureta, K. T., & Ordóñez-Valencia, E. V. (2025). Metodologías activas para el desarrollo de habilidades matemáticas: Un análisis bibliográfico. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, 6*(2), 3431–3450. https://doi.org/10.56712/latam.v6i2.3917

Vygotsky, L. S. (1978). Mind in society: *The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Eds. & Trans.). Harvard University Press. https://autismusberatung.info/wp-content/uploads/2023/09/Vygotsky-Mind-in-society.pdf

Contribution Role	Author(s)
Conceptualization	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Methodology	Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, a supporting role
Software	Programming, software development; Computer program design; Implementation of computer code and supporting algorithms; Testing of existing code components. Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, a supporting role
Validation	Verification, either as part of the activity or separately, of the full replication/reproducibility of the results/experiments and other research products. Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Formal Analysis	Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, same role
Investigation	Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, a supporting role
Resources	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Data curation	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Writing - Preparing the original draft	Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, a supporting role
Writing - Reviewing and Editing	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Display	Dr. Yolanda Moya Martínez, main role Dr. Carmen Cerón Garnica, a supporting role
Supervision	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Project Management	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role
Acquisition of funds	Dr. Yolanda Moya Martínez, equal role Dr. Carmen Cerón Garnica, same role

Appendix

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

Questionnaire applied to rural and urban students

1. Choose a high school
2. Sex
3. Age
4. Address (Rural or Urban)
5. Number of members in your family
6. Parental cohabitation status
7. Mother's educational level
8. Father's educational level
9. Mother's profession
10. Father's Profession
11. Why did you choose this high school?
12. Who is the tutor
13. Time it takes you to get to school
14. Weekly time spent studying
15. Number of failed subjects
16. Do you receive any academic support?
17. Do you receive any family support?
18. You receive private lessons
19. You participate in extracurricular activities
20. You attended daycare
21. You intend to study for a bachelor's degree
22. You have internet access at home
23. You are in a romantic relationship
24. Quality of family relationships
25. Amount of weekly free time
26. How often do you go out with friends?
27. Alcohol consumption during the week
28. Alcohol consumption during the weekend
29. Health status
30. Number of school absences
31. First semester grade