International Journal of Combinatorial Optimization Problems and Informatics, 15(5), Dec 2024, 155-160. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2024.v15i5.573

www.editada.org

Early Detection of Students at High Risk of Academic Failure using Artificial Intelligence

Antonio Álvarez Núñez, María del Carmen Santiago Díaz, Ana Claudia Zenteno Vázquez, Judith Pérez Marcial, Gustavo Trinidad Rubín Linares

Benemérita Universidad Autónoma de Puebla. Facultad de Ciencias de la Computación, Av. 14 Sur y San Claudio, Col. San Manuel, CP 72570, Puebla, Pue. México
antonio.alvarez@alumno.buap.mx, {marycarmen.santiago, ana.zenteno, judith.perez, gustavo.rubin}@correo.buap.mx.

Abstract. The academic performance of students in Mexico has a great impact on the social and economic development of the country. Early detection of students at academic risk is necessary to improve educational quality and reduce school dropouts. This work presents a proposal that uses a predictive model based on Logistic Regression to identify students at high risk of academic failure and its usefulness to provide proactive and personalized support to those who need it. In addition, an overview of the impact of Artificial Intelligence and Machine Learning in education is presented, especially in predicting student dropout and supporting academic performance, allowing us to take an important step towards a more promising and successful educational future for students. students.

Keywords: Failure, School Dropout, Academic Performance, Artificial Intelligence, Logistic Regression.

Article Info
Received May 10, 2024.
Accepted Nov 20, 2024.

1 Introduction

Education in Mexico is in a constant process of change and restructuring, seeking to respond to social demands, correct social inequalities, train subjects prepared for a globalized world and meet the requirements of the Organization for Economic Growth and Development (OECD) [1]. Therefore, academic performance is of interest to both institutions and the government, due to its effects on social and economic development. Understanding academic performance in educational institutions has allowed us to demonstrate the real production of students in formal activities and requires the analysis of their forms of expression, in terms of student delay, desertion and low rates of effectiveness, being considered components of school failure, so it is a constant challenge to maintain and improve academic quality. As a consequence, the experiences of school failure in individuals constitute a deprivation of the development of cognitive, personal and social capacities, which limit opportunities for material and symbolic goods, restrict the aspirations of individuals and the real possibilities of satisfying them [2]. Based on this, the opportunity is perceived to detect early students who present indicators of academic risk and thus increase the effectiveness and significance in the intention to support students in a preventive manner [3].

The educational panorama today faces significant challenges in terms of student retention and academic performance. Academic failure and school dropout continue to be obstacles that affect millions of young people around the world, limiting their future opportunities and generating a gap in access to quality and equitable education. However, in this context, a ray of hope emerges through the development and application of predictive models.

The digital revolution has allowed analytical research to be carried out thanks to the ease of capturing, processing, storing, distributing and transmitting digitalized information generated by new societies to discover interesting models and solve everyday problems in society [4]. In recent years, Artificial Intelligence (AI) techniques such as Machine Learning (ML) and Deep Learning (DL) have positively impacted the advancement of different fields of knowledge, including education [5]. Education plays a fundamental role in the development of all societies, as it enables people to increase their productivity and address problems more efficiently, often taking advantage of creative approaches. In the educational field, Machine Learning (ML) techniques have been used for various activities, including predicting student dropout and supporting student academic performance. Likewise, Artificial Intelligence allows teachers to analyze an individualized follow-up of students' actions in learning environments and thus detect in advance which students need intervention or specialized help [6].

Although AI applications in education are still in an early stage, there are numerous examples that show its potential to transform this field. In relation to the educational field, AI can accelerate personalized learning and provide support to students with special needs. At the educational system level, promising uses are emerging such as predictive analytics to reduce school dropouts and the assessment of new sets of required skills.

AI and digitalization have also generated a new demand for complex skills, which are less susceptible to automation. These skills include higher cognitive aspects such as creativity and critical thinking.

1.1 Tools within the field of computer science and data science.

Artificial Intelligence (AI) is a field of study that focuses on developing computational systems capable of interacting with the environment through skills such as visual perception and voice recognition, as well as intelligent behaviors such as processing and selecting information for decision-making with specific objectives. In parallel, data mining is a key tool for understanding student behavior and generating effective educational models. By describing, predicting, and generalizing data collected in the educational context, relevant patterns and trends can be identified. From these observations, learning activities can be suggested, notifications about student performance can be sent, and group work can be encouraged to improve teaching-learning. The data comes from various sources

2 Theoretical Framework and State of the Art

2.1 Artificial Intelligence in Education

AI systems are designed to operate with different levels of autonomy. The phases of the AI system lifecycle consist of [12]:

- Planning, design, data collection, processing, model creation and interpretation
- Verification and validation.
- Deployment.
- Operation and monitoring.

The algorithmic power of AI is also used to create predictive and diagnostic models to support decisions and generate feedback at the establishment (school, university, etc.) or educational system (district, region, country, etc.) level [12]. Many of the AI applications are models or profiles of learners that allow prediction, for example, the likelihood of a learner dropping out of a course or being admitted to a program, in order to offer timely support or provide feedback and guidance on content-related matters throughout the learning process [13].

Artificial Intelligence offers several advantages for detecting and preventing failure rates in schools [14], such as:

- Early identification of low performance patterns.

- Personalization of teaching and support for each student.
- Accurate feedback and recommendations to improve learning.

2.2 Successful implementations of AI in preventing failure.

Several success stories have been identified worldwide, for example, in Finland, a project has been implemented that uses AI to detect early students at risk of failing. The system collects and analyzes academic and behavioral data to identify patterns that indicate learning difficulties. Another notable case is Singapore, where an AI-based virtual tutoring system has been implemented that provides individualized educational support to students. The system analyzes the progress of each student, identifies areas of difficulty, and provides learning resources tailored to their specific needs [17].

In Mexico, the Tec. de Monterrey has implemented an adaptive learning platform based on AI ALEKS, teachers have presented an initiative that proposes the use of an adaptive learning platform as a leveling method in mathematical skills, for evaluation and online learning. The adaptive learning platform is designed to personalize the learning experience of each student, identifying their areas of strength and weakness in mathematics [20].

Another project that has been developed is an Early Warning System (SisAT) in a Mexican secondary school that aims to detect students at risk of dropping out in time and provide appropriate interventions to prevent it. Important points of SisAT include the analysis of longitudinal data on academic performance and other relevant factors, identification of exact indicators to predict dropout, generation of alerts for timely interventions, offering personalized support, constant monitoring of progress and collaboration between teachers, administrators, parents and educational policy makers [21].

2.3 Problem statement.

Education is one of the factors that most influences the progress of people and societies, in addition to providing knowledge, enriching culture, spirit, values and everything that characterizes us as human beings [6].

School dropout is a complex phenomenon that can be attributed to various causes, among which are the following:

- Economic difficulties.
- Need to work and study.
- Poor academic performance.
- Failing subjects.

3 Methodology

The methodology was developed in general in 4 phases, each of the phases is subdivided into different tasks as shown in Figure 1.

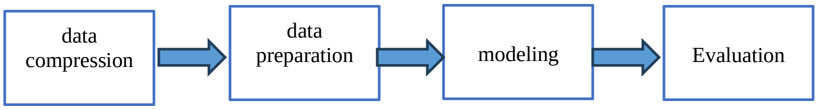

Figure 1. Methodology.

Phase 1: Understanding the data: In this initial phase, as shown in Figure 1, we sought to obtain an overview of the structure and content of the data, ensuring that relevant information about the students, their grades in different subjects, identification data, and other academic and previous performance variables was captured. A careful review was carried out to verify that there was no missing, incoherent, or erroneous data. Likewise, possible biases or imbalances in the data set that could affect the accuracy of the model were identified. Once the data was obtained, it was described using statistical and visualization techniques. The size of the data set, the number of records and characteristics present, as well as the distribution of grades and other variables were analyzed.

Phase 2: Data preparation: In this phase, a thorough preparation and adaptation of the data was carried out to ensure that it was in the appropriate form and ready to be used in the predictive model. New variables were generated from existing ones, such as grade point average, allowing for a more complete perspective of students' academic performance.

Phase 3: Modeling: First, a careful selection of the most appropriate modeling techniques was carried out, taking into account the specific characteristics of the problem and the available data. After a thorough analysis, Logistic Regression was chosen as the core algorithm of the model due to its distinctive features and its ability to deal with non-linear data, which was relevant in our educational context. Furthermore, its highly interpretable nature provides confidence in the results obtained and allows educators and decision-makers to design personalized support strategies for those students at risk.

Using the training set previously prepared in the Data Preparation phase, the model adjusted its parameters during the training process to find the best relationship between the variables and the labels of the students who pass or fail. This fine-tuning process allowed the model to learn from historical data, acquiring the knowledge necessary to make accurate predictions on new and unseen data, based on their corresponding identification data and grades obtained in different subjects, such as mathematics, computer science, English, and philosophy. To ensure the effectiveness of our model, we split the data into training and test sets, preprocessing them using feature scaling to obtain more robust and accurate results. We trained the model with the training set and evaluated its performance using the test set, defining key metrics such as Accuracy, Precision, Recall, and F1 Score for a rigorous assessment of its prediction ability. These metrics gave us a holistic view of the model's performance by evaluating both the ability to correctly classify positive cases and negative cases, thus ensuring a complete and reliable assessment of its ability to identify students at risk of academic failure.

Phase 4: Evaluation: In this phase, a rigorous evaluation of the model was carried out to determine its ability to detect students at high risk of academic failure early. The results obtained were analyzed and the modeling process was reviewed to ensure the quality and reliability of the model.

The evaluation was based on the previously defined metrics, which allowed measuring key aspects of the model's performance, such as Accuracy, which indicated the proportion of correct predictions over the total data, while Precision measured the proportion of correct positive predictions over the total positive predictions. Recall represented the proportion of positive cases correctly identified over the total positive data, and F1 Score acted as a harmonic mean between Precision and Recall.

4 Results

Below are the results obtained when evaluating the functioning and performance of our predictive model for the early detection of students at high risk of academic failure.

The test results showed that our predictive model has achieved outstanding levels in the key evaluation metrics, the accuracy was 98%, indicating that the predictions made by the model were correct, demonstrating its ability to correctly identify students at risk of failure and minimize false positives. The Recall obtained 100%, indicating that the model is highly effective in detecting all at-risk students among the positive cases present in the data. Likewise, the F1 Score yielded a value of 99%, evidencing a balance between the model's ability to correctly classify positive and negative cases, confirming its performance in the prediction task.

The data used to train and test the model was created by simulating that of an educational institution, with 241 students and the following attributes: firstname, lastname, email, grades in mathematics, computer science, English and philosophy. This data set represented a sample of the student population, which allowed the model to learn complex and generalizable patterns.

The results shown that the model has been highly effective in classifying students who passed and failed. Most students were correctly classified as passed, which reinforces the reliability of the model in this task.

We shows the subjects in which students are having the most difficulties, showing the number of students with low grades in each of them.

From the test results it is possible to identify students at risk of academic failure and those who have been failed. The validation of our predictive model was carried out with a set of tests that was not used during the training process. The results were highly encouraging, with an accuracy of 98%. This means that 98% of the predictions made by the model were correct. This high accuracy suggests that the model is reliable and capable of making accurate predictions about students at risk.

Furthermore, the accuracy of the model was 98%, indicating that these 98% of students that the model identified as "high risk" of academic failure were actually in that situation. This result is fundamental, as it guarantees that the model is not overestimating the number of students at risk and will allow educational institutions to concentrate their efforts and resources on those students who really need it.

Recall and F1 Score: The model's Recall was 100%, meaning that it did not miss any students at risk of academic failure. In other words, the model did not make any Type II errors, which involves not detecting an at-risk student when it should have. This high Recall demonstrates the model's sensitivity in detecting at-risk students, which is essential for early detection and implementing timely interventions.

The model's F1 Score was 99%, indicating an excellent ability of the model to balance precision and recall. The F1 Score is used to evaluate a model's performance in situations where there is an imbalance between classes (in this case, at-risk students and non-risk students). A score close to 100% suggests that the model strikes an optimal balance between the ability to identify at-risk students and avoiding false positives.

5 Conclusions and future work

The predictive model based on Logistic Regression has proven to be highly effective in the early detection of students at risk of academic failure, with an Accuracy, Recall and F1 Score close to 100%. Its application in educational institutions allows the implementation of early intervention strategies, providing personalized tutoring and allocating additional resources to at-risk students, contributing to a more inclusive and effective education. This tool has proven to be valuable in improving retention and academic success, positively impacting the educational future of students. The use of Machine Learning techniques in the identification of at-risk students represents an important step towards a more effective education oriented towards the success of each student, allowing data-driven decision-making to strengthen the educational process.

For future research, other machine learning algorithms could be explored and the inclusion of other factors that influence academic performance could be considered. Although the model is promising, it is essential to recognize that it is not a definitive solution, but rather a valuable tool to support educators and counselors in making informed decisions. With a continued focus on improvement and research, we hope that our model will contribute to a stronger and more equitable education system in the future.

References

1. Noyola González, A., & Espinoza Guajardo, J. R. (2012). La relación entre el modelo de evaluación y los indicadores de desempeño académico en los alumnos de bachillerato. *Universidad Virtual. Escuela de Graduados en Educación*.
2. Murillo García, O. L., & Luna, S. E. (2021). El contexto académico de estudiantes universitarios en condición de rezago por reprobación. *Universia, 12*(33).
3. Ferrari Carlevari, M., & Candia, C. (2021). Panel de Monitoreo del Desempeño Académico. Herramienta para la gestión eficiente y preventiva del riesgo académico. *Universidad del Desarrollo*.

4. Chancusing Taipicaña, D. M. (2020). Modelo de análisis del rendimiento académico de la Unidad Educativa Personas con Escolaridad Inconclusa (P.C.E.I) "Monseñor Leonidas Proaño" del cantón Latacunga, a través de minería de datos. *Universidad Técnica de Cotopaxi*.
5. Cruz, E., González, M., & Rangel, C. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico, 13*(1), 77–87.
6. Avila Grajeola, V. S. (2022). Modelo de Learning Analytics para predecir rendimiento en alumnos con datos escolares. *División de Estudios de Posgrado e Investigación*.
7. Chávez Ramírez, M. R. (2021). El papel de la inteligencia artificial en la educación superior. *STEM, 24–30*.
8. Orea, S., Salvador Vargas, A., & García Alonso, M. (2005). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los K vecinos más cercanos. *Ene, 779*(73), 33.
9. McCaffrey, P. (2020). Introduction to machine learning: Regression, classification, and important concepts. En *An Introduction to Healthcare Informatics* (pp. 191–210).
10. Murphy, M. J., & El Naqa, I. (2015). What is machine learning? En *Machine Learning in Radiation Oncology* (pp. 3–11). Springer.
11. Lancrin, V. (2022). Smart education technology: How it might transform teaching (and learning). *New England Journal of Public Policy, 34*.
12. Vicent-Lancrin, S., & Van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD*.
13. Zawacki-Richter, O., Marin, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27.
14. González, R. (2021). Advantages of AI in detection and prevention of academic failure. *International Journal of Artificial Intelligence in Education, 15*(4), 78–95.
15. Ballesteros Román, A., Sánchez-Guzmán, D., & García, R. (2014). Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. *Latin American Journal of Physics Education, 7*(4).
16. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116.
17. Jones, E. (2020). Successful implementation of AI in academic failure prevention: Global case studies. *Journal of Educational Technology, 12*(2), 89–105.
18. Brundage, M., Avin, S., & Clark, J. (2016). Artificial intelligence and life in 2030: The one hundred year study on artificial intelligence. *arXiv preprint*.
19. Castrillón, O. D., Sarache, W., & Ruiz Herrera, S. (2020). Predicción del rendimiento académico por medio de técnicas de inteligencia artificial. *Formación Universitaria, 13*, 93–102.
20. Hernández, D. (2020, November 4). Profesores crean plataforma de aprendizaje con inteligencia. *Conecta*. Recuperado de https://conecta.tec.mx/es/noticias/aguascalientes/educacion/profesores-crean-plataforma-de-aprendizaje-con-inteligencia.
21. Martínez Ponce, J. C., & Correa Carrera, H. (2020). El Sistema de Alerta Temprana (SISAT) para disminuir el abandono escolar en las Escuelas Primarias Rurales de Tabasco. *Perspectivas Docentes, 31*(73), 23–30.