Predictive model based on artificial intelligence and contextual awareness to identify students at risk of dropping out of university in Panama

Modelo predictivo basado en inteligencia artificial y conciencia de contexto para identificar estudiantes en riesgo de deserción universitaria en Panamá

Um modelo preditivo baseado em inteligência artificial e consciência de contexto para identificar estudantes em risco de abandonar a universidade no Panamá.

Laury Arenales [1]
Vladimir Villarreal [2]
Juan Jose Saldana-Barrios [3]

Received: October 10th, 2024
Accepted: December 15th, 2024
Available: January 20 th, 2025

How to cite this article:
L. Arenales, V. Villarreal y J. J. Saldana-Barrios, "Predictive model based on artificial intelligence and contextual awareness to identify students at risk of dropping out of university in Panama," *Revista Ingeniería Solidaria,* vol. 21, no. 1, 2025.
doi: https://doi.org/10.16925/2357-6014.2025.01.08

Research article. https://doi.org/10.16925/2357-6014.2025.01.08

[1] Facultad de Ingeniería de Sistemas Computacionales. Universidad Tecnológica de Panamá, Panamá, Panamá.
Email: laury.arenales@utp.ac.pa
ORCID: https://orcid.org/0000-0001-7408-2107

[2] Facultad de Ingeniería de Sistemas Computacionales. Universidad Tecnológica de Panamá, Panamá, Panamá.
Email: vladimir.villarreal@utp.ac.pa
ORCID: https://orcid.org/0000-0003-4678-5977

[3] Facultad de Ingeniería de Sistemas Computacionales. Universidad Tecnológica de Panamá, Panamá, Panamá.
Email: juan.saldana@utp.ac.pa
ORCID: https://orcid.org/0000-0001-8119-4000

Abstract

Introduction: This article presents the findings from the research on the "Analysis, design, and development of a context-aware intelligent system for university dropout prediction using a microservices architecture" conducted at the Technological University of Panama in 2024. The study focuses on creating a predictive model using artificial intelligence that integrates students' academic, sociodemographic, and psychological context to identify students at risk of early dropout.

Problem: In Latin America, the educational system struggles with high dropout rates, particularly at the university level. The early years of university education are crucial, and dropping out during this period significantly impacts a country's social, labor, and economic development.

Objective: The main goal of this research was to develop a context-aware predictive model using artificial intelligence to identify students at risk of dropping out at the Technological University of Panama. The model incorporates academic, socioeconomic, and psychological factors to predict dropout risks more accurately.

Methodology: The study follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, starting with problem understanding and followed by the collection and processing of students' academic, socioeconomic, and psychological data. Three machine learning models were implemented and evaluated: Logistic Regression, Random Forest, and Gradient Boosting. Each model was tested under different scenarios to identify the most effective one.

Results: The Random Forest model in Scenario 3 demonstrated the best performance, offering a strong balance between accuracy and generalization. This model was the most effective for predicting university dropouts based on the collected data.

Conclusion: The developed predictive model is a relevant and innovative tool that can help the Technological University of Panama identify students at risk of dropping out early. By using artificial intelligence and a context-aware approach, this tool can contribute to reducing dropout rates and improve student retention.

Originality: This research introduces a novel approach to university dropout prediction by integrating context-awareness with artificial intelligence. The multidimensional nature of the model, considering academic, sociodemographic, and psychological factors, sets it apart from traditional predictive models.

Limitations: One of the key limitations of this research was the restricted availability of real psychological data, which limited the comprehensiveness of the model.

Keywords: dropout, artificial intelligence, predictive model, context, university, student.

Resumen

Introducción: Este artículo es el resultado de la investigación *Análisis, diseño y desarrollo de un sistema inteligente consciente del contexto para la predicción de deserción universitaria, utilizando una arquitectura de microservicios*, desarrollada en la Universidad Tecnológica de Panamá en 2024. Presenta el proceso de desarrollo de un modelo predictivo basado en inteligencia artificial que incorpora el contexto académico, sociodemográfico y psicológico de los estudiantes para la identificación temprana de aquellos en riesgo de deserción.

Problema: El Sistema educativo en la región Latinoamericana enfrenta altos indices de deserción, especialmente a nivel universitario, siendo los primeros años de la carrera determinantes, lo que repercute negativamente al desarrollo social, laboral y financiero de un país.

Objetivo: Desarrollar un modelo predictivo consciente del contexto basado en inteligencia artificial para identificar de manera oportuna a estudiantes en riesgo de deserción en la Universidad Tecnológica de Panamá.

Metodología: Se sigue la metodología de mineria de datos CRISP-DM, la cual comprende el problema, luego se recolectan y procesan los datos academicos, socioeconomicos y pscicologicos de los estudiantes. Se imple-

mentan tres modelo de machine learning (Regresión Logistica, Random Forest y Gradient Boosting) evaluados en diferentes escenarios.

Resultados: En los resultados obtenidos de los tres modelos de machine learning, se evidencia que el modelo de Random Forest en el Escenario 3 tiene un equilibrio optimo entre precisión y generalización.

Conclusión: Este modelo predictivo es una herramienta relevante y pionera en la identificación temprana de estudiantes en riesgo de deserrción de la UTP.

Originalidad: Esta investigación introduce un enfoque multidimensional e innovador a la predicción de la deserción universitaria al integrar la consciencia de contexto e inteligencia artificial.

Limitaciones: la disponibilidad de datos psicologicos reales para la investigación es limitada y restringida.

Palabras claves: deserción, inteligencia artificial, modelo predictivo, contexto, universidad, estudiante.

Resumo

Introdução: Este artigo é o resultado da análise, projeto e desenvolvimento de um sistema inteligente sensível ao contexto para prever taxas de evasão universitária, utilizando uma arquitetura de microsserviços, desenvolvido na Universidade Tecnológica do Panamá em 2024. Apresenta o processo de desenvolvimento de um modelo preditivo baseado em inteligência artificial que incorpora o contexto acadêmico, sociodemográfico e psicológico dos alunos para a identificação precoce daqueles em risco de evasão.

Problema: O sistema educacional na América Latina enfrenta altas taxas de evasão, especialmente no nível universitário. Os primeiros anos da faculdade são cruciais, impactando negativamente o desenvolvimento social, trabalhista e financeiro de um país.

Objetivo: Desenvolver um modelo preditivo sensível ao contexto baseado em inteligência artificial para identificar prontamente alunos em risco de evasão na Universidade Tecnológica do Panamá. Metodologia: A metodologia de mineração de dados CRISP-DM, que abrange o problema, é seguida. Dados acadêmicos, socioeconômicos e psicológicos dos alunos são então coletados e processados. Três modelos de aprendizado de máquina (Regressão Logística, Floresta Aleatória e Reforço de Gradiente) são implementados e avaliados em diferentes cenários.

Resultados: Os resultados obtidos com os três modelos de aprendizado de máquina mostram que o modelo Floresta Aleatória no Cenário 3 atinge um equilíbrio ideal entre precisão e generalização.

Conclusão: Este modelo preditivo é uma ferramenta relevante e pioneira para a identificação precoce de alunos em risco de abandono do Ensino Médio (UTP).

Originalidade: Esta pesquisa apresenta uma abordagem multidimensional e inovadora para a previsão da evasão universitária, integrando a consciência contextual e a inteligência artificial.

Limitações: A disponibilidade de dados psicológicos reais para pesquisa é limitada e restrita.

Palavras-chave: evasão, inteligência artificial, modelo preditivo, contexto, universidade, aluno.

1. INTRODUCTION

The withdrawal of a student from a formal educational system before obtaining the corresponding final degree is known as dropout or educational abandonment, whether at the primary, secondary, or university level [1], [2].

From a more quantifiable perspective, dropout is defined as the quantitative comparison between initial enrollment and the number of students graduating in the final year [3]. According to the National Ministry of Education, a student is categorized as having dropped out of university when they have not enrolled for two or more consecutive academic periods (semesters) in their initially registered program and have not graduated [2].

University dropout is a complex issue that has garnered significant attention in the educational sector worldwide due to its repercussions on a country's academic, social, and professional spheres [4], [5]. The World Bank [6] highlights that the loss of assets in education could lead to a decrease in income by up to $1.7 trillion in the Latin American region, accompanied by a 15% increase in dropout rates. In 2018, Panama had the fourth-highest higher education dropout rate in Latin America, with 30% of students abandoning their studies, surpassed only by Bolivia, Colombia, and Ecuador [6].

In Panama, the growing enrollment rate at universities cannot be fully supported by the public and private sectors responsible for addressing this demand, potentially creating obstacles for students to complete their studies. The University of Panama exhibits the highest dropout rate at 14%, followed by the Technological University of Panama with 10%, according to a study conducted by Díaz-Tejedor [7] in 2018. University dropout not only affects the individual but also society, the community, educational institutions, and the labor market, directly impacting the social and economic development of the country.

Aligned with this issue, the question under the framework of the Sustainable Development Goals arises: What measures should be taken to prevent students from being left behind in the face of high university dropout rates? The research question for this study is formulated as follows: Can the implementation of a university dropout prediction model based on artificial intelligence technologies and context-awareness optimize the timely identification of students at risk of dropping out at the Technological University of Panama?

The main objective of this research is to develop a context-aware predictive model based on artificial intelligence to timely identify students at risk of dropping out at the Technological University of Panama.

According to a report by the Inter-American Development Bank [8], which sought to diagnose higher education in Panama, several key premises highlight the importance of developing an intelligent, context-aware university dropout prediction model. Panama lacks sufficient statistics and reliable assessments to properly analyze the outcomes of the current educational system, creating an urgent need for more effective data collection systems to enable informed, evidence-based decision-making [9].

Additionally, there is a need to improve the traditional education system by aligning objectives with new realities, focusing on innovation to ensure that trained human resources can meet the evolving needs of the labor market and society [7].

An innovative aspect of this research is its multidimensional approach to data collection on students, incorporating psychological, economic, demographic, personal, and prior education factors. This comprehensive perspective provides a deeper understanding of the variables influencing academic performance and university dropout risk [10], [11]. It also facilitates the identification of patterns and correlations that might not have been previously considered, enhancing the effectiveness of interventions and educational policies through more accurate predictions [12].

1.1 Theoretical Framework

One of the primary definitions of dropout is provided by [13], who states that "university student dropout should not only be understood as the definitive abandonment of classrooms but also as the abandonment of academic formation, which has serious social repercussions."

It is essential to conceptualize the term context, as it has multiple definitions and can be interpreted from various fields of knowledge [14], [15]. Derived from the Latin term contextus, it refers to everything surrounding an individual, whether tangible or intangible, from which a fact is interpreted or understood. The environment is where the necessary actions are carried out to complete a task or activity, whereas the context is where the relevant elements of the environment interact. "The preparation that equips the context to react appropriately by obtaining the correct information, providing convenient and adapted features" is what is understood as context awareness [16].

Machine Learning (ML) is a sub-discipline of Artificial Intelligence that enables computers to learn from input data to make predictions or generate results, facilitating decision-making based on that data. At its core, ML can be understood as the ability to automatically learn through experience, whether for general tasks or specific purposes [17].

1.2 Background and Related Studies

In the study conducted by Mussida [18] at the Polytechnic University of Milan, an analytical learning tool was developed with the objectives of identifying students at high risk of dropping out of university, providing support to these students, and implementing strategies to mitigate the factors leading to a high risk of dropout within

the system. This tool features a predictive analytics dashboard based on a multilevel linear regression model developed in R, offering university staff and students predictive information about student status and the probable causes of potential university dropout. This tool incorporates both static information (personal data) and dynamic information (exam grades) from students.

On the other hand, Cevallos et al. [19] proposed a university dropout prediction model for mid-level university students, based on economic, personal (lack of adequate vocational guidance), and institutional factors. This model is structured in three stages of the IBM SPSS Modeler methodology and validated through cross-validation. The study aimed to compare the effectiveness of two predictive techniques: Bayesian networks and decision trees within the context of Educational Data Mining. The results showed that Bayesian networks outperformed decision trees across all metrics outlined in the model. An important takeaway from the study is the recommendation for university students to engage in sports to mitigate possible stress factors in student life.

The study presented by Merchan-Rubiano [20] developed a predictive model for academic performance in engineering students. This model included demographic data and entrance exam scores (ICFES) in Colombia. For this study, algorithms such as Random Forest and J48 were employed, highlighting the importance of identifying significant variables to predict first-year performance. A key feature of this research is the identification of problem causes, allowing for preventive strategies to support at-risk students.

In the context of Massive Open Online Courses (MOOCs), Zheng [21] proposed a university dropout prediction model based on fused function, weighting, and time series with FWTS-CNN convolutional neural networks. Unlike previous models, this approach extracts, classifies, and weights behavioral features from the student's learning records according to their importance. This creates a convolutional neural network model based on behavioral time series. Results from this model and the FWTS-CNN methodology showed a high precision rate, outperforming standalone convolutional neural networks.

In the area of context awareness for teaching and learning processes, the field of ubiquitous learning (u-learning) stands out. Within this domain, the systematic mapping of context-aware ubiquitous learning environments by Rabelo [22] provides a detailed description of these environments. These systems can collect, interpret, and use student data to adjust behavior in response to the student's learning needs. Additionally, they provide access to learning resources anytime, anywhere, and on any device [23]. U-learning environments can be classified into various contexts [24]:

computational context (device, network, resources), user context (personal data, social situation), physical context (light, temperature, location), and temporal context (time, date, year).

2. METHODOLOGY AND DEVELOPMENT

The success of the data mining process and the development of a predictive model is determined by a structured methodology that ensures reliable and reproducible results over time. For the development of the predictive model for university dropout prediction, the CRISP-DM data mining methodology [25] was followed. This methodology aims to provide a structured and systematic process for project development in various contexts. It consists of six phases, which are detailed below:

2.1 Phase 1: Business Understanding

The first fundamental stage of this research methodology focuses on understanding the business, which is necessary to align the project's goals to achieve a significant impact. This phase is aimed at comprehending the objectives and requirements of the project. For this project, the business understanding was addressed in Section 1. Introduction of this document.

2.2 Phase 2: Data Understanding

In this phase, a data collection and processing process is carried out to identify and define the key variables that will be implemented in the predictive model. This process is conducted in close collaboration with various departments at the Technological University of Panama, each providing specific data sources related to students' academic information (General Directorate of Information and Communication Technologies - DITIC), psychological aspects (Psychological Guidance Directorate - DOP), and economic well-being (Student Welfare Directorate - DBE). Meticulousness and ethics in handling this data are fundamental principles guiding the entire process, ensuring the integrity and confidentiality of the information. The data collected from the various departments are described below:

2.2.1 Data Collection

General Directorate of Information and Communication Technologies (DITIC): This department provides essential academic and socioeconomic information about the students [26]. This includes data such as academic records, gender, age, socioeconomic level, previous studies, and demographic information. These data are obtained from existing records, either in digital or physical format, and are processed and structured according to the project's needs. The data collected spans the period (2013–2023), totaling 22,615 records.

Student Welfare Directorate (DBE): Collecting current and historical information from the Student Welfare Directorate [27] in each regional center is essential for a comprehensive understanding of students' needs. Data were collected both physically and digitally across the nine regional centers of UTP nationwide. These included data on economic aid related to food, transportation, scholarships, phone recharges, tablets, glasses, and compensatory work. The data collected spans the period (2015–2023), totaling 6,187 records.

Psychological Guidance Directorate (DOP): To collect data from the Psychological Guidance Directorate [28], informed consent was requested from students willing to participate in this study, using random samples nationwide. The data provided by the Directorate included results from the Aptitude Tests (BGPA) and the Interest Test administered to students upon entering UTP. The data collected spans the years (2019, 2022, and 2023), totaling 5,321 records.

2.2.2 Data Analysis and Evaluation

Initially, a descriptive analysis of the previously collected data is performed, primarily classifying variables as continuous or categorical [29], [30]. Continuous variables refer to data that can take any value within a range and are numerical. Categorical variables have a limited number of values and are generally not numerical, although they may be numerically coded. The data collected are classified based on the type of variable (see ***Table 1***).

e-ISSN 2357-6014 / Vol. 21, no. 1 / january-april 2025 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

Table 1. Data Classification by Variable Type

VARIABLE TYPE	DATA
CATEGORICAL	'CENTRO', 'FACULTAD', 'CARRERA', 'TIPO_CARRERA', 'SEXO', 'TURNO', 'PROVINCIA_RESIDE', 'PSICOLOGIA', 'PRECALCULO', 'MATBASICA', 'IVEU', 'INGRESO FAMILIAR', 'GRADUADO'
CONTINUOUS	'AÑO', 'EDAD', 'PROMEDIO_COLEGIO', 'PRUBICACION', 'ULTIMO_PERIODO', 'TRANSPORTE', 'ALIMENTICIA', 'LENTES', 'TRABAJO COMPENSATORIO', 'CANASTA NAVIDEÑA', 'BECA', 'TABLETS', 'RECARGA', 'GENERAL', 'VERBAL', 'NUMERICA', 'ESPACIAL', 'PERCEPTUAL', 'OFICINESCA', 'CL', 'CTA', 'EA', 'II', 'MA', 'SCS', 'ANIO_ULT_MATRICULA'

Source: own work

For continuous variables, an initial descriptive analysis is performed, utilizing central measures [31] (mean, median, standard deviation, range, and interquartile range [32]). This preliminary analysis highlights notable results in some variables, providing better insights into the sample students. For instance, most students are around 19 years old, and the average GPA of half of the students entering the university is approximately 4.2.

For categorical variables, a frequency analysis of the data is conducted, identifying that most UTP students in the sample are enrolled in engineering programs (11,375 students), followed by bachelor's degrees (10,655 students), and finally technical programs (584 students), with a male predominance.

It is equally important to conduct a missing value analysis [33], identifying the quantity and proportion of missing data in each variable of the dataset. These missing values can affect the quality of analyses and the accuracy of predictive models. Initially, a column count is performed to quickly identify the variables most affected by missing values, showing the percentage of missing data. In this case:

- Less than 10% missing data: Variables like 'CARRERA' (6.05%) and 'FACULTAD' (2.78%).
- 10% to 50% missing data: Variables like 'MATBASICA' (61.97%), 'PRECALCULO' (54.29%), and 'PROVINCIA_RESIDE' (36.49%) are critical, and imputation methods such as the mean or median may be considered.
- More than 50% missing data: Variables related to students' psychological tests, with high percentages of missing values due to the low data concentration collected from the Psychological Guidance Directorate. However, methods like data multiplicity are considered essential, as these variables are crucial for the predictive model.

Outlier identification is carried out using techniques such as the Interquartile Range (IQR) method. Outliers are identified in continuous variables like AGE and SCHOOL_GPA, suggesting possible data entry errors during data collection.

To verify the quality of the collected data, data consistency is evaluated, ensuring uniformity and coherence within the dataset.

2.3 Phase 3: Data Preparation

This phase is crucial for this project, as the quality of the data used for predictive models directly impacts the precision and robustness that can be achieved. The goal of this phase is to establish an optimal dataset for implementation in the machine learning model under evaluation. The following steps are critical in determining the model's outcomes.

For the selection of the data that will constitute the input dataset for the predictive model, the preliminary analysis conducted earlier was considered. The variables CEDULA and IDESTUDIANTE were removed to ensure the anonymization of records for training the model. The selected variables include: AÑO, CENTRO, FACULTAD, TIPO_CARRERA, EDAD, SEXO, TURNO, PROVINCIA_RESIDE, PRUBICACION, PSICOLOGIA, PRECALCULO, MATBASICA, IVEU, FACULTAD_PREFERENCIA, INGRESO_FAMILIAR, ULTIMO_PERIODO, GRADUADO, TRANSPORTE, ALIMENTICIA, LENTES, TRABAJO_COMPENSATORIO, CANASTA_NAVIDEÑA, BECA, TABLETS, RECARGA, GENERAL, VERBAL, NUMERICA, ESPACIAL, PERCETUAL, OFICINESCA, CL, CTA, EA, II, MA, SCS, ANIO_ULT_MATRICULA.

Imputation techniques [34], [35] were applied to the selected variables, depending on their type, as described below: For continuous numerical variables with less than 50% missing data, Mean and Median imputation methods were employed [36]. Specifically, the Mean was used for variables like AÑO and INGRESO_FAMILIAR to maintain the central tendency of the original dataset. The Median was used for variables like EDAD, PROMEDIO_COLEGIO, and PRUBICACION, due to their susceptibility to outliers, minimizing the impact on the original data structure.

Normalization techniques were applied to ensure that variables have comparable ranges during model analysis. The following normalizations were performed: The AÑO variable was normalized using the Z-Score method [37], which centers data around the mean with a standard deviation of 1. The EDAD variable was normalized using Min-Max scaling [38], scaling values between 0 and 1. This increases the variable's uniformity and ensures it does not dominate others during modeling, as outliers were previously detected in this variable. The PRUBICACION, ANIO_ULT_MATRICULA,

and ULTIMO_PERIODO variables were normalized using Z-Score [37] to maintain consistency across variables, ensuring values are uniformly adjusted to the mean and enabling fair comparisons in time-dependent variables during modeling.

Following the dataset analysis and the significance of each variable, new columns were created, contributing highly representative information for data analysis: The INTERES column contains the student's primary area of interest based on the results of the psychological test. This test outputs numerical values representing the six faculties of the Universidad Tecnológica de Panamá. The highest score indicates the student's primary area of interest. This column is categorical, housing the initials of the resulting faculty. Consequently, the six columns CL, CTA, EA, II, MA, and SCS were excluded.

The DESERTADO column, which is vital for data analysis, indicates whether a student has dropped out. This is based on the project's identification methodology, which defines a student as dropped out if they have not enrolled for two periods (one year) in the education system. The value 'SI' was assigned if the student did not graduate, and their last enrollment was over a year ago. This column serves as the target variable for predictive modeling.

To finalize the proper structuring of the data, additional transformations were carried out on categorical variables, such as CENTRO, FACULTAD, TIPO_CARRERA, SEXO, TURNO, PROVINCIA_RESIDE, PSICOLOGIA, PRECALCULO, MATBASICA, IVEU, FACULTAD_PREFERENCIA, and INTERES. The One Hot Encoding method [39], [40] was applied to these variables. This method converts categorical variables into a numerical format so that machine learning algorithms can process them directly, representing each category as a separate binary column. This approach helps mitigate the introduction of biases or artificial hierarchies among categories.

2.4 Phase 4: Modeling

In the previous phase, the dataset was prepared for training the models to be developed. For this research, the following models were selected based on an analysis of related studies on student dropout prediction, where the best-performing models were:

- **Logistic Regression** [41]: This is a linear model particularly suited for binary classification problems, such as dropout prediction. It performs well on balanced datasets and is adaptable to regularization.

- **Random Forest** [42]: This is a robust algorithm composed of multiple decision trees, offering higher resistance to noise and tolerating data non-linearity.
- **Gradient Boosting** [43]: This model sequentially builds trees, with each new tree correcting the errors of the previous one. It can recognize more complex non-linear relationships than Random Forest; however, it is more prone to overfitting.

Three diverse scenarios were established to explore how class balancing, model inputs, and contextual features affect predictions. The scenarios are described as follows:

- **Scenario 1: Entire Balanced Dataset**
  This scenario uses the complete dataset, balancing the data since approximately 63% of the records did not drop out, and the remaining records are presumed to have dropped out. Such imbalance in prediction algorithms can favor the majority class, failing to correctly identify the minority class. The SMOTE (Synthetic Minority Over-sampling Technique) [44] was employed for balancing. This technique generates new synthetic samples from the minority class in a controlled manner, reducing overfitting compared to other techniques like oversampling.
- **Scenario 2: Entire Balanced Dataset without Psychological Contextual Features**
  Similar to the previous scenario, SMOTE was applied to balance the dataset. However, this scenario seeks to identify the importance of students' psychological context by removing these features from the predictive models. This approach evaluates whether demographic, socioeconomic, and academic data alone can produce solid predictions, considering the complexity of collecting psychological information.
- **Scenario 3: Specific Feature Selection**
  This scenario employs the RFECV (Recursive Feature Elimination with Cross-Validation) [45] approach, identifying variables that contribute most to dropout prediction. No data balancing is applied in this scenario, focusing on filtering features to reduce noise, improve generalization, and avoid overfitting. Irrelevant or redundant variables are eliminated, emphasizing those with predictive value and minimizing overfitting risks.

With the established scenarios and algorithms, the development of each predictive model was performed using Jupyter Notebook [46], a tool that enables the execution of Python code [47], a robust programming language for data processing. To ensure portability and scalability of the Jupyter environment, it was implemented within a Docker container [48], facilitating the future deployment of the model across diverse systems.

2.5 Phase 5: Evaluation

The evaluation of each implemented algorithm under the established scenarios is critical to determine the best combinations of scenarios and algorithms for this research. The following evaluation metrics are defined for each algorithm [49], [50]:

- **Accuracy (Overall Precision)**: The percentage of correct predictions made by the model overall.
- **Precision (Positive Precision)**: The proportion of correctly predicted positive cases.
- **Recall (Sensitivity or True Positive Rate)**: The proportion of true positives correctly identified.
- **F1 Score**: The harmonic mean of precision and recall.
- **ROC-AUC (Receiver Operating Characteristic Curve – Area Under the Curve)**: The ability of the model to distinguish between classes.
- **Log-Loss (Logarithmic Loss)**: Measures the quality of probabilistic predictions.
- **Cross-Validation**: Evaluates the model's generalization capacity by assessing performance across different dataset partitions.

With these evaluation metrics defined and their importance acknowledged in this research, the analysis of the results obtained from the algorithm development is presented in Section 3, RESULTS, of this study.

3. RESULTS

The results obtained for each of the prediction models developed in each of the scenarios are described below:

Table 2. Results of Algorithms in Scenario 1

Algorithm	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Log-Loss	Cross – Validation Accuracy
Logistic Regression	98.89%	99.90%	97.89%	98.88%	99.86%	3.50%	98.88% (+/- 0.14%)
Random Forest	98.95%	100%	97.91%	98.94%	99.78%	5.73%	98.91% (+/- 0.14%)
Gradient Boosting	98.87%	99.75%	97.99%	98.86%	99.88%	3.23%	98.89% (+/- 0.11%)

Source: own work

For **Scenario 1**, which uses the entire input dataset with balanced data, the results in ***Table 2*** show nearly perfect values across all three models. This suggests a potential **overfitting problem**, particularly in the Random Forest model, which yielded a perfect precision score, indicating that the model might be memorizing data instead of predicting appropriately. Despite the low Log-Loss values, the nearly perfect performance across all metrics is a clear indicator that the models may fail to generalize well in real-world settings.

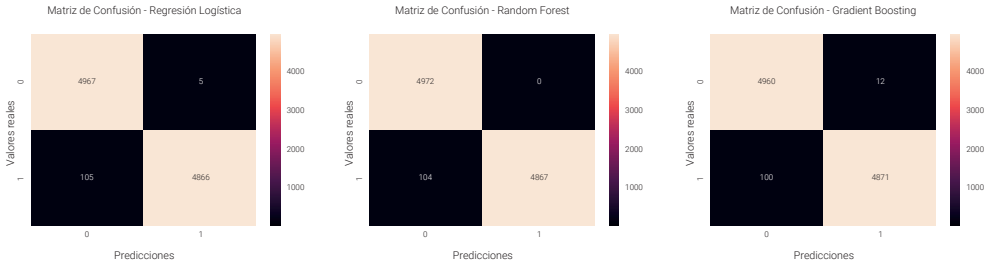

Figure 1. Confusion Matrices for Algorithms in Scenario 1
Source: own work

The confusion matrix in Figure 1 for the Scenario 1 algorithms reveals: **Logistic Regression**: Although it performed well, false negatives indicate that some dropout students were not identified correctly. **Random Forest**: The perfect absence of false positives reinforces the overfitting hypothesis. **Gradient Boosting**: While its false positives are slightly higher compared to other models, its high precision still suggests overfitting.

Table 3. Results of Algorithms in Scenario 2

Algorithm	Accuracy	Precisión	Recall	F1 Score	ROC-AUC	Log-Loss	Cross – Validation Accuracy
Regresión Logística	98.47%	99.86%	97.08%	98.45%	99.82%	4.17%	98.48% (+/- 0.20)
Random Forest	98.53%	99.57%	97.49%	98.52%	99.76%	5.65%	98.61% (+/- 0.19)
Gradient Boosting	98.48%	99.39%	97.57%	98.47%	99.83%	4.39%	98.65% (+/- 0.08)

Source: own work

Compared to Scenario 1, Scenario 2 results (***Table 3***) are slightly lower numerically. However, these results exhibit **greater robustness**, indicating better balance between model fit and its generalization capacity. The slight decrease in performance might be attributed to the exclusion of psychological context variables, which likely carry significant predictive weight.

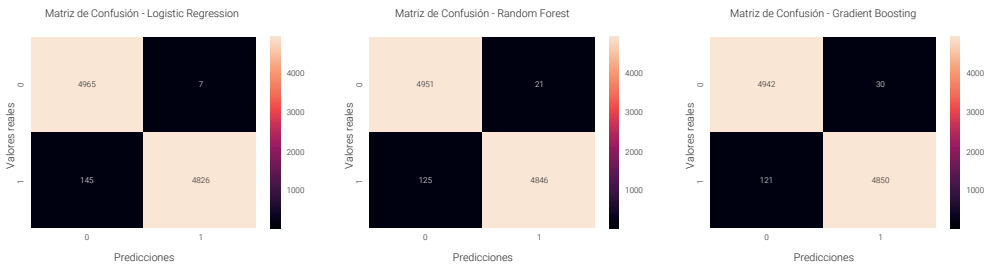

Figure 2. Confusion Matrices for Algorithms in Scenario 2
Source: own work

Regarding the results of the confusion matrices for the algorithms in **Scenario 2**, as shown in *Figure 2*, for Logistic Regression, there was an increase in the number of false negatives, representing a minimal loss in precision. For Random Forest, a slight increase in errors is observed (125 false negatives and 21 false positives). For Gradient Boosting, a slight decrease in precision is observed; however, its ability to generalize is better than in the previous scenario.

The results obtained for **Scenario 3** are presented in ***Table 4***. In this scenario, the input variables were selected using the RFECV technique. This technique optimizes cross-validation and focuses predictive models on those variables that truly provide significant value. This minimizes the risk of overfitting and noise in the results. In this case, a decrease in precision is noted across all three implemented algorithms. This could be due to the reduced number of input variables, but it indicates a better balance

between predictive capacity and the algorithms' ability to generalize, resulting in more realistic and less overfitted outcomes.

Table 4. Results of Algorithms in Scenario 3

Algorithm	Accuracy	Precisión	Recall	F1 Score	ROC-AUC	Log-Loss	Cross – Validation Accuracy
Regresión Logística	85.41%	85.91%	72.87%	78.85%	91.53%	34.80%	85.36% (+/- 0.66%)
Random Forest	91.25%	89.60%	86.64%	88.09%	96.97%	23.79%	91.25% (+/- 0.69%)
Gradient Boosting	87.94%	84.33%	83.16%	83.74%	95.75%	24.70%	87.94% (+/- 0.73%)

Source: own work

Analyzing the results of the algorithms for this scenario, it is observed that the Random Forest algorithm presents a good balance between precision (89.60%), recall (86.64%), and ROC-AUC (96.97%), which suggests that it can identify dropouts without overfitting to the training data, reinforcing its predictive capability in real-world settings.

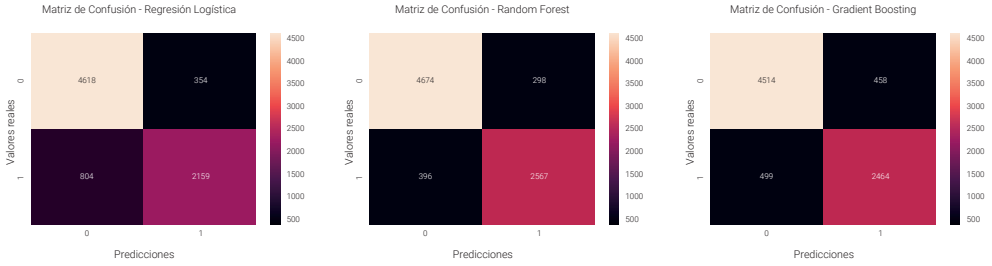

Figure 3. Confusion Matrices for Algorithms in Scenario 3.
Source: own work

Finally, the confusion matrices for the algorithms in Scenario 3 are shown in *Figure 3*. For Logistic Regression, there is an increase in the number of false positives and false negatives, which suggests greater difficulty in making predictions with fewer variables. Regarding Random Forest, it shows a smaller number of false negatives and false positives, making it more balanced with better generalization results. As for Gradient Boosting, although its number of false positives is higher than the other algorithms, it offers competitive performance. However, compared to the other scenarios, it performs worse in this one.

After conducting a thorough analysis of the results of the algorithms implemented in each of the scenarios and the values of the metrics obtained, Scenario 3 is presented as the one that shows the best balance between the predictive capability and generalization of the algorithms, as, although the metrics are inferior to the other scenarios, it does not exhibit elevated overfitting.

4. DISCUSSION AND CONCLUSIONS

The data mining process has been a fundamental part of this research, enabling the preparation of a robust and reliable dataset through various phases. Processes such as imputing missing values, handling outliers, and normalizing variables were critical in ensuring the quality of the data used in the predictive model.

The multidimensional approach allowed for the development of a context-aware predictive model that incorporates traditional aspects, such as secondary school academic performance and economic factors, as well as a more innovative focus on emotional and psychological well-being. This approach strengthens the research by enhancing our understanding of university dropout rates at the Technological University of Panama (UTP) and offering potential solutions to mitigate these rates.

After conducting an in-depth analysis of the results of the algorithms implemented in each scenario and evaluating the metrics obtained, Scenario 3 emerges as the one that offers the best balance between predictive capability and algorithm generalization. Although its metrics are slightly inferior to those of other scenarios, it does not exhibit excessive overfitting and presents itself as the most realistic scenario for implementation in a production environment.

Regarding the algorithms implemented, Random Forest demonstrated superior results in the metrics. It is a model capable of accurately identifying students at risk of dropping out without overfitting, as evidenced by the balance between precision, recall, and the F1 Score. This characteristic is essential for ensuring the model performs effectively with future data.

The university dropout prediction model was achieved using machine learning algorithms that are primarily employed for this type of problem and are not overly complex to implement. Three algorithmic models—Logistic Regression, Random Forest, and Gradient Boosting—were developed across three different scenarios to analyze the behavior of the data and the results of the prediction models. The results suggest that the Random Forest model implemented in Scenario 3 is the most realistic and effective, offering an optimal balance between precision and generalization, while minimizing the risk of overfitting observed in other scenarios.

This modeling achievement strengthens the value of the project, demonstrating that it is possible to predict with a high degree of accuracy which students are at risk of dropping out, without relying on excessively specific or overly revealing data. The model's precision and effectiveness were validated through various evaluation metrics, where the Random Forest in Scenario 3 delivered the most realistic results. The comparison between scenarios and prediction algorithm models highlights the importance of carefully balancing detailed and general data.

Finally, through the use of artificial intelligence technologies, this project represents a significant advancement in addressing university dropout at UTP. It facilitates the timely identification of students at risk of dropping out. Every step of the process—from identifying key variables to evaluating scenarios—demonstrates a strict commitment to the objectives set and the use of innovative tools that enhance predictive capacity and system efficiency.

It is recommended to expand the variables feeding the prediction model to make it more robust, such as including grades, student attendance, and relevant psychological factors, along with a more comprehensive evaluation of individual well-being. This would further improve the system's contextualization of each student, thereby increasing its utility.

5. ACKNOWLEDGEMENTS

L. Arenales is supported by a scholarship from the National Program for the Strengthening of Postgraduate Studies of the National Secretariat of Science, Technology, and Innovation (SENACYT) of Panama. V. Villarreal and J.J. Saldana-Barrios are members of the National System of Researchers (SNI) of SENACYT. This research is part of the project "Analysis, Design, and Development of an Intelligent Context-Aware System for the Prediction of University Dropouts, Using a Microservices Architecture," which is fully funded by contract No. DDCCT No. 125-2023 received from the National Secretariat of Science, Technology, and Innovation (SENACYT) of Panama.

6. REFERENCES

[1] P. A. Flores Mora, "Abordando el Desafío de la Deserción: una revisión sistemática de la Deserción en Instituciones de Educación Superior," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 8, no. 1, pp. 9813–9823, Apr. 2024. doi: 10.37811/cl_rcm.v8i2.10299.

[2] E. Himmel, "Modelos de Análisis de la deserción estudiantil en la educación superior," *Calidad en la Educación*, pp. 1, 2002.

[3] I. Quintero, "Análisis de las causas de deserción universitaria," Bogota, Colombia, 2016.

[4] K. S. Selim and S. S. Rezk, "On predicting school dropouts in Egypt: A machine learning approach," *Educ Inf Technol (Dordr)*, vol. 28, no. 7, pp. 9235–9266, Jul. 2023. doi: 10.1007/s10639-022-11571-x.

[5] A. F. Núñez-Naranjo, "Analysis of the determinant factors in university dropout: a case study of Ecuador," *Front Educ (Lausanne)*, vol. 9, p. 1444534, Oct. 2024. doi: 10.3389/FEDUC.2024.1444534/BIBTEX.

[6] World Bank, "Acting Now to Protect the Human Capital of Our Children: The Costs of and Response to COVID-19 Pandemic's Impact on the Education Sector in Latin America and the Caribbean," 2021. [Online]. Available: www.worldbank.org

[7] P. Díaz and A. Tejedor, "Programas de análisis y prevención de la deserción estudiantil universitaria dirigidos al contexto panameño", pp. 1, Panamá, 2018. doi: http://ridda2.utp.ac.pa/handle/123456789/5530.

[8] L. Reisberg, "Diagnóstico de la educación superior en Panamá: Retos y Oportunidades," Panamá, pp. 1, Mar. 2021. doi: http://dx.doi.org/10.18235/0003329.

[9] A. A. Pérez Aguirre, M. De Gracia, M. Aguirre, I. Ordas, and A. Aranzazu, "LA DESERCIÓN EN LA EDUCACIÓN SUPERIOR EN PANAMÁ Y SUS CAUSAS," Universidad Internacional de Ciencia y Tecnología, Jan. 2023, pp. 314–325. doi: 10.47300/actasidi-unicyt-2022-47.

[10] N. Reyes and A. Meneses, "Una revisión crítica de los factores psicosociales asociados al abandono universitario en primer año.", pp. 1.

[11] N. Chiarino *et al.*, "Abandono y permanencia estudiantil en universidades de Latinoamérica y el Caribe: Una revisión sistemática mixta," *Actualidades Investigativas en Educación*, vol. 24, no. 2, pp. 1–37, May 2024. doi: 10.15517/aie.v24i2.57306.

[12] F. Sáez and J. Mella, "Revisión sistemática sobre intención de abandono en educación superior.", pp. 1.

[13] G. Paramo and C. Correa, "Deserción estudiantil universitaria. Conceptualización," *Revista Universidad EAFIT Colombia*, vol. 35, pp. 65–78, Jul. 2012.

[14] M. Biancardi, “Complejidad del concepto de contexto,” 1998.

[15] S. W. Nava-Díaz and G. Chavira, “Hacia los entornos conscientes del contexto: Un marco de trabajo para etiquetar consciencia del contexto,” *Revista Colombiana de Computación*, vol. 14, no. 2, pp. 61–78, Nov. 2013.

[16] N. Díaz and S. S. M. Wilfrido, “Modelado de un ambiente inteligente: un entorno consciente del contexto a través del etiquetado.,” pp. 1, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:177669972

[17] T. Mitchell, *Machine Learning*, pp. 1, vol. 1. 1997.

[18] P. Mussida and P. L. Lanzi, “A computational tool for engineer dropout prediction,” in *IEEE Global Engineering Education Conference, EDUCON*, IEEE Computer Society, 2022, pp. 1571–1576. doi: 10.1109/EDUCON52537.2022.9766632.

[19] E. Cevallos Medina, C. Barahona Chunga, J. Armas-Aguirre, and E. Gradón, “Predictive model to reduce the dropout rate of university students in Perú: Bayesian Networks vs. Decision Trees,” *5th Iberian Conference on Information Systems and Technologies (CISTI)*, vol. 15, pp. 1–7, Jun. 2020. doi: 10.23919/CISTI49556.2020.9141095.

[20] S. Merchán Rubiano, A. Beltrán Gómez, and J. Duarte García, “Engineering Students´ Academic Performance Prediction using ICFES Test Scores and Demo-graphic Data,” *Ingeniería Solidaria*, vol. 13, no. 21, pp. 53–61, Jan. 2017. doi: 10.16925/in.v13i21.1729.

[21] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, “MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series,” *IEEE Access*, vol. 8, pp. 225324–225335, 2020. doi: 10.1109/ACCESS.2020.3045157.

[22] Á. Rabelo Lopes, R. De Sousa, D. Carvalho, and R. Vaccare, “Context-aware Ubiquitous Learning Literature Systematic Mapping on Ubiquitous Learning Environments,” *International Symposium on Computers in Education (SIIE)*, pp. 1–6, 2017. doi: 10.1109/SIIE.2017.8259662.

[23] Á. R. Lopes, D. C. de Oliveira, R. C. de Sousa Aguiar, and R. T. Vaccare Braga, “Context-aware ubiquitous learning: Literature systematic mapping on ubiquitous learning environments,” in *2017 International Symposium on Computers in Education (SIIE)*, 2017, pp. 1–6. doi: 10.1109/SIIE.2017.8259662.

[24] Z. Yu, X. Zhou, and L. Shu, “Towards a Semantic Infrastructure for Context-Aware e-Learning,” *Multimedia Tools Appl.*, vol. 47, no. 1, pp. 71–86, Mar. 2010. doi: 10.1007/s11042-009-0407-4.

[25] P. Chapman, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," DaimlerChrysler, pp. 1, 2000.

[26] "Dirección General de Tecnología de la Información y Comunicaciones | Universidad Tecnológica de Panamá.", pp. 1, Accessed: Jun. 22, 2024. [Online]. Available: https://utp.ac.pa/direccion-general-de-tecnologia-de-la-informacion-y-comunicaciones

[27] "Dirección de Bienestar Estudiantil | Universidad Tecnológica de Panamá." Accessed: Jun. 22, 2024. [Online]. Available: https://utp.ac.pa/direccion-de-bienestar-estudiantil

[28] "Dirección de Orientación Psicológica | Universidad Tecnológica de Panamá." Accessed: Jun. 22, 2024. [Online]. Available: https://utp.ac.pa/direccion-de-orientacion-psicologica

[29] I. Hodashinsky and R. Ostapenko, "Extracting Fuzzy Classifier Rules from Mixed Continuous and Categorical Data," in *2024 X International Conference on Information Technology and Nanotechnology (ITNT)*, 2024, pp. 1–6. doi: 10.1109/ITNT60778.2024.10582321.

[30] A. Taha and O. M. Hegazy, "A proposed outliers identification algorithm for categorical data sets," in *2010 The 7th International Conference on Informatics and Systems (INFOS)*, 2010, pp. 1–5.

[31] R. Pavithrakannan, N. B. Fenn, S. Raman, V. Kalyanaraman, V. K. Murugananthan, and J. Janarthanan, "Imputation Analysis of Central Tendencies for Classification," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–7. doi: 10.1109/IEMTRONICS52119.2021.9422507.

[32] S. Guo, X. Wu, and Y. Li, "Deriving Private Information from Perturbed Data Using IQR Based Approach," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, p. 92. doi: 10.1109/ICDEW.2006.47.

[33] A. Sadhu, R. Soni, and M. Mishra, "Pattern-based Comparative Analysis of Techniques for Missing Value Imputation," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 2020, pp. 513–518. doi: 10.1109/ICCCA49541.2020.9250825.

[34] J. Yu, Y. He, and J. Z. Huang, "A Two-Stage Missing Value Imputation Method Based on Autoencoder Neural Network," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 6064–6066. doi: 10.1109/BigData52589.2021.9671338.

[35] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating Synthetic Missing Data: A Review by Missing Mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, 2019. doi: 10.1109/ACCESS.2019.2891360.

[36] "Tratamiento de valores vacíos II con R y Python: Estrategias de imputación estadística (moda, mediana y media). | by Nicolás Urrego | Medium.", pp. 1, Accessed: Sep. 14, 2024. [Online]. Available: https://nicolasurrego.medium.com/tratamiento-de-valores-vac%C3%ADos-ii-estrategias-de-imputaci%C3%B3n-estad%C3%ADstica-moda-mediana-y-media-2665b0f53a4c

[37] H. Chen *et al.*, "Personal-Zscore: Eliminating Individual Difference for EEG-Based Cross-Subject Emotion Recognition," *IEEE Trans Affect Comput*, vol. 14, no. 3, pp. 2077–2088, 2023. doi: 10.1109/TAFFC.2021.3137857.

[38] V. Gajera, Shubham, R. Gupta, and P. K. Jana, "An effective Multi-Objective task scheduling algorithm using Min-Max normalization in cloud computing," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 812–816. doi: 10.1109/ICATCCT.2016.7912111.

[39] A. Aich, A. Krishna, V. Akhilesh, and C. Hegde, "Encoding Web-based Data for Efficient Storage in Machine Learning Applications," in *2019 Fifteenth International Conference on Information Processing (ICINPRO)*, 2019, pp. 1–6. doi: 10.1109/ICInPro47689.2019.9092264.

[40] B. Sun, J. Wu, C. Guo, and K. Chen, "A One-Hot Encoding Approach for Signal Integrity Enhancement of Intra-chip Interconnects," in *2024 25th International Conference on Electronic Packaging Technology (ICEPT)*, 2024, pp. 1–6. doi: 10.1109/ICEPT63120.2024.10668712.

[41] B. Ujkani, D. Minkovska, and L. Stoyanova, "Application of Logistic Regression Technique for Predicting Student Dropout," in *2022 31st International Scientific Conference Electronics, ET 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ET55967.2022.9920280.

[42] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[43] E. Cruz, M. González, and J. C. Rangel, "Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión.," *Prisma Tecnológico*, vol. 13, no. 1, pp. 77–87, Feb. 2022. doi: 10.33412/pri.v13.1.3039.

[44] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, pp. 1, Mar. 2013. doi: 10.1186/1471-2105-14-106.

[45] N. Awasthi, P. R. Gautam, and A. K. Sharma, "RFECV-DT: Recursive Feature Selection with Cross Validation using Decision Tree based Android Malware Detection," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10725127.

[46] V. F. Ochkov, A. Stevens, and A. I. Tikhonov, "Jupyter Notebook, JupyterLab – Integrated Environment for STEM Education," in *2022 VI International Conference on Information Technologies in Engineering Education (Inforino)*, 2022, pp. 1–5. doi: 10.1109/Inforino53888.2022.9782924.

[47] P. Prathanrat and C. Polprasert, "Performance Prediction of Jupyter Notebook in JupyterHub using Machine Learning," in *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 2018, pp. 157–162. doi: 10.1109/ICIIBMS.2018.8550030.

[48] F. B. Fava *et al.*, "Assessing the Performance of Docker in Docker Containers for Microservice-Based Architectures," in *2024 32nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 2024, pp. 137–142. doi: 10.1109/PDP62718.2024.00026.

[49] A. Suryanarayanan, A. Chala, L. Xu, G. Shobha, J. Shetty, and R. Dev, "Design and Implementation of Machine Learning Evaluation Metrics on HPCC Systems," in *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, 2019, pp. 1–7. doi: 10.1109/CSITSS47250.2019.9031056.

[50] M. F. Uddin, "Addressing Accuracy Paradox Using Enhanched Weighted Performance Metric in Machine Learning," in *2019 Sixth HCT Information Technology Trends (ITT)*, 2019, pp. 319–324. doi: 10.1109/ITT48889.2019.9075071.