ESCUELA NAVAL DE CADETES
"ALMIRANTE PADILLA"
Honor y Tradición

OnBoard Knowledge Journal

Research article

Prediction of Course Failure in University Students Using Machine Learning Techniques: Comparison of Supervised Models

Predicción de Reprobación de Asignaturas en Estudiantes Universitarios mediante Técnicas de Machine Learning: Comparación de Modelos Supervisados

Inés Figueroa-Sarmiento [1] Fabio Romero-Gómez [1] and Andrés Solano-Barliza [1]*

[1] Faculty of Engineering, Universidad de La Guajira, Riohacha, 440001, Colombia; ifigueroas@uniguajira.edu.co; fdromero@uniguajira.edu.co; andresolano@uniguajira.edu.co

* Correspondence: andresolano@uniguajira.edu.co

Citation: Figueroa, I., Romero, Fabio., Solano, A. Machine Learning Prediction of Course Failure. *OnBoard Knowledge Journal* **2026**, *2*, 1. https://doi.org/10.70554/OBJK2026.v02n02.01

Received: 03/06/2026, Accepted: 15/07/2026, Published: 30/07/2026

DOI: https://doi.org/10.70554/OBJK2026.v02n02.01

Abstract: The use of digital platforms among Colombian college students continues to grow, and with that, the question of whether it affects retention and academic performance has resurfaced. In Riohacha, this relationship had not been studied using artificial intelligence tools at the institutional level. For this study, a dataset was constructed based on a survey administered to 300 undergraduate students at the University of La Guajira. Twenty-one sociodemographic, behavioral, and digital perception variables were captured and organized using the CRISP-DM methodology. The target variable is binary: whether the student has failed at least one course. Three supervised classification algorithms (Decision Tree, KNN, and Random Forest) were compared, integrated into pipelines with SMOTE and five-fold stratified cross-validation. Random Forest yielded the best results, with 62% accuracy, an F1-Score of 0.606, and a ROC AUC of 0.682. Feature importance of analyzing the level of concentration while studying, daily hours of cell phone use, and preferred social network as the most significant predictors.

Keywords: Machine learning; Decision tree; K-nearest Neighbors; Random Forest; SMOTE; Cross-validation; Academic failure; Social media; Feature importance

Resumen: El uso de plataformas digitales entre universitarios colombianos sigue creciendo, y con eso ha vuelto a abrirse la pregunta de si afecta la permanencia y el rendimiento académico. En Riohacha esa relación no se había estudiado con herramientas de inteligencia artificial a nivel institucional. Para este trabajo se construyó un conjunto de datos a partir de una encuesta aplicada a 300 estudiantes de pregrado de la Universidad de La Guajira. Se capturaron 21 variables sociodemográficas, conductuales y de percepción digital organizadas bajo la metodología CRISP-DM. La variable objetivo es binaria: si el estudiante ha reprobado al menos una asignatura. Se compararon tres algoritmos de clasificación supervisada (Árbol de Decisión, KNN y Random Forest) integrados en pipelines con SMOTE y validación cruzada estratificada de cinco pliegues. Random Forest dio el mejor resultado, con 62% de exactitud, F1-Score de 0.606 y

© 2026 by authors.
Licensed by Escuela Naval de Cadetes "Almirante Padilla", COL.
This article is freely accessible and distributed under the terms and conditions of Creative Commons Attribution (https://creativecommons.org/licenses/by/4.0/).

ROC AUC de 0.682. El análisis de importancia de características señaló al nivel de concentración al estudiar, las horas diarias de uso del celular y la red social preferida como los predictores con mayor peso.

Palabras clave: Aprendizaje automático; Árbol de decisión; K-nearest Neighbors; Random Forest; SMOTE; Validación cruzada; Reprobación académica; Redes sociales; Importancia de características

1. Introduction

Transition into and persistence in higher education are conditioned by multiple vectors of academic vulnerability, among which course failure stands out as the earliest and most irreversible predictor of permanent withdrawal from the system [17]. In this regard, failure in individual courses triggers a chain effect characterized by the prolongation of graduation trajectories, an increase in institutional operating costs, and socioeconomic strain on students' families [1]. Within the context of Colombian higher education, where aggregate dropout rates have historically ranged between 40% and 50% according to official records, course failure operates as a critical symptom of institutional disengagement [11]. Consequently, the development of predictive tools capable of identifying students at imminent risk at an early stage has become a management imperative, as it enables the implementation of targeted pedagogical interventions before academic delay becomes statistically irreversible [20] [3].

Although passing and failing formally represent binary states within grading systems, the determinants underlying these outcomes form a multicausal and heterogeneous framework. Adverse academic performance does not arise from a single isolated factor; rather, it emerges from the complex interaction among gaps in prior cognitive competencies, socioeconomic disparities, deficiencies in self-regulation strategies for independent study time, and, increasingly, the systematic interference of mobile devices in processes of focused attention in the classroom [24]. Because these causes combine differently in each individual, anticipating risk through linear statistical models or simple descriptive approaches is methodologically insufficient. Conceptualizing academic performance through a dichotomous approach—pass versus fail—offers a strategic advantage for university early-warning systems, as it makes it possible to operationalize and direct institutional tutoring resources toward the affected student population, thereby optimizing retention efforts instead of distributing them uniformly across the general population [13].

In this regard, measuring academic performance without considering digital consumption produces incomplete results. Traditional methods, such as linear regression or cumulative grade averages, assume linear relationships among variables and homogeneous variances; however, human behavior does not follow such rules. Digital stimulation competes with the ability to concentrate and with the cognitive self-regulation required for independent study [21], [22]. Social media platforms rely on algorithms designed to keep users engaged for as long as possible, within the logic of the attention economy [9]. Therefore, modeling contemporary academic performance while omitting the dimension of sociotechnical consumption leads to biased and incomplete conclusions. Traditional predictive frameworks assume static and independent relationships among variables, a theoretical assumption that conflicts with the dynamics of the so-called attention economy [4].

Digital micro-content platforms such as TikTok, Instagram, and WhatsApp deploy algorithmic architectures explicitly designed to prolong screen time, directly competing with the attentional control and working-memory processes that are essential for independent study. Studies such as [23] demonstrate that behavioral traces and digital browsing habits contain highly predictive signatures that are strongly correlated with academic failure when analyzed through supervised learning techniques. Within this ecosystem, unsupervised hyperconnectivity acts as a vector of learning fragmentation that reduces the efficiency of the time devoted to autonomous learning, making maladaptive social media use an essential predictive variable in the design of any modern early-warning system.

Supervised Machine Learning classifiers identify nonlinear patterns among multiple variables and can be applied to new data without relying on manually written rules. In other countries, Decision Trees,

K-Nearest Neighbors (KNN), and Random Forest models have been used to study academic performance, although most studies have focused on continuous grade averages rather than course failure as a binary event.

The objective of this article is to design a predictive model based on supervised machine learning to anticipate the risk of course failure among students at the University of La Guajira, using sociodemographic variables, digital behavior, and social media consumption. Three classifiers (Decision Tree, KNN, and Random Forest) are compared using metrics oriented toward the model's practical usefulness in university intervention.

The remainder of this article is organized as follows. Section 2 summarizes the main contributions of the study. Section 3 reviews previous research on academic performance prediction, course failure, and the influence of digital behavior using machine learning techniques. Section 4 describes the research design and includes the data understanding and description, data preparation, class-imbalance treatment, supervised modeling, and feature-importance analysis presented in Sections 4.1–4.5. Section 5 reports the experimental findings, including the comparison of the supervised models and the analysis of the most influential predictors in Sections 5.1 and 5.2, respectively. Section 6 discusses the implications of the results, Section 7 presents the main limitations of the study, and Section 8 provides the conclusions and directions for future research.

2. Contributions

This study presents a supervised machine learning approach for predicting course failure among university students using sociodemographic, academic, behavioral, and digital-use variables. The main contributions of this work to the existing body of knowledge are threefold:

ii. Unlike previous studies that primarily focus on predicting continuous academic outcomes, such as grade point averages or final course grades, this research addresses course failure as a binary classification problem. This formulation is directly aligned with the requirements of institutional early-warning systems, since it enables universities to identify students who may require timely tutoring, academic counseling, or preventive support before failure contributes to academic delay or dropout.

ii. The study provides a comparative and methodologically consistent evaluation of three supervised learning algorithms: Decision Tree, K-Nearest Neighbors, and Random Forest. The models were implemented using pipelines that integrate SMOTE and five-fold stratified cross-validation, ensuring that synthetic observations were generated exclusively from the training data in each fold and preventing information leakage into the test sets. In addition, variables directly derived from course failure were removed before modeling. Under this evaluation framework, Random Forest achieved the best overall performance, with an accuracy of approximately 62%, an F1-Score of 0.606, and a ROC AUC of 0.682.

ii. This work contributes empirical evidence from a university context in the Colombian Caribbean that has received limited attention in previous machine learning studies. Using primary data collected from 300 undergraduate students at the University of La Guajira, the analysis identifies concentration while studying, mobile-phone use, preferred social media platform, academic program, semester, socioeconomic characteristics, and study habits as relevant predictors of course failure. These findings provide an interpretable basis for the future development of institutional early-warning tools and suggest that the manner in which students regulate attention and interact with digital technologies may be more informative than screen time alone.

3. Related Works

The use of Machine Learning to predict academic performance has expanded in recent years, facilitated by the greater availability of institutional data and lower computational costs. López-García et al. [14] demonstrated that the early detection of academic failure is possible using ML and institutional data, achieving AUC values above 0.70.

Ramzan et al. [18] developed SocialNet to predict academic performance using social media data and Random Forest, achieving 81% accuracy. Usage time and content type emerged as the most influential

predictors, a finding that is particularly relevant to the present study, in which digital consumption plays a central role.

Al Mamun et al. [2] applied Gradient Boosting to 1,200 surveys and achieved 93% accuracy using cross-validation. This result differs substantially from ours; however, the differences can be explained by sample size, the variables included, and the specific context of Riohacha.

Martínez et al. [16] assessed behavioral and personality factors among engineering students using Decision Trees and Support Vector Machines (SVM), obtaining F1-Scores above 78%, although their study relied on limited longitudinal data. Gil-Vera [8] implemented neural networks at a Colombian university and outperformed traditional statistical methods, indicating that ML has significant potential within the national context.

At the local level, Martelo et al. [15] examined the influence of social media on academic performance using descriptive and inferential statistics and identified correlations among frequency of use, type of social network, and academic outcomes. Contreras et al. [7] used machine learning to predict success and failure in engineering programs, reporting promising results through binary classification. In Bangladesh, Hemal et al. [12] found that Random Forest provided the best predictive performance when assessing the impact of internet use on academic achievement, with results similar to those obtained in the present study.

Overall, Random Forest and ensemble methods tend to achieve the strongest performance in this type of problem. The present study differs from previous work in that it addresses course failure as a binary outcome, incorporates SMOTE within cross-validation pipelines, and uses primary data from a Colombian university context that had not previously been examined using these techniques.

4. Methodology

This study employed a quantitative predictive approach, structured into four phases following the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining): domain understanding, data understanding, data preparation, and modeling and evaluation [6]. The computational analysis was conducted in Python 3 using Google Colab, a cloud-based platform that provides free access to computing resources. The libraries used included pandas for tabular data manipulation, NumPy for efficient numerical operations, scikit-learn for the implementation of Machine Learning algorithms, imbalanced-learn for the application of SMOTE, and Matplotlib and Seaborn for exploratory visualization and graph generation.

4.1. Data Understanding and Description

The dataset was constructed from a self-administered digital survey developed in Google Forms and distributed to 300 undergraduate students enrolled at the University of La Guajira during the 2026-1 academic term. A total of 300 valid responses were obtained after applying data-cleaning and quality-control criteria. The instrument was validated by five experts in educational technology and artificial intelligence, who assessed the relevance, clarity, and adequacy of the items, thereby ensuring content validity.

The dataset contains 21 variables organized into five sections: general information, time spent on social media, academic habits, academic performance, and perceived impact. These variables include sociodemographic information, patterns of mobile-device and social-media use, academic behaviors, and perceptions of the impact of digital consumption.

The target variable of the study is `course_failure`, a binary variable that is, a variable that can take only two values, which indicates whether the student has failed at least one course (Yes = 1, No = 0). Table 1 presents the complete data dictionary:

Table 1. Description and data types of the variables included in the dataset.

#	Variable	Description	Statistical Type	Python Type
1	`edad_rango`	Student age range	Ordinal	`category`
2	`genero`	Student gender	Nominal	`category`
3	`programa`	Academic program	Nominal	`category`
4	`semestre`	Current semester enrolled	Discrete	`int`
5	`estrato`	Socioeconomic stratum	Ordinal	`category`
6	`horas_celular_dia`	Daily hours of mobile-phone use	Ordinal	`category`
7	`horas_redes_dia`	Daily hours spent on social media	Ordinal	`category`
8	`red_principal`	Most frequently used social media platform	Nominal	`category`
9	`uso_redes`	Primary purpose of social media use	Nominal	`category`
10	`frec_celular_no_acad_clase`	Frequency of non-academic mobile-phone use during class	Ordinal	`category`
11	`frec_celular_acad_clase`	Frequency of academic mobile-phone use during class	Ordinal	`category`
12	`horas_estudio_dia`	Hours of independent study per day	Ordinal	`category`
13	`nivel_concentracion`	Concentration level while studying	Ordinal	`category`
14	`frec_distraccion_celular`	Frequency of mobile-phone distraction	Ordinal	`category`
15	`frec_uso_acad_celular`	Frequency of academic mobile-phone use	Ordinal	`category`
16	`uso_acad_celular_tipo`	Type of academic mobile-phone use	Nominal	`category`
17	`reprobacion`	Has the student failed any course?	Binary	`bool`
18	`num_reprobadas`	Number of failed courses	Ordinal	`category`
19	`redes_afectan_rendimiento`	Do social media platforms affect academic performance?	Binary	`bool`
20	`tipo_impacto`	Type of perceived impact	Nominal	`category`
21	`intento_reducir_uso`	Has the student attempted to reduce mobile-phone use?	Binary	`bool`

Figure 1 presents the distribution of the target variable (course failure). It can be observed that 51.5% of the students reported having failed at least one course, whereas 48.5% indicated that they had not failed any courses. Although the class imbalance is moderate, with a difference of approximately three percentage points, the application of SMOTE remains advisable to balance the classes and prevent the model from becoming biased toward the majority class during training, particularly when combined with stratified cross-validation, which preserves the class proportions within each fold.

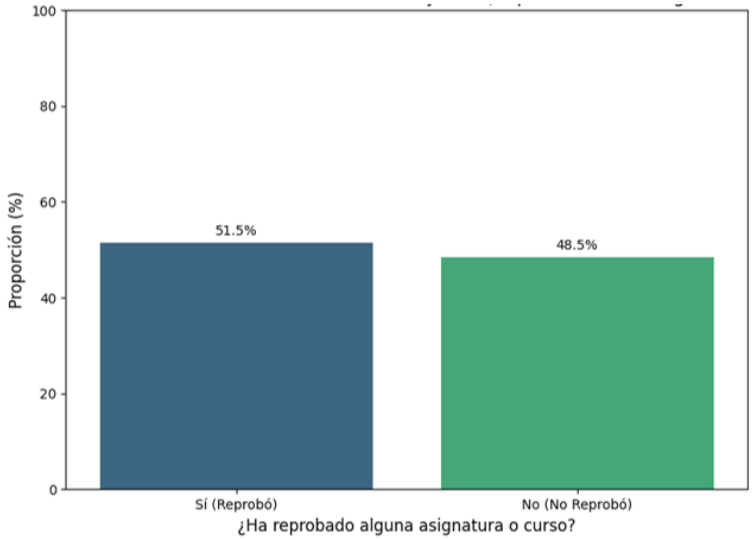


Figure 1. Distribution of the target variable.

4.2. Data Preparation

The original column corresponding to the question "Have you failed any subject or course?" was renamed `Failed_Courses_Target`, and its values ("Yes" and "No") were converted to 1 and 0, respectively. Four columns were removed to prevent data leakage: `Number_Failed_Courses` and `Failed_Courses_Yes`, which were directly derived from the target variable; Timestamp, which had no predictive value; and Current Academic Average, which was highly correlated with course failure and could therefore have biased the model. The variable "If your answer is yes, how many?" was also removed after the exploratory analysis because it functioned as an indirect source of data leakage.

Categorical variables were transformed using one-hot encoding (`get_dummies` with `drop_first=True` to prevent collinearity), resulting in a total of 67 numerical variables. To reduce dimensionality, Principal Component Analysis (PCA) was applied while retaining 95% of the explained variance, reducing the dataset to 52 principal components. This orthogonal linear transformation preserves most of the original information and mitigates sparsity-related problems in high-dimensional spaces.

Despite the reduction to 52 PCA components, the ratio between the number of variables and the sample size remained moderate. No additional feature-selection techniques, such as Lasso or recursive feature elimination, were applied for two reasons. First, Random Forest handles moderate-dimensional datasets effectively because it considers only a random subset of variables at each split (`max_features='sqrt'`). Second, Gini-based feature importance calculated from the original variables prior to PCA makes it possible to examine the predictive contribution of each category without sacrificing interpretability. The application of PCA primarily benefited KNN, as discussed in Section 4.1.

4.3. Treatment of Class Imbalance

The target variable exhibited class imbalance, with more students reporting no course failures than students reporting at least one failure. When class imbalance is present, models trained using standard methods tend to favor the majority class and lose sensitivity in detecting students who did fail a course. To address this issue, SMOTE [16] was incorporated into each pipeline. SMOTE identifies neighboring observations from the minority class in the feature space and generates new synthetic data points through interpolation. By integrating SMOTE into the pipeline, synthetic samples are generated exclusively from the training data within each cross-validation fold, thereby preventing contamination of the test data.

4.4. Supervised Modeling

Three supervised classification algorithms were tested using their default configurations. The Decision Tree algorithm (DecisionTreeClassifier, `random_state=42`) was selected because of its interpretability: it applies a sequence of binary decisions until reaching a final classification. KNN (KNeighborsClassifier, default parameters) assigns a class according to the most frequent label among the nearest neighbors. Random Forest (RandomForestClassifier, `random_state=42`) combines multiple decision trees, can capture complex interactions, and provides estimates of feature importance.

Each model was encapsulated in a scikit-learn Pipeline that applies SMOTE to the training data, fits the classifier, and generates predictions. Evaluation was performed using StratifiedKFold with five folds, shuffle=True, and `random_state=42`, preserving the class distribution within each fold. The evaluation metrics were accuracy, precision, recall, F1-Score, and ROC AUC. All reported values correspond to the average across the five folds.

4.5. Feature Importance Analysis

The Random Forest model [15] was retrained on the 67 original variables prior to PCA, using all 300 observations, to extract feature importance values based on Gini impurity reduction. Gini impurity measures the degree of class heterogeneity within a tree node; lower values indicate purer nodes in which the classes are more clearly separated. For each variable, its importance is calculated according to the extent to which it reduces the weighted average Gini index across the nodes where that variable is used to create a binary

split, normalized by the number of samples reaching those nodes. Variables that produce purer splits receive higher Gini importance values.

The features were ranked in descending order of importance, making it possible to identify those with the greatest capacity to differentiate between students who had failed at least one course and those who had not. This analysis improves model interpretability and facilitates the identification of key factors for designing effective pedagogical interventions.

5. Results

5.1. Comparison of Supervised Models

Table 2 presents the average results obtained through five-fold stratified cross-validation with SMOTE integrated into the training pipelines. Each value corresponds to the mean of the five evaluation iterations performed on the test data from each fold, ensuring that the results are not biased by any particular data split.

Table 2. Performance metrics of the evaluated classification models.

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Decision Tree	0.580710	0.597782	0.564113	0.578997	0.581252
KNN	0.567705	0.586427	0.532056	0.557300	0.592696
Random Forest	0.620328	0.652794	0.570766	0.605581	0.681593

Random Forest achieved a ROC AUC of 0.682 and an accuracy of 62%, representing the best performance among the three models. Its advantage stems from averaging multiple trees trained on random subsamples, which reduces variance and improves generalization. The detailed metrics were 65.2% precision, 57.1% recall, and an F1-Score of 0.606. An AUC of 0.682 means that the model has a 68.2% probability of correctly ranking a randomly selected student who failed a course above a randomly selected student who did not. Some individual folds reached values as high as 0.78, which is expected when working with survey data and a moderate sample size. KNN achieved 56.8% accuracy, 58.6% precision, 53.2% recall, an F1-Score of 0.557, and a ROC AUC of 0.593. Its performance was the lowest among the three models. Although PCA reduced the 67 original variables to 52 components, KNN remained sensitive to the presence of components with limited class separability, which affected its ability to identify meaningful nearest neighbors. The Decision Tree produced the lowest values in nearly all metrics: 58.1% accuracy, 59.8% precision, 56.4% recall, an F1-Score of 0.579, and a ROC AUC of 0.581. The most likely explanation is overfitting, as deep individual trees tend to memorize patterns specific to the training data that do not generalize effectively, and cross-validation was not sufficient to fully mitigate this effect.

Figure 2 presents the comparison among the three models. Random Forest outperformed the other two across all five metrics, particularly in ROC AUC (0.682) and precision (65.2%). KNN ranked second, whereas the Decision Tree produced the lowest values, confirming its tendency to overfit in high-dimensional settings.

Figure 3 presents the confusion matrices for Random Forest across the cross-validation folds. The model maintains a balance between correct and incorrect classifications, although some variation is observed, reflecting the heterogeneity of the survey data. The consistency across folds suggests that the model exhibits stable performance.

Figure 4 shows that the Random Forest AUC of 0.682 represents the average across the five folds, although some individual folds reached values as high as 0.78. Such variation is expected in cross-validation when working with survey data and a sample of 300 observations.

5.2. Feature Importance (Random Forest)

Among the 20 most influential variables, having a "Very High" concentration level emerged as the strongest protective factor. This was followed by using a mobile phone for 7 to 12 hours per day, having WhatsApp as the primary social media platform, the number of hours devoted to studying, and the use of

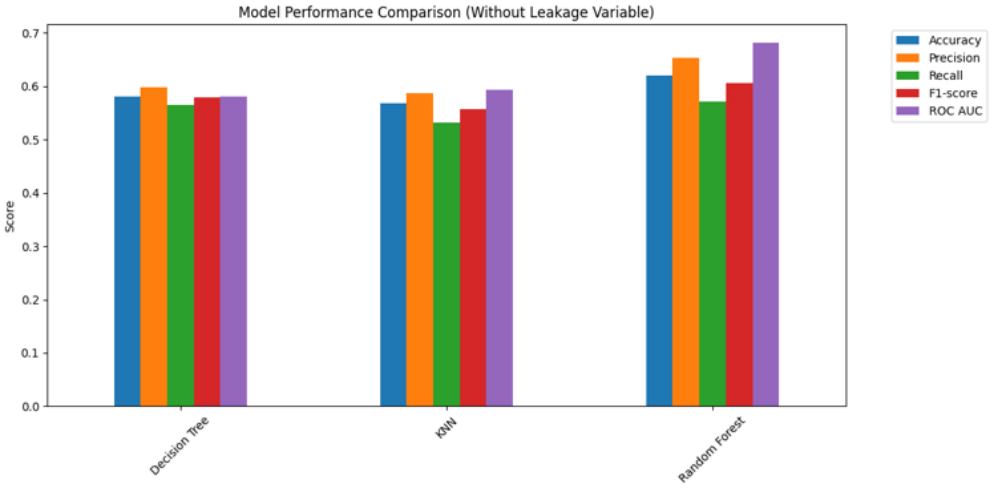


Figure 2. Comparison of metrics for supervised models (5-fold CV with SMOTE).

Figure 3. Confusion Matrices by Fold (Random Forest).

social media for entertainment purposes. To obtain these importance scores, the Random Forest model [15] was retrained on the 300 observations and the 67 original variables using Gini impurity reduction.

Figure 5 presents the hierarchical relevance of the predictor variables in the Random Forest model and provides a finding of substantial heuristic value: the qualitative, attitudinal, and perceptual dimensions of technology use exhibit significantly greater discriminatory power than chronological indicators or measures based solely on usage volume. The three variables with the greatest weight in the calculation of Gini impurity correspond to Likert-scale items associated with the self-perceived benefits of digital technologies: time saved on academic tasks (`likert_ahorro_n`), perceived improvement in academic performance (`likert_rendimiento_n`), and enhanced understanding of curricular content (`likert_comprension_n`). Statistically, the prominence of these three variables suggests that habituation and the resulting risk of inattention leading to course failure are strongly mediated by the construct of "perceived usefulness." This finding is consistent with the classical assumptions of Technology Acceptance Models (TAM), according to which habits are shaped not by the tool itself, but by the functional value attributed to it by the student.

It is equally revealing to examine the variables located in the middle and lower segments of the hierarchy. Traditional characteristics such as Current Semester and Socioeconomic Stratum, which have historically been regarded as structural determinants of academic failure in the conventional literature, appear here to be subordinate to dynamic behavioral variables within the digital ecosystem. These include self-reported concentration levels (`Concentration_Low` and `Concentration_Medium`) and the frequency of distraction in the classroom or during independent study sessions. This ranking highlights a paradigm shift in contemporary learning analytics: fixed socioeconomic and demographic factors provide the broader contextual framework, whereas patterns of daily interaction with mobile devices and the associated vulnerability of attentional control operate as more immediate triggers of adverse academic performance.

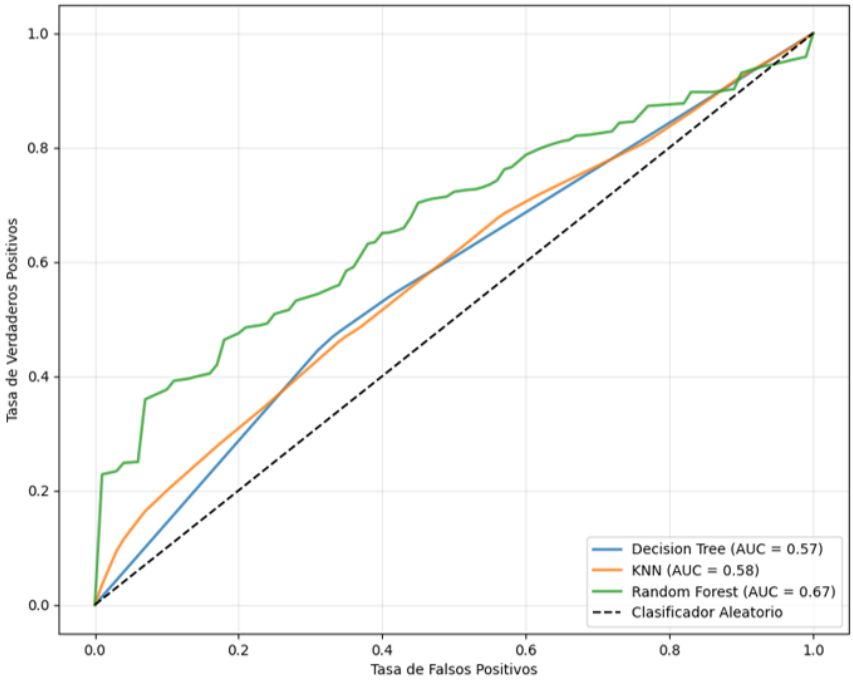


Figure 4. Average ROC curves for each model.

Finally, the relatively low importance of variables that directly quantify connection time, such as the total number of hours spent on a mobile phone or on specific platforms, reinforces the central argument of this study. The Random Forest model independently indicates that chronological intensity is not the critical risk factor. A student may accumulate a high number of screen hours without compromising their academic trajectory, provided that a clear separation between digital and academic environments is maintained. The true source of vulnerability associated with course failure lies in cognitive conflict, manifested as an inability to regulate attention when externally generated notification flows compete directly with working-memory processes during learning activities.

6. Discussion

The Random Forest model achieved a ROC AUC of 0.682 and an accuracy of 62%. These results indicate that course failure can be predicted to a certain extent using the variables measured in this study, although a substantial proportion of the variance remains unexplained. This is unsurprising, as course failure depends on factors that are not adequately captured through surveys, including student motivation, teaching quality, and personal circumstances that may vary from one academic term to another.

Among all the variables analyzed, concentration during study emerged as the strongest predictor. Students who reported a "Very High" level of concentration were less likely to fail courses. Research on self-regulated learning [10] and cognitive load theory supports the finding that maintaining attention during study acts as a protective factor, probably because it reflects more developed metacognitive skills [1]. Similarly, the authors in [10] identified related factors, such as study hours and device use, as predictors of academic risk among students. In this regard, the quality of the time devoted to studying appears to be more decisive than the number of hours alone. This finding calls on faculty members to design highly engaging learning environments that strengthen students' involvement in class and promote sustained attention to the subject matter [19].

Using a mobile phone for 7 to 12 hours per day ranked as the second most important predictor. This finding suggests the existence of a threshold beyond which mobile-device use begins to interfere with academic performance. Spending one or two hours on a mobile phone is not equivalent to spending nearly

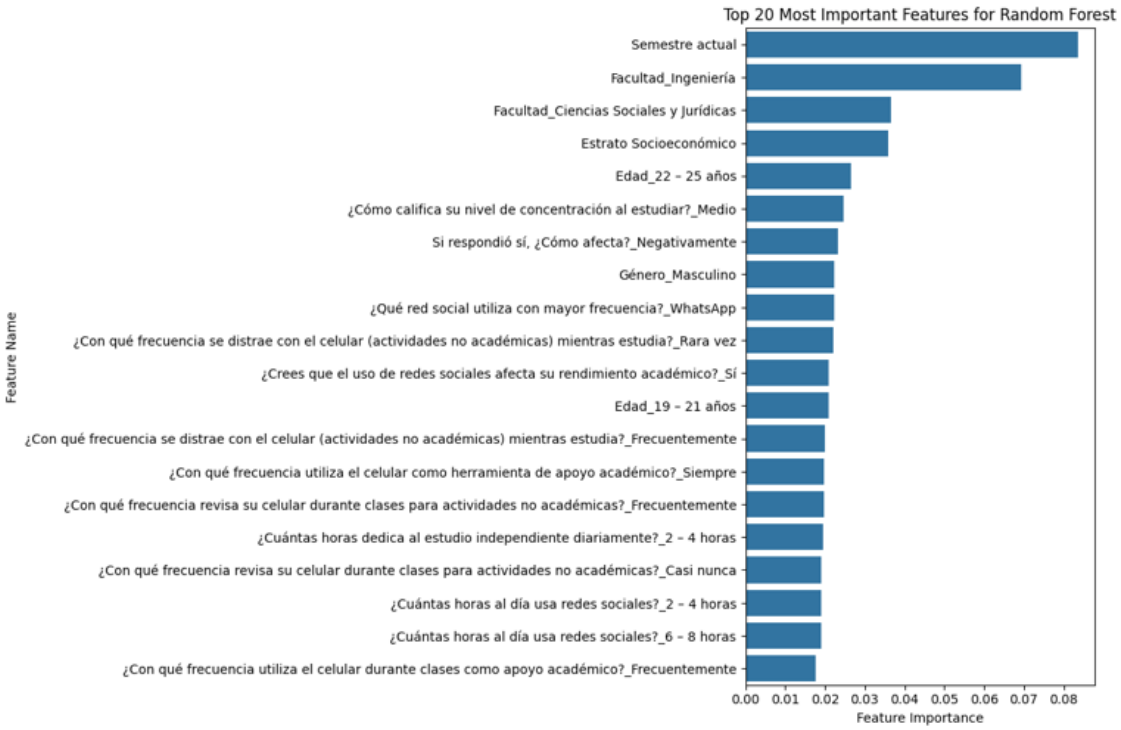


Figure 5. Top 20 most important variables according to Random Forest.

half of the day using it. The attention economy framework [4] helps explain this pattern: digital platforms are designed to sustain user engagement, and when mobile-phone use exceeds seven hours per day, the time available for studying may be considerably reduced.

Random Forest outperformed both the Decision Tree and KNN across all evaluation metrics. Ensemble methods are better suited to survey data characterized by high dimensionality and nonlinear relationships, as demonstrated by the model comparison. The Decision Tree exhibited overfitting, whereas KNN, even after dimensionality reduction through PCA, was unable to overcome the algorithm's inherent limitations for this type of data. Incorporating SMOTE into the pipelines was necessary to prevent the models from overlooking students who had failed at least one course.

A ROC AUC of 0.682 may appear low when compared with standards in fields such as computer vision or medicine, where values above 0.90 are often sought. However, in the social sciences and education, where data frequently rely on self-reports and the phenomena under study are inherently multifactorial, values between 0.60 and 0.70 may still be considered useful and meaningful [19], [5]. The model is not intended to replace instructors' professional judgment, but rather to serve as an early-warning tool that helps direct academic support resources toward the students who need them most.

7. Limitations

This study has several limitations that should be considered when interpreting its findings. The first concerns the cross-sectional design of the dataset. The data were collected at a single point in time during the 2026-1 academic term, making it impossible to capture the evolution of digital behaviors and their effects over time. Course failure is a process that develops throughout the semester, and assessing it at only one point does not reveal how patterns of mobile-phone use, concentration, or study strategies change as the academic term progresses.

The second limitation is the reliance on self-reported data. All study variables, including hours of mobile-phone use, concentration level, and preferred social media platforms, were reported by the students themselves. This introduces potential social desirability bias and recall errors. A student may underestimate

the number of hours spent using a mobile phone or overestimate their concentration level, not necessarily because of intentional misreporting, but because self-perception does not always correspond to actual behavior.

The third limitation is that the sample was drawn from a single institution in Riohacha. The University of La Guajira serves a population with specific socioeconomic and cultural characteristics, and patterns of digital technology use may differ across other Colombian or Latin American university contexts. This limits the generalizability of the findings.

Finally, although PCA was applied to reduce the 67 original variables to 52 components, the sample of 300 observations remains modest. With larger datasets, the models might capture more complex relationships and achieve better predictive performance. In addition, interpretability is reduced when working with PCA components rather than the original variables.

8. Conclusions

An AUC of 0.682 and an accuracy of 62% may appear modest when compared with the performance expected in fields such as computer vision, where AUC values often exceed 0.90. However, data in education and the social sciences differ substantially: they are frequently derived from self-reports, exhibit high variability, and reflect phenomena that are inherently multifactorial. In the literature on Machine Learning applied to education, models with AUC values between 0.60 and 0.70 are considered useful. A 68.2% probability of correctly distinguishing a student at risk means that the model may be used as an early-warning tool not as an automated decision-making system, but as a support mechanism that enables the university to provide tutoring before academic delay becomes difficult to overcome.

The results indicate that course failure can be predicted with 62% accuracy and a ROC AUC of 0.682 using Random Forest with SMOTE and cross-validation. The model outperformed both the Decision Tree and KNN across all evaluation metrics, which is consistent with the view that ensemble methods are better suited to high-dimensional survey data. The most influential variables (concentration, hours of mobile-phone use, and preferred social media platform) provide concrete insights for the design of intervention programs.

The following directions are proposed for future research: (a) expanding the sample by replicating the study at other universities in the Colombian Caribbean region to assess the generalizability of the models; (b) exploring algorithms with greater representational capacity, such as XGBoost, an advanced sequential ensemble method; LightGBM, an efficient gradient-boosting framework; and deep neural networks, which use multiple layers of artificial neurons to learn highly complex transformations and may capture more subtle patterns; (c) incorporating longitudinal data through cohort studies that follow the same students over time, making it possible to model academic trajectories and behavioral changes; (d) integrating data from institutional learning management systems, such as Moodle or Blackboard, including content access, forum participation, assignment submission, and response-time records, to enrich the characterization of academic behavior; and (e) developing an early-warning system integrated into the institutional platform that deploys the Random Forest model in production and retrains it periodically, for example, every academic term to maintain predictive accuracy as student behavior patterns evolve.

Author Contributions: **Andrés Solano-Barliza** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration.

Inés Figueroa-Sarmiento Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.

Fabio Romero-Gómez Validation, Investigation, Resources, Data curation, Writing – original draft.

All authors have read and agreed to the published version of the manuscript. Please refer to the CRediT taxonomy for the definitions of the terms. Authorship should be limited to those who have made substantial contributions to the reported work.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable, since the present study does not involvehuman personnel or animals.

Informed Consent Statement: This study is limited to the use of technological resources, so nohuman personnel or animals are involved.

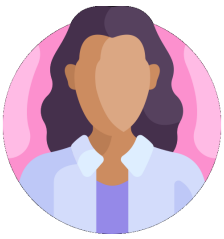
Conflicts of Interest: Under the authorship of this research, it is declared that there is no conflict of interest with the present research.

References

1. Al-Alawi, L., Al Shaqsi, J., Tarhini, A., and Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: The case of college students on academic probation. *Education and Information Technologies*, 28(10):12407–12432.
2. Al Mamun, A. et al. (2024). Impact of the use of social media among university students using machine learning. In *Communication and Intelligent Systems. ICCIS 2023*, Lecture Notes in Networks and Systems. Springer.
3. Barliza, A. S., Gómez, I. C., Caballero, J. M., and Muñoz, S. V. (2025). Towards a national artificial intelligence policy in colombia: A comparative analysis of international frameworks. *OnBoard Knowledge Journal*, pages 1–13.
4. Bhargava, V. R. and Velasquez, M. (2021). Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly*, 31(3):321–359.
5. Bowers, A. J. and Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, 24(1):20–46.
6. Chapman, P. et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Technical report, SPSS.
7. Contreras, L., Fuentes, H., and Rodríguez, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, 13(5):233–246.
8. Gil-Vera, V. D. and Quintero-López, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información Tecnológica*, 32(6):221–228.
9. Goldhaber, M. H. (1997). The attention economy and the net. *First Monday*, 2(4).
10. Gordon, M. S. and Ohannessian, C. M. (2024). Social media use and academic achievement among early adolescents. *Youth & Society*.
11. Gutiérrez, A. A. D., Vélez Díaz, J. F., and López, J. M. (2021). Indicadores de deserción universitaria y factores asociados. *EducaT: Educación Virtual, Innovación y Tecnologías*.
12. Hemal, S. H., Khan, M. A. R., Ahammad, I., et al. (2024a). Predicting the impact of internet usage on students' academic performance using machine learning techniques in bangladesh perspective. *Social Network Analysis and Mining*, 14:66.
13. Hemal, S. H., Khan, M. A. R., Ahammad, I., Rahman, M., Khan, M. A. S., and Ejaz, S. (2024b). Predicting the impact of internet usage on students' academic performance using machine learning techniques in bangladesh perspective. *Social Network Analysis and Mining*, 14(1):66.
14. López-García, A., Blasco-Blasco, O., Liern-García, M., and Parada-Rico, S. E. (2023). Early detection of students' failure using machine learning techniques. *Operations Research Perspectives*, 11:100292.
15. Martelo, R. et al. (2017). Incidencia de las redes sociales en el rendimiento académico de los estudiantes de la universidad. *Revista Espacios*, 38(51).
16. Martinez, R. et al. (2019). Use of machine learning to measure the influence of behavioral and personality factors on academic performance. *IEEE Latin America Transactions*, 17(3).
17. Nti, I. K. et al. (2022). Prediction of social media effects on students' academic performance using machine learning algorithms (MLAs). *Journal of Computers in Education*, 9(2):195–223.
18. Ramzan, M. et al. (2025). From social media reactions to grades: A machine learning-based SocialNet analysis for academic performance prediction. *EAI Endorsed Transactions on AI and Robotics*.
19. Solano-Barliza, A. D. (2023). Enseñanza de la analítica de datos usando aprendizaje basado en proyectos colaborativos: Teaching data analytics using collaborative project-based learning. *Formación Universitaria*, 16(6):23–32.
20. Solano-Barliza, A. D. (2024). Quantitative analysis of the perception of using ChatGPT artificial intelligence in the teaching and learning of colombian-caribbean undergraduate students. *Formación Universitaria*, 17(3):129–138.
21. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.
22. Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4):295–312.
23. Xu, X., Wang, J., Peng, H., and Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, 98:166–173.

24. Yu, H., Yang, W., Xu, N., and Du, Y. (2021). Advertising strategy and contract coordination for a supply chain system: Immediate and delayed effects. *Kybernetes*, 52(1):235–261.

Authors' Biography



Inés Figueroa-Sarmiento is a Systems Engineering student at the Universidad de la Guajira. Her academic interests focus on software development, data analytics, machine learning, and their applications in education. She has participated in research projects related to information technologies and data science.

Fabio Romero-Gómez is a Systems Engineering student at the Universidad de la Guajira. His interests include the development of computing solutions, data mining, machine learning, and artificial intelligence applications for solving problems in education and engineering. He has participated in academic and research projects in these areas.

Andrés Solano-Barliza holds a Ph.D. in Information and Communication Technologies (ICT) from Universidad de la Costa, Colombia, and a Ph.D. in Computer Science and Security Mathematics from Universitat Rovira i Virgili, Spain. He earned a B.Sc. in Systems Engineering from the Universidad de la Guajira, Colombia, in 2014; a postgraduate specialization in Analytics and Big Data from Corporación Universitaria Iberoamericana, Colombia, in 2023; and a master's degree in ICT Pedagogy from the Universidad de la Guajira, Colombia, in 2019. His research interests include recommender systems, data analytics and Big Data, machine learning, artificial intelligence, pedagogy, and the educational use of information and communication technologies. He is currently a full-time professor in the Faculty of Engineering at the Universidad de la Guajira, Colombia.

Disclaimer/Editor's Note: Statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and not of the OnBoard Knowledge Journal and/or the editor(s), disclaiming any responsibility for any injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content.