Systems and Soft Computing 7 (2025) 200352

Contents lists available at ScienceDirect

Systems and Soft Computing

journal homepage: www.journals.elsevier.com/soft-computing-letters

# Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors

Reymark D. Deleña [a,*], Norniña J. Dia [b], Redeemtor R. Sacayan [c], Joseph C. Sieras [b], Suhaina A. Khalid [b], Amer Hussien T. Macatotong [b], Sacaria B. Gulam [d]

[a] Faculty, Department of Information Technology, College of Computer Studies, Mindanao State University – Iligan Institute of Technology, Iligan City 9200, Philippines
[b] Faculty, Department of Information Sciences, College of Information and Computing Sciences, Mindanao State University – Main Campus, Marawi City, Lanao Del Sur, 9700, Philippines
[c] Faculty, Department of Mathematics and Statistics, College of Science and Mathematics, Mindanao State University-Iligan Institute of Technology, Iligan City 9200, Philippines
[d] Faculty, Department of Computing Sciences, College of Information and Computing Sciences, Mindanao State University – Main Campus, Marawi City, Lanao Del Sur, 9700, Philippines

ARTICLE INFO

*Keywords:*
Student retention
Educational data mining
Machine learning algorithms
Data visualization
Higher Education Analytics
Data-driven Education
Sociodemographic Factors
Academic Factors
eXtreme Gradient Boosting

ABSTRACT

Student dropout remains a persistent challenge in higher education, particularly in developing regions where institutional resources for intervention are limited. This study explores the predictive potential of machine learning (ML) algorithms in identifying students at risk of dropping out using historical academic and socio-demographic data from Mindanao State University–Main Campus, covering a ten-year period (2012–2022). A total of 482 student records and 146 variables were preprocessed using Power BI and prepared via the CRISP-DM methodology before being modeled in Jupyter Notebook. Ten ML algorithms such asExtreme Gradient Boosting (XGBoost), Gradient Boosting (GB), Artificial Neural Network (ANN), Decision Tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB) were evaluated using both train-test split and 5-fold cross-validation to ensure robustness and generalizability. Results indicate that XGBoost outperformed all other models, achieving the highest cross-validated accuracy (90.66 %), F1 Score (90.72), and one of the lowest error values (Mean Square Error (MSE) = 9.34, Log Loss = 0.26). GB and ANN followed closely, demonstrating strong balance between precision, recall, and low misclassification rates. While models like Naïve Bayes recorded high recall, they also produced excessive false positives, limiting their practical use. The study offers a scalable and transferable modeling framework for higher education institutions seeking to implement early warning systems. It also emphasizes the pedagogical value of integrating educational data science into Information Technology Education (ITE) curricula to foster real-world AI application. Limitations and future directions are discussed, particularly regarding behavioral data integration and model interpretability.

## 1. Introduction

Student attrition in university settings has long been a pressing concern among educators and administrators, as it significantly affects institutional rankings, reputation, workforce readiness, and financial sustainability [1,2]. Globally, high dropout rates are associated with wasted educational resources, decreased graduation outputs, and weakened national educational goals. In many developing and under-represented contexts, particularly in Southeast Asia, challenges related to academic persistence are compounded by socioeconomic disparities and infrastructure limitations. A wide range of factors contribute to a student's decision to withdraw from a program, but academic performance and sociodemographic characteristics remain the most frequently examined variables in retention studies [3,4].

In response to this challenge, Educational Data Mining (EDM) has emerged as a key area within educational research, focusing on the application of machine learning techniques to institutional data in order to predict student outcomes and inform strategic interventions [5]. Numerous studies have established that academic records and demographic profiles are strong indicators of student persistence or risk of

* Corresponding author.
*E-mail address:* reymark.delena@g.msuiit.edu.ph (R.D. Deleña).

https://doi.org/10.1016/j.sasc.2025.200352
Received 14 July 2024; Received in revised form 15 June 2025; Accepted 14 July 2025
Available online 18 July 2025

dropout [3,4,6]. The integration of EDM and machine learning into university systems allows institutions—both locally and globally—to analyze historical student data to identify learning gaps, anticipate academic failures, and predict future academic trajectories [7]. These predictive insights can guide data-informed decisions aimed at improving institutional effectiveness and creating supportive learning environments [8], ultimately helping students stay on track toward timely graduation.

The College of Information and Computing Sciences (CICS) of Mindanao State University – Main Campus has one of the highest attrition rates within the university, which piqued the researcher's interest in this subject matter. However, the significance of this research extends far beyond the boundaries of one institution. As higher education institutions worldwide contend with increased student diversity, massification of enrollment, and post-pandemic educational disruptions, the ability to accurately model and address student attrition becomes a global imperative. Retention metrics are now pivotal not just for institutional planning but also for national education performance benchmarks, accreditation compliance, and international university rankings.

Moreover, diversity and inclusion goals, particularly disaggregated retention statistics by gender, socio-economic status, and ethnicity, have gained prominence as critical indicators of how well institutions support marginalized student populations. These data-driven approaches offer powerful tools to inform not only curriculum design and student services but also equitable resource distribution and strategic policy development. On a global scale, retention and graduation rates are among the most scrutinized indicators by ranking bodies and education quality assurance agencies, making predictive analytics in this domain both timely and globally significant.

Recent studies apply diverse predictive models for predicting academic performance, graduation, and dropout [9], enabled by the increasing availability of institutional datasets. As model performance depends heavily on dataset quality and size [10], large-scale research such as that by Alhazmi & Sheneamer [11] (275,000 students) and Rodríguez-Hernández et al. [12] (162,030 students) reinforces the reliability of data-driven modeling. Several others, Beckham et al. [13], Niyogisubizo et al. [14], Ghorbani & Ghousi [15], Marbouti et al. [16], Lam et al. [17], and Gonzalez-Nucamendi et al. [18], explore the predictive power of various machine learning models, including hybrid and ensemble approaches, in student dropout forecasting across countries such as Portugal, Colombia, Iran, and Mexico. These global efforts underscore the importance of localized yet scalable modeling approaches, especially for contexts with high dropout risk but low representation in existing literature.

This study makes several key contributions to the field of student retention analytics. First, it integrates a comprehensive set of socio-demographic and academic variables to develop predictive models for student retention within the context of a state university in the Philippines, an underrepresented setting in existing literature. Second, it conducts a rigorous comparative evaluation of ten (10) widely adopted machine learning (ML) algorithms, identifying Extreme Gradient Boosting (XGBoost) as the most accurate across multiple standardized metrics. Third, the study leverages actual student data spanning a full decade (2012–2022) from the College of Information and Computing Sciences (CICS) at Mindanao State University–Main Campus, encompassing its Computer Science, Information Technology, and Information Systems programs. This longitudinal dataset enables deeper insights into student behavior and retention trends over time, supporting more effective early intervention strategies.

The novelty of this research lies in its context-specific focus, breadth of model comparison, and practical implementation using institutional data visualized through Python in a Jupyter Notebook environment. While the effectiveness of ML in predicting student performance is well-established, much of the existing literature relies on short-term datasets, simulated environments, or limited algorithm selection. Some studies also emphasize hybrid model development without conducting systematic benchmarking. In contrast, this study addresses these gaps by employing ten ML models evaluated using six performance metrics such as Accuracy, F1 Score, Precision, Recall, Mean Squared Error (MSE), log loss, and by conducting a thorough review to identify the most relevant data inputs and modeling strategies.

Ultimately, the goal of this research is not only to achieve high predictive performance but also to provide actionable insights that inform institutional policy, guide academic interventions, and support program-level strategies to reduce student attrition and improve timely graduation rates. By contributing findings from a region with limited representation in literature, this study aims to support more inclusive and globally relevant approaches to student success analytics.

## 2. Material and methods

In conducting the study, secondary data of CICS students was sourced from the Information and Communication Technology Center (ICTC), following the guidelines established by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The data underwent the CRISP-DM process. First, an in-depth analysis of the business objectives and needs was done. Next, the quality of gathered data was examined and cleaned. Subsequently, the data was organized for modeling. Ten (10) data modeling techniques were applied to predict the outcomes which were then evaluated using six (6) evaluation metrics to identify the optimal model that provides the most accurate results.

### 2.1. Systematic review

#### 2.1.1. Data collection and extraction

For the systematic review, two (2) journal databases were used to find the recent and relevant articles necessary to conduct the systematic review. To find the relevant papers, the researcher used the keywords "Data mining" AND "Student retention" OR "Student performance" in Science Direct and "Predicting student retention" + "data mining" in IEEE Xplore. Furthermore, the search was filtered to restrict the publication year between 2013 and 2023 spanning ten (10) years; this was to ensure that all selected papers are relevant and up to date. The query resulted in a total of 364 journals (see Table 1). Subsequently, MS Excel was used to organize, sort, and eliminate duplicates from the retrieved data set

#### 2.1.2. PRISMA model

In this process as illustrated in Fig. 1, 336 records were excluded in the screening process due to its irrelevancy based on their title and abstract. The remaining articles were then assessed for eligibility based on the criteria listed in Table 2 by reviewing them in full text, and this assessment retained 21 studies listed in Table 3 .

#### 2.1.3. Key findings

Fig. 2a illustrates the frequency with which various sociodemographic factors were utilized across a set of studies. The data reveals that gender (9 studies) was the most frequently used factor, followed closely by marital status (8 studies) and age (7 studies). These results suggest that these variables are considered highly relevant in research exploring

**Table 1**
Databases and keywords used.

| No. | Database | URL | Retrieved Papers |
|---|---|---|---|
| 1 | Elseiver (Science Direct) | https://www.sciencedirect.com/ | 277 |
| 2 | IEEE Xplore Digital Library | https://www.ieeexplore.ieee.org/ | 87 |
| | | **Total** | **364** |

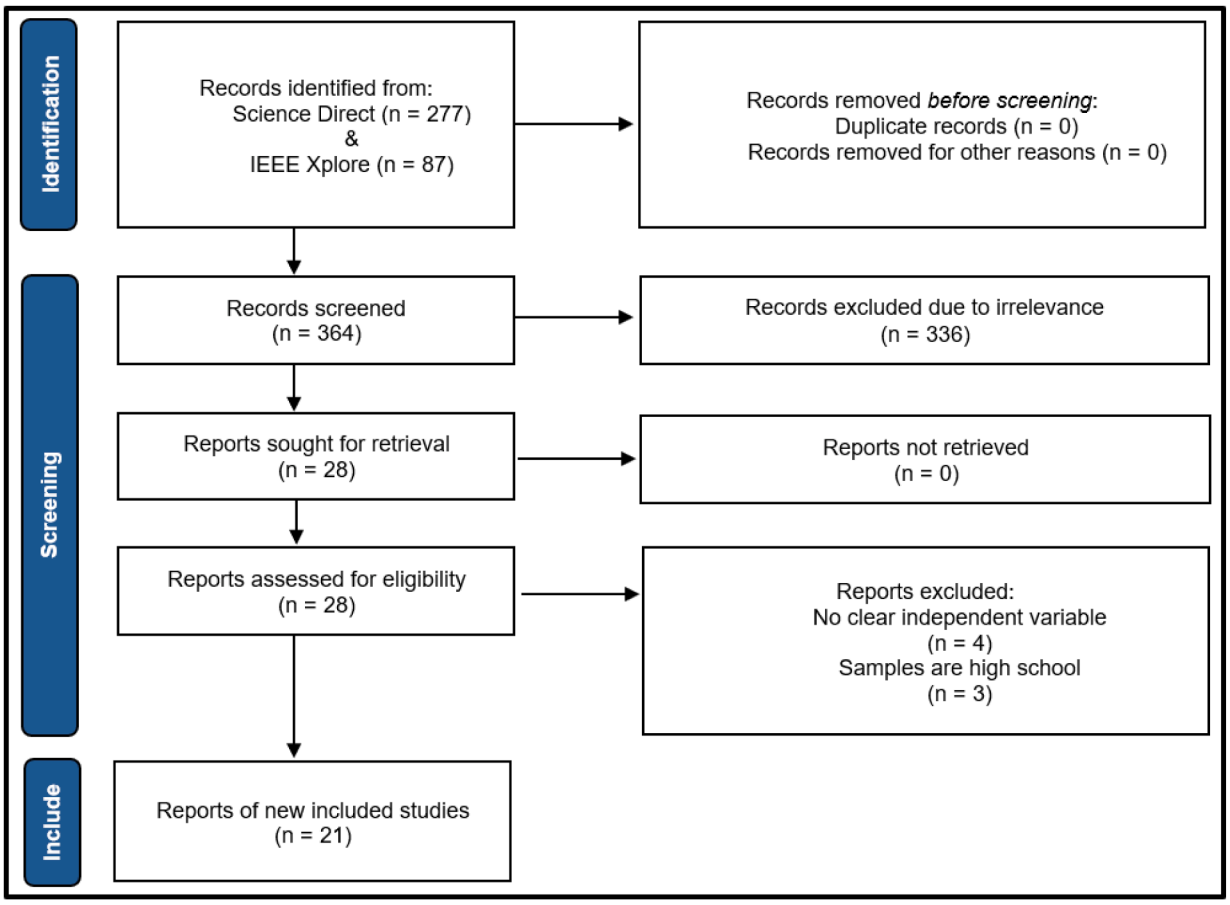



**Fig. 1.** PRISMA flow diagram illustrating the systematic review process for the selection of studies related to data mining and student retention/performance.

**Table 2**
Inclusion and exclusion criteria.

| No. | Inclusion Criteria | Exclusion Criteria |
|---|---|---|
| 1 | Using algorithms to predict student retention | Studies not utilizing predictive analytics |
| 2 | Articles between 2013 and 2023 | Articles older than 2013 |
| 3 | Sample population is college students | Sample population other than college students |
| 4 | Journal, Conference, Thesis/ Dissertation | Review articles, book chapters, etc. |
| 5 | Written in English Language | Not written in English Language |
| 6 | Open access | Paid |

social or behavioral outcomes. Other moderately used factors include parent's occupation (4 studies), family size and financial support (3 studies each), and guardian and parent's education/qualification (2 studies each). Several factors, such as address, family responsibilities, religious affiliation, parent's school issues, school, travel time to school, and working status, were only used in a single study, indicating limited application or relevance in most contexts analyzed. Further, this variation in usage frequency highlights a trend in which basic demographic variables like gender, age, and marital status are prioritized, possibly due to their availability, perceived impact, or ease of interpretation in predictive modeling and statistical analyses.

On the other hand, Fig. 2b depicts presents the frequency with which specific academic-related variables were included across different research studies. The most used academic factor was expected graduation year ($n = 8$), followed by grades or GPA ($n = 7$) and project grades ($n = 7$). These variables likely reflect students' academic performance and trajectory, making them strong predictors in educational analytics. Other academic factors with moderate usage include admission test score ($n = 4$), dropout status ($n = 3$), and high school GPA and year level ($n = 3$ each). Several variables such as enrollment status, student classification, socioemotional support, transfer status, and middle school performance were only used in one or two studies, indicating lower prevalence or relevance depending on the study's objectives. Moreover, the variety in factor usage suggests that while some indicators like grades and graduation timelines are central to academic studies, others are context-dependent and possibly underexplored. This also highlights an opportunity to investigate the value of less commonly used academic variables in future predictive or evaluative studies.

Fig. 2c shows the frequency of usage of various machine learning (ML) models in the studies, distinguishing between supervised and unsupervised learning methods. It is evident that Decision Tree ($n = 8$), Multilayer Perceptron (MLP) ($n = 8$), and Support Vector Machine (SVM) ($n = 8$) are the most widely used supervised learning models, reflecting their popularity due to interpretability, accuracy, and robustness in classification tasks. Other frequently applied models include Logistic Regression ($n = 6$) and K-Nearest Neighbor (KNN) ($n = 5$). These classical models are still highly valued for their simplicity and effectiveness. Less frequently used models, each appearing in just one study, include Bayesian Network, Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM), among others. Notably, only a few models were classified under unsupervised learning (in gray), such as K-Means Clustering ($n = 2$) and J48/JRip, indicating a predominant focus on supervised techniques in the reviewed studies. This highlights a potential research gap where unsupervised or hybrid learning approaches could be further explored.

Fig. 2d illustrates the frequency of various evaluation metrics used across studies related to predictive modeling. Accuracy emerges as the

**Table 3**
Eligible articles that employ data mining on student retention.

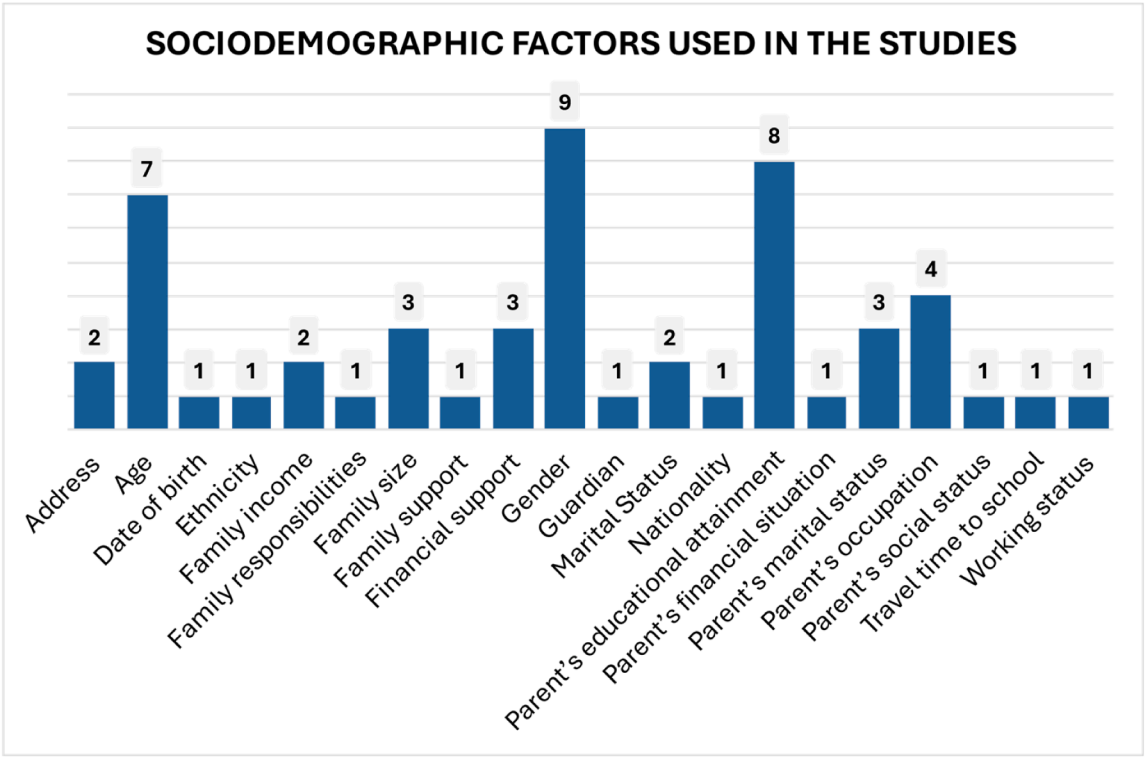
| No. | Author | Year | Article | Prediction Goal | Dataset Size | No. of Variables | Algorithm (s) | Ref. |
|---|---|---|---|---|---|---|---|---|
| 1 | *(Alhazmi & Sheneamer)* | 2023 | Early predicting of student performance in higher education | Student's performance (CGPA) | 275,000 | 16 | XGBoost, LR, SVM, KNN, RF | [11] |
| 2 | *(Beckham* et al.*)* | 2023 | Determining factors that affect student performance using various machine learning method | Factors affecting student performance (Correlation Score) | 395 | 13 | MLP, DT, RF | [13] |
| 3 | *(Alwarthan* et al.*)* | 2022 | An explainable model for identifying at-risk student at higher education | At-risk Students (CGPA) | N/A | 7 | RF, ANN, SVM | [19] |
| 4 | *(Feng* et al.*)* | 2022 | Analysis and prediction of students' academic performance based on education data mining | Student's performance (Poor, Good, Excellent) | N/A | 1 | K-means Clustering | [20] |
| 5 | *(Mariano* et al.*)* | 2022 | Decision trees for predicting dropout in Engineering Course students in Brazil | Factors affecting student performance (Correlation score), and Dropout rate | 91 | 2 | DT | [21] |
| 6 | *(Niyogisubizo* et al.*)* | 2022 | Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization | Student dropout (Dropout or Non-dropout) | 216 | 8 | RF, XGBoost, GB, FNN | [14] |
| 7 | *(Singh* et al.*)* | 2022 | Predicting student-teachers dropout risk and early identification: A four-step logistic regression approach | At-risk Students (Dropout or Non-dropout) | 1723 | 8 | LR | [22] |
| 8 | *(Marbouti* et al.*)* | 2021 | Academic and demographic cluster analysis of engineering student success | Factors affecting student performance (P-Value) | 12,053 | 10 | K-means clustering | [23] |
| 9 | *(Nabil* et al.*)* | 2021 | Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks | Student's performance (Pass or Fail) | 4266 | 1 | DNN, DT, RF, GB, LR, SVC, KNN | [24] |
| 10 | *(Prabowo* et al.*)* | 2021 | Aggregating time series and tabular data in deep learning model for university students' GPA prediction | Student's performance (GPA) | 46,670 | 7 | MLP, LSTM | [25] |
| 11 | *(Rodríguez-Hernandez* et al.*)* | 2021 | Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation | Student's performance (Level to Level 4) | 16,2030 | 11 | MLP | [26] |
| 12 | *(Uliyan* et al.*)* | 2021 | Student retention (retention rate) | GPA, Subjects' grades | 2949 | 2 | BLSTM, CRF | [12] |
| 13 | *(Fernández-García* et al.*)* | 2020 | Creating a recommender system to support higher education students in the subject enrollment decision | Student's performance (Pass or Fail) | 6948 | 14 | RF, LR, DT, SVM, KNN, MLP, GB | [27] |
| 14 | *(Ghorbani & Ghousi)* | 2020 | Comparing different resampling methods in predicting students' performance using machine learning techniques | Algorithm Performance (Model Accuracy Scores) | 650 | 19 | RF, KNN, ANN, XGBoost, SVM, DT, LR, NB | [15] |
| 15 | *(Mengash)* | 2020 | Using data mining techniques to predict student performance to support decision making in university admission systems | Student's performance (Excellent, Very Good, Good, Acceptable, Poor) | 2039 | 2 | ANN, DT, SVM, NB | [28] |
| 16 | *(Cardona & Cudney)* | 2019 | Predicting Student Retention Using Support Vector Machines | Algorithm performance (Model accuracy score) | 904 | 6 | SVM | [29] |
| 17 | *(Viloria* et al.*)* | 2019 | Integration of data technology for analyzing university dropout | Student dropout (Dropout or Non-dropout) | 19,300 | 4 | Bayesian Network, DT, NN | [30] |
| 18 | *(Lesinski & Corns)* | 2018 | Multi-objective evolutionary neural network to predict graduation success at the United States Military Academy | Graduation Status (Graduates, Late Graduates, Non-graduates) | 5100 | 7 | MLP, MOEA | [31] |
| 19 | *(Lesinski,* et al.*)* | 2016 | Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy | Graduation Status (Graduates, Late graduates, Non-graduates) | 5100 | 7 | MLP | [32] |
| 20 | *(Marbouti* et al.*)* | 2016 | Models for early prediction of at-risk students in a course using standards-based grading | At-risk Students (Pass or Fail) | 1600 | 5 | NB Classifier, SVM, KNN, LR, DT, MLP | [16] |
| 21 | *(Goga* et al.*)* | 2015 | A recommender for improving the student academic performance | Student's performance (GPA) | 7500 | 10 | RF, Random Tree, J48, Decision Stump, REPTree, JRip, OneR, ZeroR, PART, Decision Table, MLP | [33] |

most employed metric, reported in 13 studies, highlighting its popularity due to its simplicity and intuitive interpretation. Following closely, F1 Score, Precision, and Recall are each used in 11 studies. These three metrics are especially important in imbalanced datasets where focusing solely on accuracy could be misleading. Other metrics such as MSE (used in 2 studies) and a range of others (each used once) — including ROC curve, R2, AUC, MAE, and RMSE — suggest that while diversity exists in evaluation practices, there is a clear concentration around a few key indicators. This trend may reflect a preference for metrics that balance interpretability with robustness, especially in classification tasks. Interestingly, statistical tests like the Hosmer & Lemeshow test, Wald test, and Likelihood Ratio Test, though rarely used (only once each), indicate occasional application of inferential statistical methods to complement predictive accuracy. This distribution provides insight into how performance is validated in student retention and academic prediction models, pointing to a need for more consistent use of varied and task-appropriate evaluation metrics to ensure balanced model assessment.

### 2.2. CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) (Fig. 3) is widely adopted, industry - independent process model for data mining, extensively used in both practical applications and in research [34]. This cycle is composed of six phases: (1) Business Understanding which focuses on the understanding of the project objectives and requirements from a business perspective; (2) Data Understanding or determining the data to be collected and discovering insights from it; (3) Data Preparation which encompasses all activities that result in the final dataset; (4) Modeling, where various modeling techniques are chosen

**( a )** Frequency of sociodemographic factors used in the studies.




**( b )** Frequency of academic factors used in the studies.

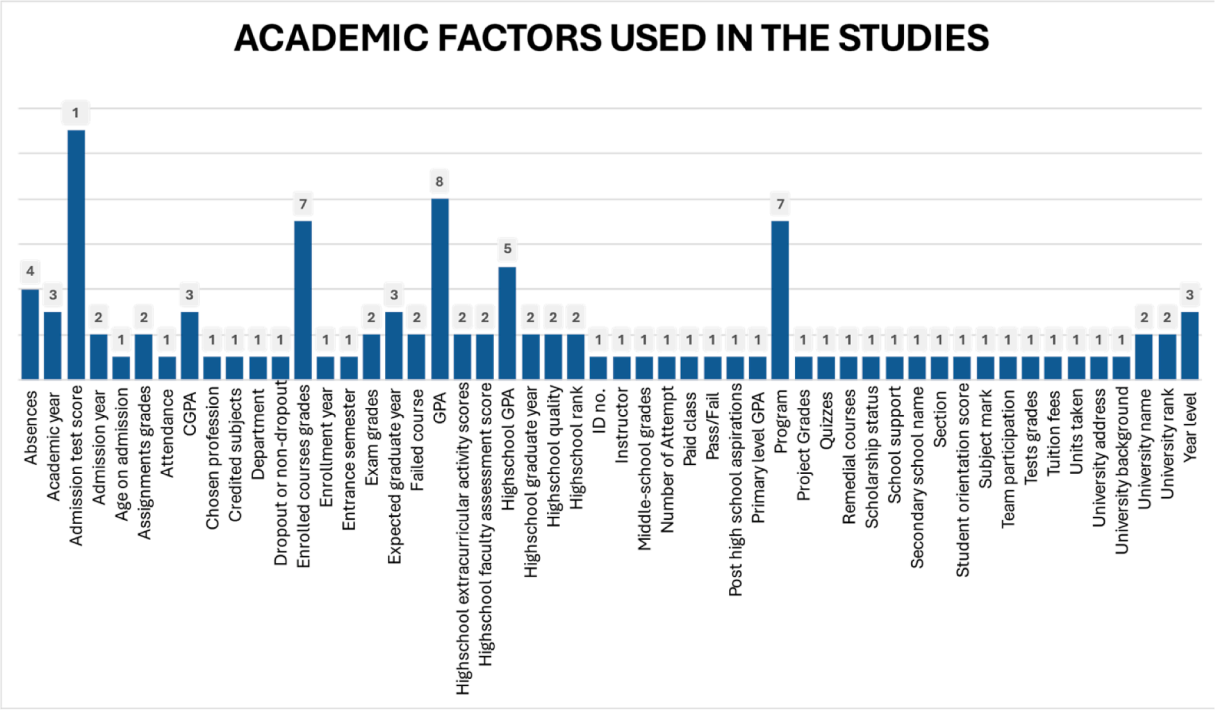
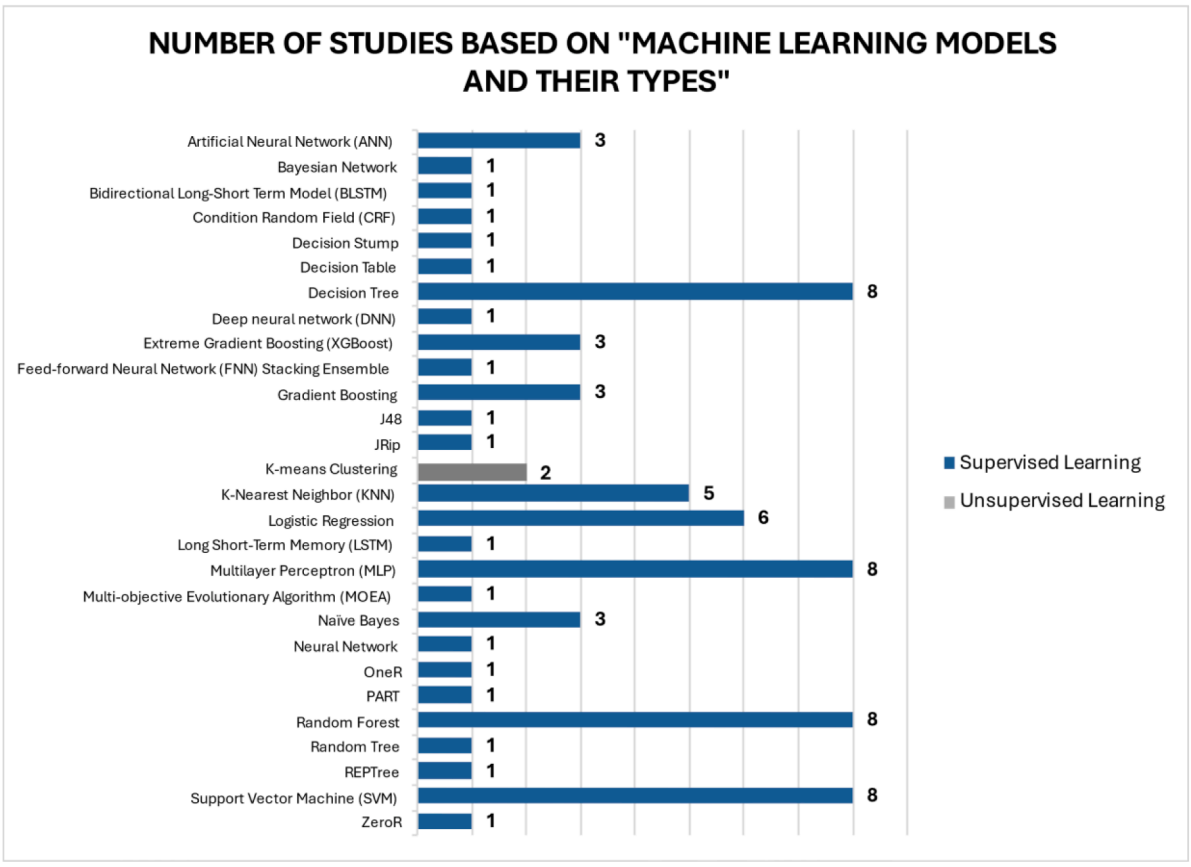



**Fig. 2.** Comprehensive summary of factors, models, and evaluation metrics used in student retention prediction studies. (a) Sociodemographic factors frequently included in predictive models, with gender, parent's educational attainment, and age among the most cited attributes. (b) Academic factors considered in prior studies, where high school GPA, failed courses, and dropout status were the most prominent predictors. (c) Machine learning models used in the literature, showing that decision tree, random forest, support vector machine (SVM), and multilayer perceptron (MLP) are the most employed, all of which fall under supervised learning. (d) Evaluation metrics applied in model assessment. Accuracy was the most frequently used metric, followed by F1 score, precision, and recall, reflecting a focus on classification performance in most studies.

**( c )** Number of studies based on machine learning models and their types.




**( d )** Distribution of Studies According to Evaluation Metrics Used in Predictive Modeling

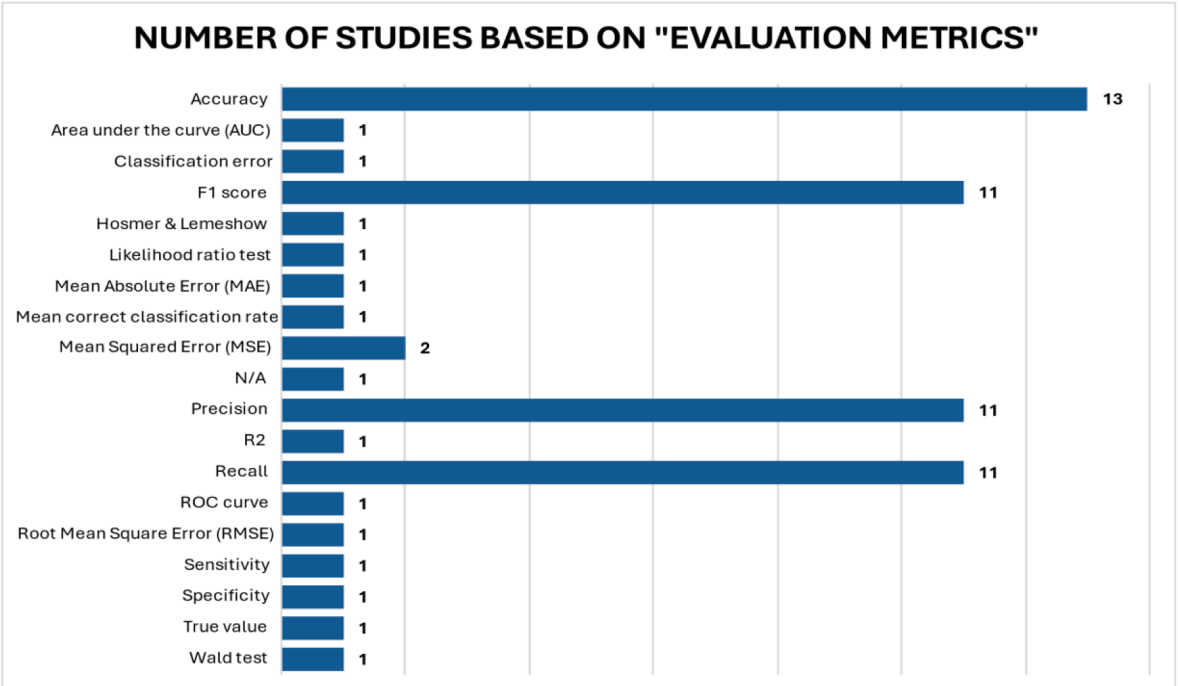



**Fig. 2.** (*continued*).

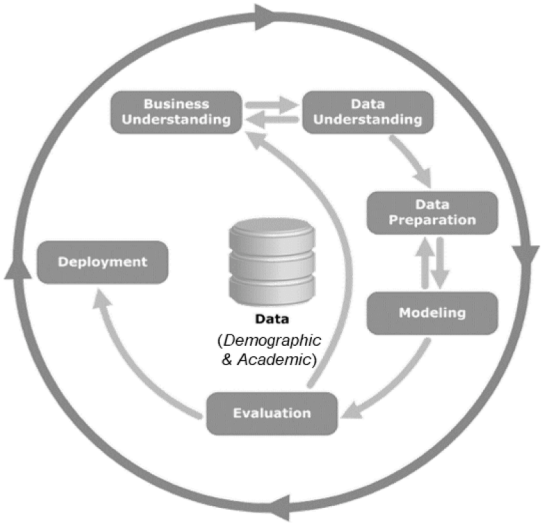



**Fig. 3.** The CRISP-DM (cross-industry standard process for data mining) methodology adopted in this study [44]. The framework consists of six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This cyclic approach ensured that the predictive modeling process remained aligned with institutional goals, data quality, and model performance throughout the project.

and applied, and their parameters are calibrated to optimal values; and (5) Evaluation, where algorithms are tested to find the optimal model [35,36].

Higher Education Institutions (HEIs) suffer intensely from students' withdrawal as they are highly interested and invested in producing competent students which could positively affect their reputation in the market. The College of Information and Computing Sciences (CICS) in MSU-Main Campus is not an exemption as most of the enrolled students took 5 years or more to pass and complete the program (shown in Fig. 4). Seemingly, half of the overall population from A.Y. 2012–2022 opted to withdraw from the course and shift to other college to graduate early. Therefore, having a strong understanding of the business aspects related to student retention is essential for educational leaders and administrators.

For data preparation stage, the data gathered from the Information and Communication Technology Center (ICTC) through the permission of the Office of the Vice Chancellor for Academic Affairs (OVCAA) of the Mindanao State University-Marawi (MSU-Main Campus) was inspected and cleaned where data was corrected, and null and erroneous values were imputed with zero (0). The multiple files were then integrated in a single file. Afterwards, the categorical data was transformed into numerical to perform the data analysis without difficulty, and the unnecessary attributes and period were excluded in the final dataset that was fed into the models used.

In the modeling stage, the ten (10) machine learning algorithms

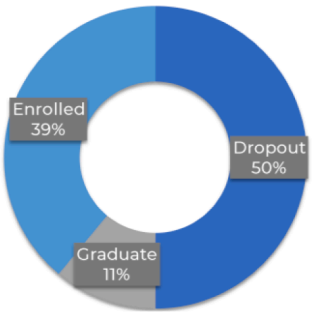



**Fig. 4.** Distribution of student academic status in the dataset.

selected were tested by setting the X and Y variables, then dividing the dataset into training and testing data. After that, the models were built for prediction using the test data.

In the evaluation stage, the algorithms were tested using the selected six (6) evaluation metrics to calculate the accuracy of the predicted values. The model with highest accuracy rate was the algorithm deployed for the final prediction of student's probability rate of withdrawal.

### 2.3. Machine learning algorithms

In predictive analytics, ML algorithms are essential tools for identifying trends and patterns in educational data, particularly in modeling student retention. These algorithms must first be trained on historical data to ensure accurate predictions and can be categorized into classification, ensemble, and neural network models. Classification algorithms such as SVM [37], Logistic Regression [22], KNN [38], Naïve Bayes [39], and Decision Tree [37] are widely used to label students as either at risk or likely to persist, each with varying assumptions and sensitivities to educational data structures. SVM and LR assume Independent and Identically Distributed (IID) data, though LR can be adapted using generalized estimating equations (GEE) to account for clustered errors. KNN is a non-parametric model sensitive to demographic clustering, while Naïve Bayes, despite its strong independence assumption, remains useful in large datasets when paired with feature selection. Decision Trees are intuitive and flexible but prone to overfitting without pruning. Ensemble models like Random Forest and Gradient Boosting [14] combine multiple learners to improve accuracy and generalization [18], with Random Forest offering simplicity, short training time, and improved forecasting accuracy, albeit with increased model complexity as more trees are added [19]. XGBoost, an optimized implementation of gradient boosting, has emerged as a leading model due to its speed, built-in regularization, early stopping, and ability to handle structured and nested data using parameters such as "group" [14]. The heterogeneity among ensemble learners helps mitigate bias and variance, making them robust in handling diverse student profiles [40].

Moreover, Neural network models are powerful algorithms inspired by the structure and function of the human brain. They process data through interconnected layers of artificial neurons, enabling the learning of non-linear relationships and adaptive weight adjustments. These models are particularly effective for complex prediction tasks in EDM environments [41]. One common variant is the ANN, which consists of an input layer, one or more hidden layers, and an output layer. Each artificial neuron receives inputs, computes weighted sums, applies activation functions, and passes outputs to the next layer. During training, these weights are adjusted iteratively to minimize prediction error. ANNs are widely applied in EDM because they can model intricate relationships between variables, even with small datasets, and capture subtle patterns in student data [28]. While standard ANN architectures assume IID data, educational datasets often include correlated or nested structures. To address this, methods such as recurrent architectures for time-series data and dropout layers for regularization are frequently used. Another variant, the MLP, expands on the ANN by incorporating multiple hidden layers, enabling deeper representation learning. An MLP typically includes an input layer for features, one or more hidden layers for transformation, and an output layer that produces classification or regression outputs. MLPs are particularly useful for solving prediction tasks related to academic performance and retention [26]. Like other neural models, MLPs assume IID inputs, and thus preprocessing steps such as batch normalization or the integration of institutional-level variables are recommended when working with hierarchical educational data.

### 2.4. Evaluation metrics

To assess the performance of the machine learning models used in this study, six evaluation metrics were employed: Accuracy, Precision, Recall, F1 Score, MSE, and log loss. Accuracy measures the proportion of correctly predicted instances, both positive and negative, relative to the total number of predictions, offering a general sense of model performance, especially in balanced datasets [42]. Precision quantifies the number of correctly predicted positive cases out of all instances predicted as positive, thereby reflecting the model's ability to minimize false positives. In contrast, Recall, or sensitivity, evaluates the model's capacity to identify all relevant positive instances, making it crucial in contexts where missing at-risk students has significant consequences. The F1 Score, which is the harmonic mean of Precision and Recall, provides a balanced assessment when there is a trade-off between the two, and is especially useful in handling imbalanced educational data [43]. Moreover, MSE is applied to measure the average squared difference between actual and predicted values in regression-based outputs, with lower values indicating more accurate models. Lastly, the log loss, used to evaluate the performance of probabilistic classification where lower log loss indicates better performance. These metrics together enable a comprehensive evaluation of classification and regression performance in predicting student retention outcomes.

### 2.5. Modeling setup

All machine learning models were implemented using Python 3.10 in a Jupyter Notebook environment. The following libraries were used: pandas for data preprocessing, scikit-learn for model training and evaluation, xgboost for gradient boosting implementation, and matplotlib/seaborn for visualization. To ensure robustness in model evaluation, a 5-fold cross-validation approach was adopted using KFold from the scikit-learn library. This method randomly splits the dataset into five equally sized folds. For each iteration, four folds were used to train the model while the remaining fold was used for validation. This process was repeated five times so that each data point served as a validation sample once. The use of cross-validation reduces the likelihood of overfitting and ensures the generalizability of results. Hyperparameter tuning was done manually for transparency, and all models were evaluated using the six standard metrics described in Section 2.4. No mathematical derivations were included, as this study focuses on the applied implementation of machine learning for institutional decision support in student retention.

### 2.6. Dataset description

The dataset utilized in this study comprises 482 anonymized student records drawn from the Information Technology Education (ITE) program at Mindanao State University – Main Campus across curriculum (see Table 4). It contains 146 variables, meticulously curated to represent a comprehensive view of each student's academic journey, demographic background, and socio-economic context. These attributes serve as predictive inputs for modeling student retention outcomes, with the DROPOUT variable encoded as a binary target (0 = retained, 1 = withdrawn).

The dataset is entirely complete, containing no missing values, which enhances its integrity for robust machine learning applications. The feature set is composed of two primary categories: (1) sociodemographic attributes, such as AGE, SEX, CITY, PROVINCE, CIVILSTATUS, MOTHERINCOME, and FATHERINCOME, which provide insight into the students' geographic and economic contexts; (2) academic attributes, encompassing raw or normalized performance indicators derived from institutional grading systems—examples include PED001 to PED012, STT071, and STT101, which likely correspond to core and technical courses specified in the university's prospectus; and (3) program-specific variables, such as PROGRAM, YEARLEVEL, and COURSE HISTORY, capturing progression within the prescribed BSIT-D track.

These academic features reflect the learning milestones embedded in the ITE curriculum, such as basic computer programming subjects, mathematics, probability and statistics for computing, data structures. and advanced database systems. Many of the performance variables are likely derived from institutional logs and standardized academic metrics, normalized within a [0,1] scale, allowing for effective comparison across subjects and semesters. The inclusion of longitudinal academic indicators and technical subject coverage enables the model to capture performance trends critical to predicting attrition.

Furthermore, the dataset's balanced class distribution (approximately 50 % dropout, 50 % retained) supports fair model training and evaluation without the need for resampling or synthetic balancing. The presence of highly granular features, such as course-level assessments and regional identifiers, adds dimensional richness to the analysis, enabling the exploration of intersectional retention risks (e.g., by geographic origin or year level).

This dataset exemplifies the growing intersection between academic analytics and educational data mining in the Philippine higher education landscape. It offers a rare opportunity to apply machine learning models to real-world academic data with direct implications for institutional policy, early warning systems, and student support interventions. By aligning variable design with curriculum structure and institutional reporting systems, the dataset ensures both contextual relevance and technical rigor for predictive modeling of student retention and also was reportedly used by this study [44]. The dataset is openly available for public access through CodeOcean [45].

**Table 4**
Summary of dataset characteristics for student retention modeling.

| Characteristic | Description |
|---|---|
| **Sample size (records)** | 482 student records |
| **Number of features** | 146 total variables |
| **Dropout rate** | 50.2 % (242 students) |
| **Retention rate** | 49.8 % (240 students) |
| **Missing values** | None (dataset is 100 % complete) |
| **Feature categories** | |
| • Sociodemographic | Age, Sex, Province, City, Civil Status, Parent's Income |
| • Academic Performance | Course grades (e.g., CCC, ITE, STT, MAT subjects), normalized on a [0,1] scale |
| • Program Progression | Year Level, Program Code, Total Units Earned, Course History |
| • Target Variable | DROPOUT (binary: 0 = retained, 1 = withdrawn) |

## 3. Results

This section presents the results of machine learning model assessments conducted on historical data collected over a decade (2012–2022) from the Information and Communication Technology Center (ICTC) of Mindanao State University-Main Campus. The data underwent preprocessing in Power BI using the CRISP-DM methodology before being imported into Jupyter Notebook for further analysis. Subsequently, the dataset was partitioned into $X$ (independent variables) and $Y$ (dependent variables) and split into training (70 %) and testing (30 %) sets. During analysis, variables highly correlated with the dependent variable were identified through Spearman coefficient (ρ or rho) in correlation analysis (Table 5) and subsequently removed to avoid multicollinearity,

**Table 5**
Rule of thumb or interpretation of Spearman's correlation value [46].

| Size of correlation | Interpretation |
|---|---|
| ±.90 to ±1.0 | Very high positive/negative correlation |
| ±.70 to ±.90 | High positive/negative correlation |
| ±.50 to ±.70 | Moderate positive/negative correlation |
| ±.30 to ±.50 | Low positive/negative correlation |
| .00 to ±.30 | Negligible correlation |

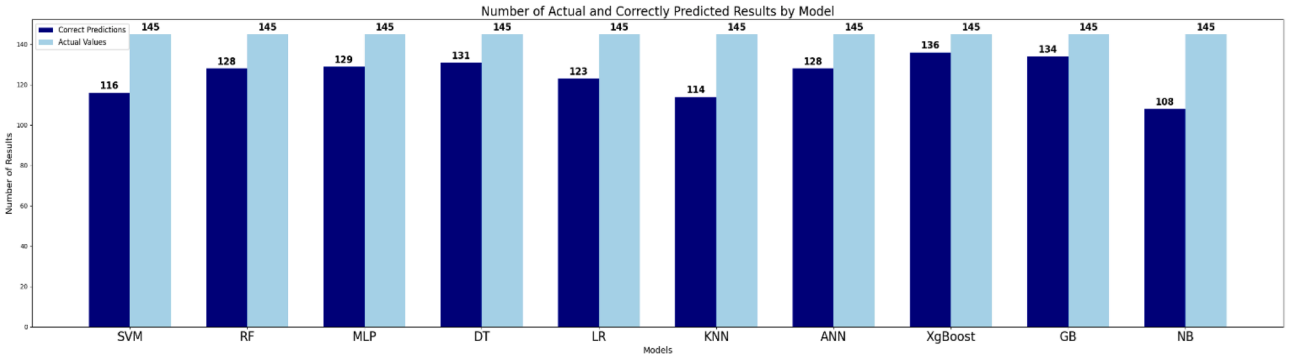



**Fig. 5.** Number of actual and correct prediciton comparison.

resulting in the retainment of 79 columns out of the initial 134. See Appendix A for the code of removing multicollinearity.

Fig. 5 illustrates the number of actual and correctly predicted results by the ten (10) machine learning algorithms. Results indicate that among the evaluated models, the XgBoost algorithm achieves the highest accuracy, correctly predicting 136 instances when compared to the actual *y*-values in the test dataset. This performance is then followed by the GB and Decision Tree (DT) models, which correctly predicted 134 and 131 instances, respectively, out of a total of 145. Further, the sample code used to assess the prediction accuracy of XgBoost is listed in Appendix B. A confusion matrix was also created to quantify correct predictions for the *y*-values (Dropout or Non-dropout).

On the other hand, Fig. 6 presents the confusion matrices for each classification model, while Table 6 provides a comparative summary of their predictive performance in identifying student dropout. Among the evaluated models, XGBoost demonstrated the most robust classification capability, achieving 67 true positives (correctly identified dropouts), 5 false positives (non-dropouts misclassified as dropouts), 4 false negatives (dropouts misclassified as non-dropouts), and 69 true negatives (correctly identified non-dropouts). These results indicate that XGBoost achieved the highest overall classification accuracy across both dropout and retention classes. As a reference, in the standard confusion matrix layout, the top-left cell represents true positives, top-right indicates false positives, bottom-left corresponds to false negatives, and bottom-right denotes true negatives.

Close behind XGBoost were the GB and ANN models, both of which exhibited strong and balanced predictive performance. The GB model yielded 66 true positives, 6 false positives, 5 false negatives, and 68 true negatives, indicating a high degree of precision and recall. The ANN model achieved 60 true positives, 12 false positives, 5 false negatives, and 68 true negatives. Although both models performed well, they incurred slightly higher misclassification rates compared to XGBoost. Nevertheless, GB and ANN demonstrated commendable levels of sensitivity (recall of actual dropouts) and specificity (correct identification of non-dropouts), supporting their utility in retention-related prediction tasks.

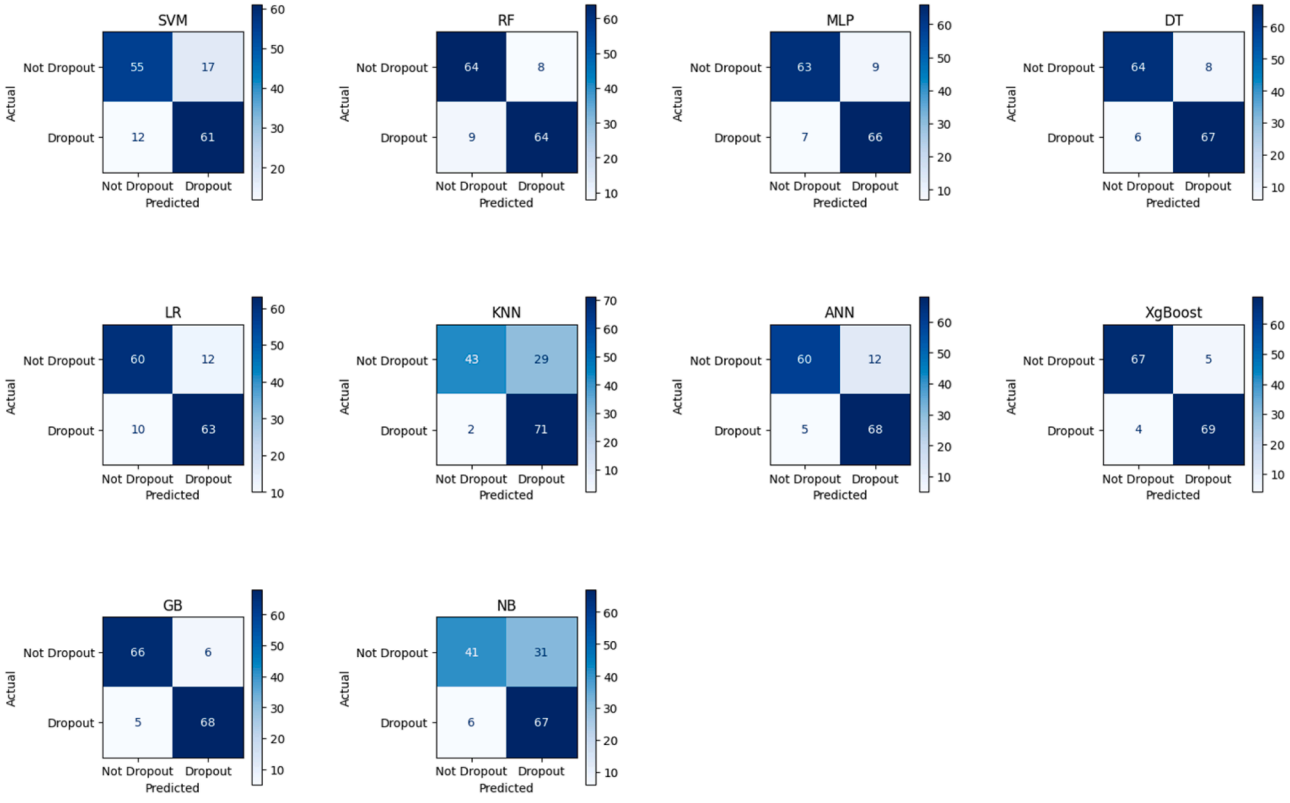



**Fig. 6.** Confusion matrix summary for each machine learning model.

**Table 6**
Confusion matrix components and dropout detection rates of machine learning models.

| Model | True Positive (Correct Dropout) | False Negative (Missed Dropout) | False Positive (False Dropout) | True Negatives (Correct Not Dropout) | Dropout Detection Rate (TPR / Recall) |
|---|---|---|---|---|---|
| SVM | 61 | 12 | 17 | 55 | 61 / (61+12) = **0.836** |
| RF | 64 | 9 | 8 | 64 | 64 / (64+9) = **0.877** |
| MLP | 66 | 7 | 9 | 63 | 66 / (66+7) = **0.904** |
| DT | 67 | 6 | 8 | 64 | 67 / (67+6) = **0.918** |
| LR | 63 | 10 | 12 | 60 | 63 / (63+10) = **0.863** |
| KNN | 71 | 2 | 29 | 43 | 71 / (71+2) = **0.972** |
| ANN | 68 | 5 | 12 | 60 | 68 / (68+5) = **0.932** |
| **XGBoost** | **69** | **4** | **5** | **67** | **69 / (69+4) = 0.945** |
| GB | 68 | 5 | 6 | 66 | 68 / (68+5) = **0.932** |
| NB | 67 | 6 | 31 | 41 | 67 / (67+6) = **0.918** |

Table 7 presents the initial evaluation results of ten machine learning models based on a standard train-test split. The assessment covers key classification metrics, such as Accuracy, Precision, Recall, and F1 Score, as well as error metrics like MSE and Log Loss. Among all models, Extreme Gradient Boosting (XGBoost) demonstrated the strongest performance, achieving the highest accuracy (93.79 %), F1 Score (93.88), and lowest MSE (6.21), with a competitive Log Loss value of 3.98. Similarly, Gradient Boosting yielded high classification scores, with 92.41 % accuracy and 92.52 F1 Score, and matched the lowest Log Loss value (3.23), suggesting well-calibrated probability estimates. Multilayer Perceptron (MLP) and Artificial Neural Network (ANN) models also performed strongly, particularly in recall (90.41 % and 93.15 %, respectively), indicating their sensitivity in identifying students at risk of withdrawal.

The Decision Tree classifier recorded competitive scores, including a 90.34 % accuracy and 90.54 F1 Score. However, its Log Loss was relatively high at 5.47, suggesting less confidence in its probability estimates despite good classification performance. Conversely, Naïve Bayes, while achieving a high recall of 91.78 %, posted the lowest precision (68.37 %) and the poorest error scores, with the highest MSE (25.52) and Log Loss (9.45), reflecting a high false positive rate and poor calibration. Support Vector Machine (SVM) similarly underperformed, with an accuracy of 80.00 %, MSE of 20.00, and a Log Loss of 5.72.

The addition of MSE and Log Loss in this preliminary evaluation provides a more nuanced view of each model's performance. While traditional classification metrics suggest strong outcomes, the error metrics, particularly Log Loss, highlight how well each model estimates class probabilities. A lower Log Loss indicates that the model is not only predicting correctly but is also confident and accurate in its probability estimates [47]. In this context, XGBoost and Gradient Boosting are shown to be not just accurate, but also well-calibrated.

However, these results are derived from a single train-test split, which may be subject to sampling bias and overfitting, especially in datasets with class imbalance or limited size [48]. The observed performance could therefore be overly optimistic. As such, further validation using K-Fold cross-validation was performed and is discussed in the succeeding section to assess the stability and generalizability of these models.

To address this, 5-fold cross-validation was applied to each model, offering a more reliable measure of generalization. When validated using this method, as shown in Fig. 7, several models exhibited performance declines. Notably, the Decision Tree classifier, which initially recorded over 90 % accuracy, saw a reduction to 84.44 %, with a corresponding drop in F1 Score and an increase in error metrics. This suggests that its original performance was likely influenced by overfitting. Similarly, XGBoost, while still the top performer, experienced a slight reduction in accuracy to 90.66 % and a modest increase in error, reflecting a more realistic estimate of its predictive strength. Naïve Bayes continued to display high recall but was again marked by poor precision and increased error, further confirming its tendency to overpredict positive cases. In contrast, models such as Random Forest, Logistic Regression, and Multilayer Perceptron maintained relatively consistent performance across both evaluation strategies, suggesting more stable generalization.

These results underscore the necessity of employing cross-validation techniques when evaluating machine learning models, especially in sensitive domains like education. While initial single-split evaluations are useful for exploratory comparisons, they do not account for variability and are prone to optimistic bias. Cross-validation mitigates these risks by ensuring each data point contributes to both training and testing, resulting in more robust and generalizable performance estimates.

Furthermore cross-validation revealed meaningful differences in model performance, correcting overestimations observed in pre-validation results. The XGBoost model remained the most reliable and consistent across all metrics, reinforcing its suitability for deployment in early warning systems aimed at reducing student attrition. These findings highlight the importance of rigorous validation when applying predictive analytics to educational decision-making.

## 4. Discussion

Student retention continues to be a persistent challenge in higher education, with significant implications for institutional rankings, reputational standing, and financial sustainability. As such, it remains a widely studied area in academic literature. Traditional approaches

**Table 7**
Assessment scores of each model.

| No. | MODEL | Accuracy ( %) | Precision ( %) | Recall ( %) | F1 score ( %) | MSE | Log Loss |
|---|---|---|---|---|---|---|---|
| 1 | Support Vector Machine | 80.00 | 78.21 | 83.56 | 80.79 | 20.0 | 5.72 |
| 2 | Random Forest | 88.28 | 88.89 | 87.67 | 88.28 | 11.7 | 3.98 |
| 3 | Multilayer Perceptron | 88.97 | 88.00 | 90.41 | 89.19 | 11.0 | 3.23 |
| 4 | Decision Tree | 90.34 | 89.33 | 91.78 | 90.54 | 9.66 | 5.47 |
| 5 | Logistic Regression | 84.83 | 84.00 | 86.30 | 85.14 | 15.1 | 8.20 |
| 6 | K-Nearest Neighbor | 83.45 | 78.82 | 91.78 | 84.81 | 16.55 | 7.21 |
| 7 | Artificial Neural Network | 88.28 | 85.00 | 93.15 | 88.89 | 10.61 | 4.23 |
| **8** | **Extreme Gradient Boosting** | **93.79** | **93.24** | **94.52** | **93.88** | **6.21** | **3.98** |
| 9 | Gradient Boosting | 92.41 | 91.89 | 93.15 | 92.52 | 7.59 | 3.23 |
| 10 | Naive Bayes | 74.48 | 68.37 | 91.78 | 78.36 | 25.52 | 9.45 |

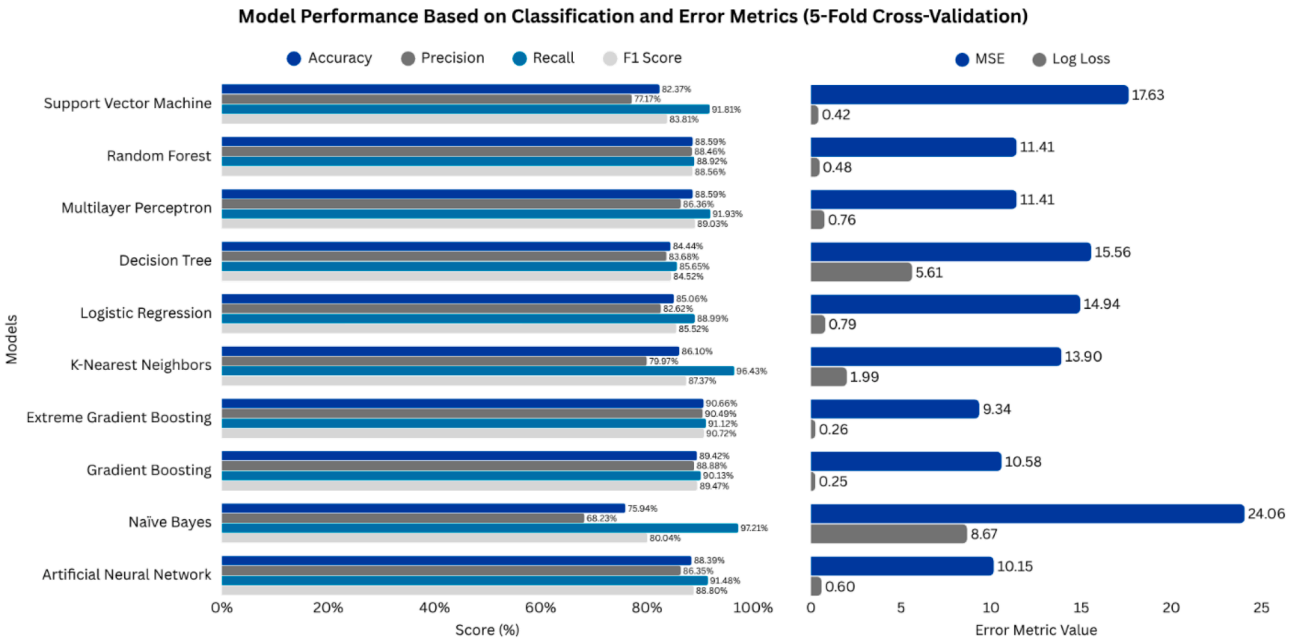



**Fig. 7.** Model performance based on classification and error metrics (5-fold cross-validation). This figure compares ten machine learning models using classification metrics—accuracy, precision, recall, and F1 score (left panel), alongside error metrics, MSE and log loss (right panel), evaluated using 5-fold cross-validation. Among the models, extreme gradient boosting (XGBoost) achieved the highest accuracy (90.66 %), F1 score (90.72), and one of the lowest error values (MSE = 9.34, log loss = 0.26), indicating superior predictive performance. Gradient boosting and multilayer perceptron (MLP) also exhibited competitive results with strong classification scores and low error values. Conversely, Naïve Bayes and support vector machine (SVM) showed higher error metrics, reflecting less consistent predictive reliability. These metrics provide a balanced view of both classification correctness and model confidence, aiding in robust model selection for student retention prediction tasks.

relying on subjective assessments and historical trends are often insufficient to capture the complexity of student attrition. To improve predictive capability and support timely interventions, ML techniques have become increasingly prominent. These methods enable institutions to extract patterns from diverse data sources, generate accurate dropout predictions, and inform data-driven policies aimed at improving student outcomes.

In this study, multiple ML models were evaluated to predict student dropout using a combination of academic and sociodemographic factors. The models were first assessed using a traditional train-test split, followed by a more rigorous 5-fold cross-validation, to ensure robustness and generalizability of results. Across both evaluation phases, Extreme Gradient Boosting (XGBoost) consistently outperformed other models. In the initial evaluation, XGBoost achieved an accuracy of 93.79 %, F1 Score of 93.88, and the lowest MSE of 6.21. After cross-validation, XGBoost maintained its top-ranking performance, achieving an average accuracy of 90.66 %, F1 Score of 90.72, and low error metrics (MSE = 9.34, Log Loss = 0.26). The confusion matrix also reinforced these findings, with XGBoost yielding 67 true positives, 69 true negatives, 5 false positives, and only 4 false negatives—the lowest total misclassifications among all models tested. Moreover, these results are strongly supported by prior literature. Study [49] found that XGBoost consistently outperformed alternative models in forecasting tasks such as stock price movements. Similarly [50], emphasized XGBoost's capacity for generalization, reduced overfitting, and enhanced predictive accuracy [51]. noted that XGBoost is uniquely capable of handling heterogeneous and noisy datasets, outperforming models that perform best only under controlled or homogeneous input conditions. Furthermore [52], reported that XGBoost achieves superior accuracy, specificity, and sensitivity compared to Logistic Regression, particularly in handling class-imbalanced datasets, an important feature in dropout prediction, where class imbalance is often present. Together, these findings and our results strongly establish XGBoost as a robust and adaptable model for educational prediction tasks.

Conversely, Naïve Bayes was consistently the weakest performer across all evaluation phases. It yielded the lowest accuracy (74.48 %) in the initial evaluation and the highest error metrics after cross-validation (MSE = 24.06, Log Loss = 8.67). Although Naïve Bayes recorded the highest recall (97.21 %), it also produced 31 false positives, indicating a tendency to over-predict dropout risk. While its simplicity and computational efficiency may still make it suitable for certain contexts or ensemble integration, the high misclassification rates suggest it may not be ideal for high-stakes educational interventions without further refinement.

The analysis also highlighted important trade-offs between parametric and nonparametric models. For example, K-Nearest Neighbors (KNN), a nonparametric algorithm, achieved the highest recall (97.2 %) but also had one of the highest false positive counts (29). Likewise, Decision Tree, another nonparametric method, showed strong recall (91.8 %) but relatively poor error metrics, reflecting sensitivity to overfitting and variability across folds. These models are highly effective in pattern detection but may require careful tuning and ensemble integration to enhance stability.

On the other hand, ensemble and hybrid models such as GB and ANN offered competitive, balanced performance. GB achieved 89.42 % accuracy post-cross-validation with a low Log Loss of 0.25, the lowest among all models, and maintained strong classification results (66 true positives, 68 true negatives). ANN followed closely with 88.39 % accuracy, 86.35 % precision, and a high recall of 91.48 %. These models demonstrate high sensitivity and specificity while maintaining fewer misclassifications, supporting their reliability for real-world applications. Also, the potential of neural network models is further supported by [53], which found that deep learning architectures, including MLP and ANN, outperform traditional statistical methods in modeling complex and nonlinear patterns in student data. However, our findings suggest that while ANN performed competitively, XGBoost retained a performance edge in terms of both classification accuracy and error minimization, suggesting greater generalizability and practical usability, especially when false positives must be minimized in institutional settings.

The integration of the predictive modeling framework developed in this study demonstrates substantial pedagogical and institutional value for ITE curricula in the Philippines. Although the analysis was conducted using data from a single institution, Mindanao State University Main Campus in Marawi City, the design and structure of the dataset establish a scalable and transferable foundation that can be readily adopted by other higher education institutions. The dataset comprises 482 complete student records and 146 variables, encompassing essential dimensions of student profiling such as sociodemographic indicators (e.g., age, sex, parental income), academic performance (e.g., normalized grades), and program progression metrics (e.g., year level, total units earned, course history). These data dimensions are consistent with widely accepted frameworks in educational data mining and learning analytics [54,55], making the model broadly applicable across academic settings.

Importantly, this framework offers a replicable blueprint for institutions aiming to develop their own student retention analytics. Prior studies emphasize that predictive models based on institutional data can effectively identify at-risk students and guide timely interventions [9–11,13–16]. Since the variables used in this study are generalizable and align with common student information systems, other universities can easily adapt their internal data infrastructures to adopt similar models. This creates a pathway for ITE programs to lead institutional analytics initiatives, fostering real-world engagement with artificial intelligence while strengthening the culture of data-informed decision--making in education.

Moreover, this study underscores the strategic role of ITE programs in driving institutional digital transformation. As digital transformation in education increasingly relies on data-driven approaches [53], equipping students and faculty with the capability to build and interpret predictive models using academic data is both timely and necessary. Engaging with real institutional data not only cultivates technical skills but also develops ethical reasoning and accountability in the use of AI in education—a point highlighted in literature on educational data science [14,18,37]. By empowering students to apply these models within their own institutions, ITE departments position themselves as active agents of change, contributing to improved educational outcomes through interpretable, context-aware, and impactful computing solutions.

Looking forward, future research on student retention should prioritize the integration of psychosocial, behavioral, and engagement-based features [56] such as attendance, motivation, learning management system (LMS) activity [57], and peer interaction data [58], to enhance predictive power. Moreover, temporal modeling using architectures like Long Short-Term Memory (LSTM) networks offers promising potential for real-time risk monitoring and longitudinal trend analysis across semesters [25]. To improve both model performance and interpretability, researchers are also encouraged to explore Deep P-Spline Models [59], Additive Gaussian Process Models [60], and hybrid ensemble strategies, which offer scalable and flexible frameworks for capturing complex non-linear patterns in large, structured educational datasets.

Furthermore, the results of this study affirm that XGBoost is the most accurate, stable, and generalizable model for predicting student dropout, outperforming both traditional and deep learning methods. However, the continued development of richer datasets, more interpretable models, and ethically aligned deployment practices will be vital in advancing the field of educational data mining and its impact on student success, particularly when integrated into IT education for both instructional and institutional innovation.

## 5. Conclusions

This study evaluated ten machine learning algorithms for predicting student dropout using historical data from MSU-Marawi spanning 2012 to 2022. Through the application of the CRISP-DM methodology, combined with Power BI for preprocessing and Jupyter Notebook for modeling, the data was rigorously prepared for analysis. After filtering variables using Spearman correlation to minimize multicollinearity, models were assessed using both a traditional train-test split and 5-fold cross-validation. Among the tested models, XGBoost consistently delivered the highest performance in both classification and error metrics, with an average cross-validated accuracy of 90.66 %, F1 Score of 90.72, and low error values (MSE = 9.34, Log Loss = 0.26). Ensemble models such as GB and ANN also showed strong and balanced results. These findings demonstrate that predictive modeling, particularly using ensemble and neural network techniques, offers significant promise for building early warning systems that can identify students at risk of attrition. Moreover, the dataset and modeling approach are structured in a generalizable format, offering a transferable framework for other higher education institutions aiming to improve student retention through analytics. From a pedagogical perspective, the integration of predictive modeling into ITE programs can cultivate technical competence in educational data science while reinforcing real-world application of AI in institutional contexts.

Despite its promising outcomes, this study is not without limitations. First, the data was sourced from a single institution, which may limit the external validity of the findings despite the dataset's broad and generalized features. Further, while 5-fold cross-validation mitigated risks of overfitting and sampling bias, more advanced validation strategies such as stratified or nested cross-validation could yield even more robust generalization. In addition, the study focused primarily on academic and sociodemographic variables; behavioral and psychosocial dimensions, such as LMS engagement, motivation, or peer interaction, were not included but are known to be significant predictors of student success. The inclusion of these features in future studies may improve the precision and contextual relevance of predictive models. Similarly, while XGBoost achieved the strongest performance, model interpretability remains a challenge for deployment in sensitive domains like education. Institutions seeking to implement such models should complement high-performance algorithms with explainable AI tools to ensure transparency, fairness, and ethical accountability. Lastly, this study reinforces the emerging role of ITE programs in driving institutional digital transformation by equipping both faculty and students with the skills to develop intelligent systems grounded in real-world data. Looking ahead, future research can build on this work by investigating hybrid ensemble techniques, Deep P-Spline architectures, Additive Gaussian Process Models, and longitudinal approaches such as LSTM networks to capture more nuanced patterns in student dropout behavior. As higher education institutions increasingly embrace data-driven decision-making, this study serves as both a foundational benchmark and a transferable framework for embedding predictive analytics into educational practice and strategic policy formulation.

## Acknowledgment

The authros would like to express their sincere gratitude to Mindanao State University - Main Campus, and the Office of the Vice Chancelor for Academic Affairs for granting permissoin to utilize students' academic and sociodemographic data for this research. Special thanks are also extended to the Mamitua Saber Institute of Research and Creation (MSIRC) for their valuable support throughout the conduct of the study.

## CRediT authorship contribution statement

**Reymark D. Deleña:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Norniña J. Dia:** Writing – original draft, Visualization, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Redeemtor R. Sacayan:** Writing – original draft, Validation, Funding acquisition, Formal analysis. **Joseph C. Sieras:** Supervision, Funding acquisition, Conceptualization. **Suhaina**

**A. Khalid:** Resources, Investigation, Funding acquisition. **Amer Hussien T. Macatotong:** Software, Resources, Funding acquisition. **Sacaria B. Gulam:** Resources, Funding acquisition, Data curation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendices

*Splitting the data into training and testing data then removing multicollinearity*

```
#TRANSFORMING THE CATEGORICAL INTO NUMERICAL DATA
from sklearn.preprocessing import LabelEncoder
#creating an instance of Label Encoder le = LabelEncoder()
#Using .fit_transform function to fit label encoder and return encoded label label = le.fit_transform(student_df['DROPOUT'])
#removing the categorized (old) column 'TARGET' and adding 'DROPOUT' which contains numerical data. student_df.drop ('DROPOUT', axis=1, inplace=True) student_df['DROPOUT'] = label
#SPLITTING THE DATA INTO TESTING AND TRAINING from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3)
#GENERATING THE CORRELATION MATRIX corrmat = student_df.corr(method = "spearman") corrmat plt.figure(figsize=(50,50)) mask = np.triu(np.ones_like(corrmat, dtype=bool)) sns.heatmap(corrmat, annot=True, annot_kws={"size":4}, mask=mask, vmin=-1, vmax=1) plt.title('Correlation Coefficient of Predictors') plt.show
#REMOVING THE HIGHLY CORRELATED VARIABLES (WITH ≥70 % THRESHOLD) def correlation(student_df, threshold): correlated_cols = set() corr_matrix = student_df.corr(method="spearman") for i in range (len(corr_matrix.columns)): for j in range (i): if abs(corr_matrix.iloc[i, j])> threshold: colname = corr_matrix.columns[i] correlated_cols.add (colname) return correlated_cols
#storing the variables with ≥70 % threshold in corr_features then removing them. corr_feature = correlation(student_df, 0.7)
X_train.drop(labels=corr_feature, axis=1, inplace=True)
X_test.drop(labels=corr_feature, axis=1, inplace=True)
#plotting the new matrix corr = X_train.corr(method="spearman") plt.figure(figsize=(30,30)) mask = np.triu(np.ones_like(corr, dtype=bool)) sns.heatmap(corr, annot=True, annot_kws={"size":4}, mask=mask, vmin=-1, vmax=1) plt.title('Correlation Coefficient of Predictors') plt.show
```

*Sample code for identifying the prediction performance of a model*

```
#importing evaluation metrics from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, mean_squared_error from sklearn.metrics import confusion_matrix
#importing the model import xgboost as xgb xgb = xgb.XGBClassifier()
#fitting the model xgb.fit(X_train, y_train)
#TESTING THE PREDICTION prediction_xgb = xgb.predict(X_test) prediction_xgb
```

*Assessing the accuracy rate of the model*

```
#checking the evaluation metrics score of the model accuracy_xgb = accuracy_score(y_test, prediction_xgb)*100 precision_xgb = precision_score(y_test, prediction_xgb)*100 recall_xgb = recall_score(y_test, prediction_xgb)*100 f1_xgb = f1_score(y_test, prediction_xgb)*100 mse_xgb = mean_squared_error(y_test, prediction_xgb)*100 print("Accuracy Score: ", accuracy_xgb) print("Precision Score: ", precision_xgb) print("Recall Score: ", recall_xgb) print("F1 Score: ", f1_xgb) print("Mean Squared Error: ", mse_xgb)
# GENERATING THE CONFUSION MATRIX cm = confusion_matrix (y_test, prediction_xgb) sns.heatmap(cm, annot=True) plt.ylabel('Prediction',fontsize=13) plt.xlabel('Actual',fontsize=13) plt.title('Confusion Matrix',fontsize=17) plt.show()
#COMPUTING THE LOG LOSS OF THE MODEL from sklearn.metrics import log_loss ll_xgb = log_loss(y_test, prediction_xgb) print('Log Loss of Extreme Gradient Boosting:', ll_xgb)
```

## Data availability

If this paper is accepted for publication, the data will be published in a separate article.

## References


[1] S. Ameri, M.J. Fard, R.B. Chinnam, C.K. Reddy, Survival analysis based framework for early prediction of student dropouts, in: International Conference on Information and Knowledge Management, Proceedings, 2016, pp. 903–912, https://doi.org/10.1145/2983323.2983351, 24-28-October-2016.

[2] S. Trivedi, Improving students' retention using machine learning: impacts and implications, Sci. Prepr. (2022), https://doi.org/10.14293/S2199-1006.1.SOR-.PPZMB0B.V2.

[3] Bineid, A.A. (2022). Predicting student withdrawal from UAE CHEDS repository using data mining methodology. 1–72. https://bspace.buid.ac.ae/handle/1234/2130.

[4] Calvert, C.E. (2014). Developing a model and applications for probabilities of student success: a case study of predictive analytics. 29(2), 160–173. https://doi.org/10.1080/02680513.2014.931805.

[5] R.S.J.d. Baker, K. Yacef, The State of educational data mining in 2009: a review and future visions, J. Educ. Data Min. 1 (1) (2009) 3–17, https://doi.org/10.5281/ZENODO.3554657.

[6] Fain, P. (2016). Data on student engagement with an LMS is a key to predicting retention. https://www.insidehighered.com/news/2016/06/13/data-student-engagement-lms-key-predicting-retention.

[7] A.I. Adekitan, E. Noma-Osaghae, Data mining approach to predicting the performance of first year student in a university using the admission requirements, Educ. Inf. Technol. 24 (2) (2018) 1527–1543, https://doi.org/10.1007/S10639-018-9839-7, 2018 24:2.

[8] M. Bucos, B. Drăgulescu, Predicting student success using data generated in traditional educational environments, TEM J. 7 (3) (2018) 617, https://doi.org/10.18421/TEM73-19.

[9] A.I. Adekitan, E. Noma-Osaghae, Data mining approach to predicting the performance of first year student in a university using the admission requirements, Educ. Inf. Technol. 24 (2) (2018) 1527–1543, https://doi.org/10.1007/S10639-018-9839-7, 2018 24:2.

[10] H. Almarabeh, Analysis of students' performance by using different data mining classifiers, Int. J. Mod. Educ. Comput. Sci. 9 (8) (2017) 9–15, https://doi.org/10.5815/IJMECS.2017.08.02.

[11] E. Alhazmi, A. Sheneamer, Early predicting of students performance in higher education, IEEE Access 11 (2023) 27579–27589, https://doi.org/10.1109/ACCESS.2023.3250702.

[12] D. Uliyan, A.S. Aljaloud, A. Alkhalil, H.S.A. Amer, M.A.E.A. Mohamed, A.F. M. Alogali, Deep learning model to predict students retention using BLSTM and CRF, IEEE Access 9 (2021) 135550–135558, https://doi.org/10.1109/ACCESS.2021.3117117.

[13] N.R. Beckham, L.J. Akeh, G.N.P. Mitaart, J.v. Moniaga, Determining factors that affect student performance using various machine learning methods, Procedia Comput. Sci. 216 (2023) 597–603, https://doi.org/10.1016/j.procs.2022.12.174.

[14] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, P.C. Nshimyumukiza, Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization, Comput. Educ. 3 (2022) 100066, https://doi.org/10.1016/j.caeai.2022.100066.

[15] R. Ghorbani, R. Ghousi, Comparing different resampling methods in predicting students' performance using machine learning techniques, IEEE Access 8 (2020) 67899–67911, https://doi.org/10.1109/ACCESS.2020.2986809.

[16] F. Marbouti, J. Ulas, C.-H. Wang, Academic and demographic cluster analysis of engineering student success, IEEE Trans. Educ. 64 (3) (2021) 261–266, https://doi.org/10.1109/TE.2020.3036824.

[17] P. Xuan Lam, P.Q.H. Mai, Q.H. Nguyen, T. Pham, T.H.H. Nguyen, T.H Nguyen, Enhancing educational evaluation through predictive student assessment modeling, Comput. Educ. 6 (2024) 100244, https://doi.org/10.1016/J.CAEAI.2024.100244.

[18] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, R.M.G. García-Castelán, Predictive analytics study to determine undergraduate students at risk of dropout, Front. Educ. 8 (2023) 1244686, https://doi.org/10.3389/FEDUC.2023.1244686.

[19] S. Alwarthan, N. Aslam, I.U. Khan, An explainable model for identifying at-risk student at higher education, IEEE Access 10 (2022) 107649–107668, https://doi.org/10.1109/ACCESS.2022.3211070.

[20] G. Feng, M. Fan, Y. Chen, Analysis and prediction of students' academic performance based on educational data mining, IEEE Access 10 (2022) 19558–19571, https://doi.org/10.1109/ACCESS.2022.3151652.
[21] A.M. Mariano, A.B. de Magalhães Lelis Ferreira, M.R. Santos, M.L. Castilho, A.C.F. L.C Bastos, Decision trees for predicting dropout in engineering course students in Brazil, Procedia Comput. Sci. 214 (2022) 1113–1120, https://doi.org/10.1016/j.procs.2022.11.285.
[22] H.P. Singh, H.N. Alhulail, Predicting student-teachers dropout risk and early identification: a four-step logistic regression approach, IEEE Access 10 (2022) 6470–6482, https://doi.org/10.1109/ACCESS.2022.3141992.
[23] F. Marbouti, J. Ulas, C.-H. Wang, Academic and demographic cluster analysis of engineering student success, IEEE Trans. Educ. 64 (3) (2021) 261–266, https://doi.org/10.1109/TE.2020.3036824.
[24] A. Nabil, M. Seyam, A. Abou-Elfetouh, Prediction of students' academic performance based on courses' grades using deep neural networks, IEEE Access 9 (2021) 140731–140746, https://doi.org/10.1109/ACCESS.2021.3119596.
[25] H. Prabowo, A.A. Hidayat, T.W. Cenggoro, R. Rahutomo, K. Purwandari, B. Pardamean, Aggregating time series and tabular data in deep learning model for university students' GPA prediction, IEEE Access 9 (2021) 87370–87377, https://doi.org/10.1109/ACCESS.2021.3088152.
[26] C.F. Rodríguez-Hernández, M. Musso, E. Kyndt, E. Cascallar, Artificial neural networks in academic performance prediction: systematic implementation and predictor evaluation, Comput. Educ. 2 (2021) 100018, https://doi.org/10.1016/j.caeai.2021.100018.
[27] A.J. Fernández-García, R. Rodríguez-Echeverría, J.C. Preciado, J.M.C. Manzano, F. Sánchez-Figueroa, Creating a recommender system to support higher education students in the subject enrollment decision, IEEE Access 8 (2020) 189069–189088, https://doi.org/10.1109/ACCESS.2020.3031572.
[28] H.A. Mengash, Using data mining techniques to predict student performance to support decision making in university admission systems, IEEE Access 8 (2020) 55462–55470, https://doi.org/10.1109/ACCESS.2020.2981905.
[29] T.A. Cardona, E.A. Cudney, Predicting student retention using support vector machines, Procedia Manuf. 39 (2019) 1827–1833, https://doi.org/10.1016/J.PROMFG.2020.01.256.
[30] A. Viloria, J.G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N.O. Llinas, M. A. David, Integration of data technology for analyzing university dropout, Procedia Comput. Sci. 155 (2019) 569–574, https://doi.org/10.1016/j.procs.2019.08.079.
[31] G. Lesinski, S. Corns, Multi-objective evolutionary neural network to predict graduation success at the United States military academy, Procedia Comput. Sci. 140 (2018) 196–205, https://doi.org/10.1016/J.PROCS.2018.10.329.
[32] G. Lesinski, S. Corns, C. Dagli, Application of an artificial neural network to predict graduation success at the United States military academy, Procedia Comput. Sci. 95 (2016) 375–382, https://doi.org/10.1016/J.PROCS.2016.09.348.
[33] M. Goga, S. Kuyoro, N. Goga, A recommender for improving the student academic performance, Procedia - Soc. Behav. Sci. 180 (2015) 1481–1488, https://doi.org/10.1016/j.sbspro.2015.02.296.
[34] C. Schröer, F. Kruse, J.M. Gómez, A systematic literature review on applying CRISP-DM process model, Procedia Comput. Sci. 181 (2021) 526–534, https://doi.org/10.1016/J.PROCS.2021.01.199.
[35] N. Caetano, P. Cortez, R.M.S. Laureano, Using data mining for prediction of hospital length of stay: an application of the CRISP-DM methodology, Lect. Notes Bus. Inf. Process. 227 (2015) 149–166, https://doi.org/10.1007/978-3-319-22348-3_9.
[36] Wirth, R., & Hipp, J. (2000). Crisp-dm: towards a standard process model for data mining.
[37] V. Vijayalakshmi, K. Venkatachalapathy, Comparison of predicting student's performance using machine learning algorithms, Int. J. Intell. Syst. Appl. 11 (12) (2019) 34–45, https://doi.org/10.5815/IJISA.2019.12.04.
[38] S. Hansun, J. Tanuwijaya, LQ45 stock index prediction using k-nearest neighbors regression, Int. J. Recent Technol. Eng. 8 (3) (2019) 2277–3878, https://doi.org/10.35940/ijrte.C4663.098319 (IJRTE).
[39] E. Osmanbegovic, M. Suljic, Data mining approach for predicting student performance, J. Econ. Bus. 10 (1) (2012) 3–12. https://www.researchgate.net/publication/242341193_DATA_MINING_APPROACH_FOR_PREDICTING_STUDENT_PERFORMANCE.
[40] A.M. Durán-Rosal, T. Ashley, J. Pérez-Rodríguez, F. Fernández-Navarro, Global and diverse ensemble model for regression, Neurocomputing 647 (2025) 130520, https://doi.org/10.1016/J.NEUCOM.2025.130520.
[41] R.E. Uhrig, Introduction to artificial neural networks, in: Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics 1, 2002, pp. 33–37, https://doi.org/10.1109/IECON.1995.483329.
[42] Accuracy, I.S.O., Of Measurement Methods and Results—Part 1: General principles and Definitions, International Organization for Standardization, Geneva, Switzerland, 1994.
[43] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (1) (2020) 1–13, https://doi.org/10.1186/S12864-019-6413-7/TABLES/5.
[44] N.J. Dia, J.C. Sieras, S.A. Khalid, A.H.T. Macatotong, J.M. Mondejar, E.R. Genotiva, R.D. Delena, EduGuard RetainX: an advanced analytical dashboard for predicting and improving student retention in tertiary education, SoftwareX 29 (2025) 102057, https://doi.org/10.1016/j.softx.2025.102057.
[45] N. Dia, R. Delena, EduGuard RetainX: an Advanced Analytical Dashboard For Predicting and Improving Student Retention in Tertiary Education (Version V1), CodeOcean, 2024. https://codeocean.com/capsule/2453835/tree/v1. </Dataset>.
[46] M. Mukaka, Statistics corner: a guide to appropriate use of correlation coefficient in medical research, Malawi Med. J. 24 (3) (2012) 69–71.
[47] Setia, M. (2023, September 14). Log loss - logistic regression's cost function for beginners. https://www.analyticsvidhya.com/blog/2020/11/binary-cross-entropyaka-log-loss-the-cost-function-used-in-logistic-regression/.
[48] H. Babaei, M. Zamani, S. Mohammadi, The impact of data splitting methods on machine learning models: a case study for predicting concrete workability, Mach. Learn. Comput. Sci. Eng. 1 (1) (2025), https://doi.org/10.1007/s44379-025-00021-3.
[49] Dey, S., Kumar, Y., Saha, S., & Basak, S. (2016). Forecasting to classification: predicting the direction of stock market price using Xtreme gradient boosting. https://doi.org/10.13140/RG.2.2.15294.48968.
[50] P. Carmona, F. Climent, A. Momparler, Predicting failure in the U.S. banking sector: an extreme gradient boosting approach, Int. Rev. Econ. Finance 61 (2019) 304–323, https://doi.org/10.1016/J.IREF.2018.03.008.
[51] I. Babajide Mustapha, F. Saeed, Bioactive molecule prediction using extreme gradient boosting, Molecules 21 (8) (2016), https://doi.org/10.3390/MOLECULES21080983.
[52] Hanif, I. (2020). Implementing extreme gradient boosting (XGBoost) classifier to improve customer churn prediction. https://doi.org/10.4108/EAI.2-8-2019.2290338.
[53] M.F. Flayyih, H. Hassan TOUT, Predictive analytics model for students' grade prediction using machine learning. a review, EDRAAK 2023 (2023) 1–4, https://doi.org/10.70470/EDRAAK/2023/001.
[54] C. Romero, S. Ventura, Educational data mining and learning analytics: an updated survey, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10 (3) (2020), https://doi.org/10.1002/widm.1355.
[55] G. Siemens, P. Long, Penetrating the fog: analytics in learning and education, Educ. Rev. 46 (5) (2011) 30. https://eric.ed.gov/?id=EJ950794.
[56] B. Chahar, S.R. Jain, V. Hatwal, Mediating role of employee motivation for training, commitment, retention, and performance in higher education institutions, Probl. Perspect. Manag. 19 (3) (2021) 95–106, https://doi.org/10.21511/ppm.19(3).2021.09.
[57] N. Nithiyanandam, S. Dhanasekaran, A.S. Kumar, D. Gobinath, P. Vijayakarthik, G. V. Rajkumar, U. Muthuraman, Artificial intelligence assisted student learning and performance analysis using instructor evaluation model, in: 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 1555–1561, https://doi.org/10.1109/icesc54411.2022.9885462.
[58] S. Xiao, A. Shanthini, D. Thilak, Instructor performance prediction model using artificial intelligence for higher education systems, J. Interconnect. Netw. 22 (Supp03) (2021), https://doi.org/10.1142/s0219265921440035.
[59] N.Y. Hung, L. Lin, V.D. Calhoun, Deep P-Spline: theory, fast tuning, and application, arXiv (Cornell Univ.) (2025), https://doi.org/10.48550/arxiv.2501.01376.
[60] L. Lin, V.R. Joseph, Transformation and additivity in gaussian processes, Technometrics 62 (4) (2019) 525–535, https://doi.org/10.1080/00401706.2019.1665592.