Predicting Student Retention and Dropout Rates in Cronasia Foundation College Inc. Using Educational Data Mining and Machine Learning Regression Techniques

Leomil Jay Duran

Graduate School, University of the Immaculate Conception, Davao City, Philippines

DOI: https://dx.doi.org/10.47772/IJRISS.2025.91100456

Received: 01 December 2025; Accepted: 09 December 2025; Published: 18 December 2025

ABSTRACT

This study investigated the potential for machine learning (ML) and educational data mining (EDM) capabilities in predicting retention and dropout rates at Cronasia Foundation College Inc. (CFCI). Predicting student persistence is a founding element of improving and enabling retention efforts in higher education institutions. However, understanding what contributes to retention and dropout still presents complications. Hence, this study crafted predictive models via the analysis of historical academic records, levels of engagement, socio-economic level, and psychological components and examined a dataset of 9,100 student records (75% training and 25% testing). According to the performance of the models analyzed using a total of five machine learning classifiers (Decision Trees, Random Forest, Support Vector Machines, Neural Networks, and Logistic Regression), and the models have been analyzed using F1-score, recall, accuracy, and precision. The accuracy from the models we analyzed from the highest being the model that was Neural Network with 80.42%, which had precision of 0.840, recall of 0.895, and F1-score of 0.867 for retention (Class 0); and precision of 0.692, recall of 0.582, and an F1-score of 0.632 for dropout (Class 1). Random Forest and Decision Tree had similar accuracy with Random Forest's accuracy being 79.90% with an F1-score for dropout of 0.622, and Decision Tree's accuracy was 80.58%. Logistic regression performed with the lowest accuracy of 73.98%, and had poor recall associated with dropout; inducing action for a number of academic leaders to begin intervening and take responsibility in helping retain students prior to their exit point, or after their first semester. The study found that retention was most strongly related to important variables such as intrinsic academic performance, attendance, and scholarship status. The results of the study can aid in data-based decision-making in higher education by helping institutions develop focused programs to increase retention and decrease dropout.

Keywords: Student Retention, Student Dropout, Educational Data Mining, Machine Learning, Predictive Analytics

INTRODUCTION

The issues of student retention and dropout rates are increasingly important topics in higher education institutions throughout the world. Student retention through graduation is not only a criterion of institutional success (or lack thereof), but it has also been regarded as a measure of academic success for students [8][6]. There are many factors that can affect whether a student remains in or withdrawals from education including their academic results, financial security, mental health, and socio-economic class background [47][17]. While there are intervention programs in place to support at-risk students, many institutions struggle to successfully identify those who require help effectively and early enough before the situation is irreversibly damaged. This combination of corroborated and correlative relationships shows that institutional commitment and enhanced services are essential strategies to help students persist [6][10]. Consequently, there is an indeterminate gap regarding proactive identification or prediction of students who drop out as a result of life challenges that could be effectively addressed using data-driven approaches [44][11]. The advances and improvements by student support services notwithstanding, limited use of predictive analytics in higher education limits these data analytics methodologies usefully [11][1]. Addressing this gap in knowledge is important in supporting greater educational success, institutional effectiveness and efficiency, as the lost goals of many students (and their institutions) are caused by unconnected challenges they are unable to navigate [13][6].

This research will employ educational data mining (EDM) and machine learning (ML) techniques to study dropout and retention within Cronasia Foundation College, Inc. (CFCI). By studying past academic records or transcripts, demographics, and behavioral data from students, we will develop models to predict when students are at risk of dropping out. The models developed will predict student retention and dropout using different machine learning algorithms. The research and findings of this study will be beneficial in informing recommendations for the Institution to be able to develop an intervention for students predicted to dropout or not persist, thus improving student success. Moreover, this research will, ultimately, contribute to the use of predictive analytics and improve institutional decision making in a data-driven culture of retention management.

The use of machine learning in educational analytics has the potential to be disruptive by predicting and addressing student attrition with large-scale data and algorithms uncovering subtle behaviors consistent with disengagement and risk of failure [15]. In particular, predictive models facilitated by educational data mining can help education institutions to identify at-risk dropouts earlier, which helps deliver timeliness and tailor intervention to improve retention rates [18]. The research study on predicting retention and dropout rates is an example of this way of modeling, and expired education data analytics to explore existing literature on predicting dropout; using academic take rates, engagement rates and demographics, the research addresses historical challenges of dealing with imbalance, and missing data in educational data sets as well as improving the predictive accuracy using robust machine learning models [14]. This comprehensive use of machine learning serves to highlight the valuable opportunity of developing new education systems to adaptively improve evidence-based approaches to student learning.

Research Questions

This study aims to predict student retention and dropout rates using educational data mining and machine learning regression techniques. Specifically, it seeks to answer the following questions.

1. What are the primary factors that predict student retention and dropout rates at CFCI in terms of:

Academic Performance

Socioeconomic Status

Student Engagement

Psychological and Behavioral Factors

2. What is the effectiveness of the various machine learning regression models in predicting student retention and dropout rates at CFCI?

Decision Trees

Random Forest

Support Vector Machines (SVM)

Neural Networks

3. Which machine learning regression model provides the most accurate prediction of student retention and dropout rates based on the factors?

4. How can the predicted retention and dropout rates from machine learning regression models be used to design targeted interventions that improve student retention and reduce dropout rates at CFCI?

Significance of the Study

This study will be beneficial to the following:

Cronasia Foundation College Inc. will benefit by acquiring data-driven insights to enhance its student retention programs. The study will assist administrators in developing by leveraging machine learning models, the institution can optimize student support services, improve academic performance, and increase overall institutional effectiveness.

The students. Students will benefit from the study's findings since they shed light on the variables that affect their perseverance and academic achievement. Students can make educated decisions about their education and seek assistance when necessary if they are aware of the main issues influencing retention. Predictive models can also facilitate tailored academic interventions, which will ultimately enhance students' performance and general wellbeing.

The Teachers. Teachers will obtain a better grasp of risk factors and student retention trends as a result of this study. Teachers can use the findings to identify and assist pupils who are at danger by putting data-driven methods into practice. With this information, educators can implement focused teaching strategies, provide more academic help, and foster a more encouraging learning atmosphere.

The Parents. Parents will use intervention strategies and proactive initiatives to reduce dropout rates. provides useful details regarding the factors influencing their child's scholastic trajectory. If they understand the elements that predict retention and dropout, they can better guide and assist their children. This study encourages more communication between parents and the school to foster an environment where students can thrive academically and personally.

The Researchers. This study will contribute to the growing body of research on educational data mining and machine learning in higher education. Researchers can use the findings as a foundation for future studies that explore new predictive models and intervention strategies. The study will also provide valuable empirical data that can be used to improve retention prediction methodologies.

The Future Researchers. Future research in the areas of student retention and educational analytics will use this study as a guide. The results can be expanded upon, machine learning methods can be improved, and other variables affecting student persistence can be investigated by future researchers.

Scope and Delimitation

This study addresses student retention and dropout at Cronasia Foundation College Inc. (CFCI) through either educational data mining or machine learning regression analysis techniques. The scope of this research involves examining students' academic records, demographic characteristics, as well as behavioral information as an attempt to explain relationships between these variables, from retention/dropout trends. The study will identify different machine learning regression type models (e.g. Decision Trees, Random Forest, Support Vector Machines, Neural Networks), to analyze their results, as it relates to students' retention and dropout.

LITERATURE REVIEW

Due to their inability to retain students until the completion of their degree, many higher education organizations explicitly define problems relating to the issues of student drop out and student retention. Student success is affected by a number of factors, including social engagement, socio-economic status, mental health, and academic performance. Institutions are now researching more possibilities to predictive dropout probabilities, due to improvements in educational data mining (EDM) and machine learning (ML) capabilities. This literature review incorporates studies relating to EDM, retention prediction, and the applicability of ML methods for dropout prediction.

Factors Influencing Student Retention and Dropout Rates

Student retention and drop-out rates are influenced by a host of variables including, but not limited to, academic achievement, finance, mental health, and extracurricular participation. Academic achievement, in particular Grade Point Average (GPA), has been demonstrated to be strongly correlated with student retention according to a substantial body of research. The role of academic achievement in the retention equation is further evident

due to the tendency of students with lower GPAs to have higher drop-out rates. For example, research shows students who successfully complete first-year seminar courses have higher retention rates and improved GPAs; some studies report retention rates as high as 98% for students in seminar courses [32]. In addition, a study found amotivation is a significant predictor of retention in first-year college and college GPA, which indicates colleges and universities should focus on improving retention through motivation [25].

Retention is also strongly relevant to socioeconomic status. Financial insecurity often presents difficulties for lower-income students, magnifying stress levels and negatively affecting performance and commitment to completion of their same-level education. When looking at ways to improve retention of economically impoverished students, research has indicated many support strategies that can be adopted, including creating supportive educational contexts and providing resources [23], and collectively the results identify the need to counteract the structural barriers that low-SES students face to reduce dropping out.

It is becoming increasingly apparent that the impact mental health issues have on student retention cannot be underestimated. Anxiety, depression, and stress can significantly affect a student's ability to engage with his/her education, which can contribute to disengagement and ultimately increased student dropout rates [12]. Given this evidence, educational institutions should provide mental health support services that can alleviate these barriers and improve student well-being.

Engagement in extracurricular activities represents a key factor affecting student retention. It has been demonstrated in many studies that students who engage in extracurricular activities gain benefits that extend far beyond their academic experience. Improvements in social skills, development, and sense of community within their academic communities have been shown to occur, indicating that engagement in extracurricular activities has many potential advantages. For example, studies indicate that engagement in extracurricular activities builds social responsibility, confidence and self-worth, and can significantly reduce dropout rates [19][20]. Engaging in extracurricular activities has been shown to foster both skill development and peer support and can provide a positive impact on academic and non-academic outcomes, including in studies that separate these effects [24][46].

Student retention and dropout rates exist in a unique intersecting ecosystem of academic success, financial circumstances, mental health, and extracurricular participation. Offering focused academic support, mental health, and active engagement in extracurricular activities that connect students to their social and academic surroundings can support an improved retention strategy.

Educational Data Mining and Machine Learning in Retention Prediction

Educational Data Mining (EDM) has developed using sophisticated machine learning strategies and presented significant advances in prediction of student retention. Academic performance, engagement behaviors, and socioeconomic status have all been recognized as student-related variables and machine learning has been used to create nuanced insights that traditional statistical models have not identified. Recent research has shown how multiple machine learning models, particularly ensemble models, have improved the accuracy of retention predictions.

Utilizing decision trees and/or its variants to predict student retention has been a prevalent strategy in EDM. For example, Hadiyanoor et al. found strong utility for the C4.5 algorithm in classifying student majors, with implications for reducing secondary education drop-outs [4]. As demonstrated by Khalilia et al. decision trees are valuable for parsing through academic profiles, providing useful insights into elements effecting student performance, or consequently retention rates. Furthermore, Goundar et al. employed support vector machines (SVM) and random forests to assess patterns of communication with students and improved the predictive power of student retention models [41]. These works confirm the valuable contribution of decision tree algorithms like Random Forests in parsing the complexity of educational data sets relevant to retention analytics.

Comparative studies of ensemble learning techniques, and especially Random Forests and Gradient Boosting machines (GBM), have produced promising results in improving the accuracy of predictions in regard to student dropout. Wong and Yip noted the important of finding a balance between data characteristics and the machine

learning techniques we select, emphasizing that decision trees, when possible, combined with ensemble or combined machine learning techniques demonstrate a substantial increase in model accuracy compared to other machine learning algorithms while considering multiple predictors (e.g. GPA and engagement) [27] . Similarly, Desfiandi and Soewito presented an ensemble model that effectively combined multiple classifiers that they showed was superior to other techniques in prediction time-to-graduation, also an important measure of retention [3]. The ensemble methods take advantage of the strengths of individual classifiers, giving high performance across a wide range of educational datasets.

The likelihood of a student dropping out of school is commonly performed using regression techniques where logistic regression has been widely used with additional methods such as decision trees and SVMs, etc., gaining popularity. For example, Nadeem and Palaniappan studied postgraduate dropout rates using a number of machine learning methods and showed that decision trees provided some significant accuracy measures [34]. Furthermore, SVM has been useful in predicting academic performance, but additional research is needed to identify its particular suitability within the scope of retention prediction frameworks [36]. These studies suggest the significance of merging the multitude of variety of machine learning methods to allow practitioners to leverage and bolt together the data to develop better practices and processes around student retention activities.

Machine Learning Models for Dropout Prediction

Recently, the application of machine learning (ML) models to predict student dropout has received considerable attention within educational technology and analytics. This field utilizes different algorithms to make predictions based on the characteristics of at-risk students through multiple pathways (academic performance; demographics and psychological aspects). Random Forests (RF) and Logistic Regression (LR) are two common examples of algorithms and models used for this purpose due to their robustness and interpretability allowing for a broad variety of data to be incorporated [42]. These models have the capacity to analyze complex multidimensional data, enabling postsecondary institutions and educators to intervene quickly to assist students who are deemed at-risk of dropping out.

Deep Learning (DL) techniques have also become useful for dropout prediction. Neural Networks, in particular, are appreciated for their ability to detect complex patterns in large datasets which many ML methods can miss [28]. However, it is worth noting that DL models involve significant computational capacity and often large datasets in order to perform well. Recent studies have successfully used Neural Networks to find weak indicators of dropout, demonstrating the promise and challenge of using Neural Networks [5][22].

The dropout data is an imbalance data issue, and although there are solutions, is a significant hurdle. Traditionally, there are significantly fewer dropouts than there are students who stay school; thus, there can be bias when predicting from the models [42]. There are also various techniques for balancing out the datasets - oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) can yield improvement in their performance and reliability [39]. SMOTE generates synthetic samples for the minority class (the dropouts) in order to obtain higher predictive accuracy from several machine learning algorithms to meet the needs of dropout prediction models with actual dropouts [39].

Many methodologies have been established that effectively identify predictors of dropout. For instance, some research on MOOCs has described various dropout predictors and used multiple ML and DL methods illustrating that multi-VI based approaches often produce higher accuracy results [48]. Also, the combination of deep learning techniques with traditional machine learning methods create stronger predictive models and stimulate more flexibility and robustness for enlarging training data with oversampling techniques [39].

Challenges in Dropout Prediction

A central issue in dropout prediction is the interpretability of machine learning models. While approaches such as Random Forests and Gradient Boosting provide some type of insight because feature importance scores can explain a model's decision-making, many of the more holistic models, i.e., neural networks, can be characterized as "black boxes." This is a concern for educators and administrators who want to make data-informed decisions about interventions, and who need to understand the "why" behind predictions [9]. Ifenthaler and Yau note the

need for learning analytics systems that explain the prediction mechanisms behind identifying at-risk students [9]. Otherwise, practical implementations of these models in educational settings remain ambiguous and may facilitate their use in real-world decision-making.

Data availability and quality are additional significant barriers affecting dropout prediction models. Many educational contexts face issues of missing and incomplete data, especially in the non-academic realms such as mental health or social participation [31][30]. According to Mihăescu and Popescu, incomplete datasets can severely undermine the precision of predictive models, meaning this research must improve the data acquisition methodologies used, along with quality assurance [33]. To help resolve these gaps, appropriate imputation methods or linkage with other data sources could be employed [45]. Also, some datasets exist, such as the PEEK dataset, that can be examined for new avenues for improving learner engagement, which can ultimately support improving prediction models [40]. Educational data mining (EDM) and machine learning combine well to facilitate thorough analyses; although, the barriers such as dataset imbalance remain a challenge that needs to be addressed, particularly to maximize the opportunities for these new strategies [43].

In addition, the utilization of a wider array of data sources is important to increase the effectiveness of predicting dropouts. Current models largely rely on academic performance metrics alone which overlook important social and psychological drivers of engagement and withdrawal from courses. This sentiment has been stated in the literature assertively that said models should utilize more holistic data including social data, the school was not using data to measure student well-being [45][38][26]. Razghandi et al. detailed the importance of using many data analytics methods and data types said these types of data and method were also useful to gain a more nuanced level of understanding when it came to dropout and the creation targeted interventions [43].

RESEARCH METHODOLOGY

Research Design

This study follows a quantitative research design using data mining and machine learning regression techniques to predict student retention and dropout rates at Cronasia Foundation College Inc. The primary goal is to explore the relationships between various student-related features and their likelihood to either stay enrolled or drop out. The study aims to identify the key factors influencing student retention, including academic performance, socioeconomic status, student engagement, and psychological factors, and then build predictive models using machine learning techniques.

Data Collection

The data for this study were collected from Cronasia Foundation College Inc.'s student management system, which includes historical records on students' academic performances, demographics, attendance, financial status, and extracurricular participation. The dataset comprised the following:

- Academic performance data: GWA, semesters enrolled
- Demographic data: Age, gender, family income, socioeconomic status, scholarship
- Engagement data: Attendance rates, social engagements
- Psychosocial data: Survey responses related to student satisfaction, mental health status, and emotional well-being from guidance

The study will consider students who have completed at least one academic year at the institution and have data records available for analysis.

Data Preprocessing

During the preprocessing, several essential steps were performed to prepare the dataset for machine learning. First, missing values in the SEMESTERS ENROLLED and SCHOLARSHIP columns were handled by imputing

missing entries. While the SCHOLARSHIP column, which had a "nan" string, was cleaned by substituting NaN before being imputed with the most frequent category using the SimpleImputer method, the SEMESTERS ENROLLED column was imputed using the median. Then, in order to prevent the creation of erroneous ordinal associations, One-Hot Encoding was used for multi-category columns like COURSE and SCHOLARSHIP, whereas Label Encoding was used for binary categories such as GENDER, PARTICIPATION IN CLASS, and SOCIAL ENGAGEMENT.

Additionally, StandardScaler was used to scale numerical features such as GWA, SEMESTERS ENROLLED, and FAMILY INCOME in order to guarantee that all features were on the same scale, which is essential for some machine learning methods. To guarantee that only the best predictors of student retention and dropout were included in the model, feature selection was finally carried out using SelectKBest to preserve the most pertinent characteristics. The dataset was ready for reliable model training thanks to these preprocessing procedures, which successfully addressed problems like missing data, improper encoding, and feature scaling.

Exploratory Data Analysis (EDA)

To further understand the dataset and the factors influencing dropout rates and student retention, a variety of methodologies were used. In the first phase, the frequency of each category in the DROP OUT STATUS column was shown using a count plot to visualize the distribution of student retention and dropout rates. This visualization showed the balance between students who stayed enrolled and those who dropped out, providing an initial understanding of the data's class distribution.

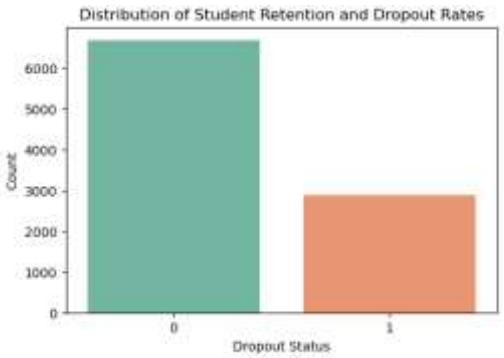

Figure 1: Distribution of Student Retention and Dropout Rates

Figure 1 shows the distribution of student retention and dropout rates using a count plot. This visualization illustrates that a larger proportion of students remain enrolled compared to those who drop out. The DROP OUT STATUS column is clearly split between two categories (0 = retained, 1 = dropped out). The most common category is student retention, indicating that the institution generally keeps its student body fairly steady. This distribution would imply that elements such as extracurricular activities, academic support, and scholarship possibilities help retain students.

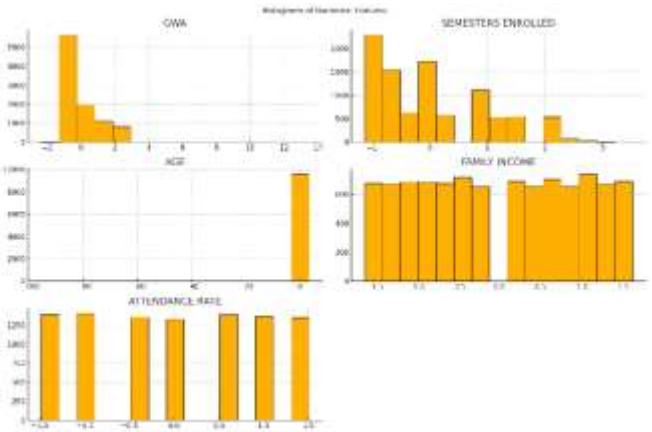

Figure 2: Histograms of Numerical Features

Figure 2 shows histograms for numerical features like GWA, ATTENDANCE RATE, and FAMILY INCOME. The majority of students have a GPA in the median range, with a small bias towards lower grades, according to the GWA histogram. This may suggest that a sizable portion of students are at risk of dropping out as a result of academic difficulties. The majority of students had high attendance rates, according to the ATTENDANCE RATE histogram, which is consistent with the idea that students who attend class frequently are more likely to remain enrolled. Contrarily, FAMILY INCOME indicates a bias toward lower-income households, which may have an impact on dropout rates because of financial limitations.

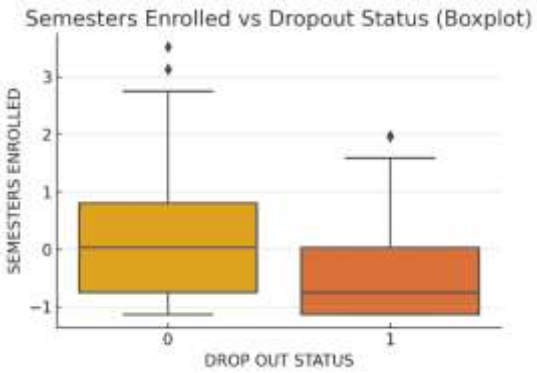

Figure 3: Box Plots of GWA and Semesters Enrolled by Dropout Status

Figure 3 presents box plots of GWA and SEMESTERS ENROLLED against DROP OUT STATUS. The GWA plot demonstrates that academic performance is a significant predictor of student retention, with dropouts typically receiving worse grades. In a similar vein, dropouts typically had fewer semesters enrolled, which may suggest that prior academic or personal difficulties caused them to leave the school. These charts offer compelling visual proof of the variables affecting student retention.

Model Development and Evaluation

The data set was submitted for model building and evaluation after it was preprocessed, including feature scaling, categorical encoding, imputation of missing values, and feature selection. Several machine learning models were developed in this phase to estimate dropout and retention rates for students.

Logistic regression was first used for binary classification and as a baseline model for estimating whether students but would drop out or stay enrolled. The logistic function provides the likelihood of an event occurring based on a linear combination of input features. Logistic regression has the ability to identify linear relationships between the outcome variable and the predictors, despite its basic nature.

Decision trees were subsequently used because they produce a more interpretable model which intends to split the data set based on the most important features at each node, ultimately classifying whether students are categorized as retained or have dropped out. Decision trees are easy to understand and interpret, they will sometimes overfit, particularly with deeper trees.

Random Forests, which utilize the predictions of multiple decision trees to combine them in order to lower the risk of overfitting as well as increase the robustness of the model were used to help solve issue. The final classification outcome is determined by averaging or voting from predictions of numerous decision trees constructed from bootstrapped samples of the dataset as part of the random forests.

Support Vector Machines (SVM) were also employed based on the concept of finding the optimal hyperplane to maximize the margin between classes. SVM is also effective in high dimensional domains as well as in cases when the dimensions are not linearly separable, since it can transform the input space through the kernel method. The SVM classifier optimizes the decision boundary to minimize the classification error, where the SVM classifies the data points by which side of the hyperplane they land on.

Lastly, Neural Networks (specifically the Multi-Layer Perceptron Classifier) were employed to model the

potentially complex relationships in the data. Neural networks can model non-linear patterns among the features and interactions that simpler models (like logistic regression) will miss. The backbone of the learning process in neural networks is adjusting weights, or parameters, by minimization of errors in prediction to the loss function using a mechanism known as backpropagation. Although neural networks tend to be more computationally demanding, they have the potential to outperform traditional approaches when modelling complex datasets.

The prediction models were evaluated based on their accuracy and precision, recall, and F1-score, and the outcomes were compared to help isolate the best performing model to forecast student retention and dropout. The evaluation highlighted the advantages and disadvantages of each method and provided guidance for deciding the best model for the real-world use.

Results Interpretation

Accuracy:

With an accuracy of 80.58%, the Decision Tree classifier was the most accurate model. Neural networks (80.42%) and logistic regression (73.98%) were next in line. All models appear to have performed quite well, based on the relatively similar accuracy scores of Random Forest (79.90%) and SVM (79.48%). Decision Tree, on the other hand, is notable for its accuracy, showing that it was best at differentiating between students who dropped out and those who stayed enrolled.

Precision:

With precision values of 0.84 for Class 0 (Retention) and 0.69 for Class 1 (Dropout), the neural network produced the best results for both classes. This high precision shows that the Neural Network was very successful in correctly recognizing both retained students and dropouts. Precision is a measure of a model's ability to correctly identify positive cases. In comparison to the Neural Network, the Decision Tree model did well for Class 0 (0.83), but its precision for Class 1 was marginally worse (0.71), indicating that the Neural Network is better at accurately identifying dropouts.

Recall:

With the highest recall of 0.91 for Class 0 (Retention), logistic regression showed a strong ability to identify students who were retained. However, as seen by its poor recall for Class 1 (0.32), Logistic Regression has trouble detecting dropouts. This suggests that students who are at risk of dropping out can be predicted less accurately using logistic regression. However, with a more balanced recall of 0.90 for Class 0 and 0.58 for Class 1, the Neural Network model proved to be a reliable one for detecting both dropouts and retained pupils.

F1-Score:

The F1-Score provides a thorough assessment of a model's performance by striking a balance between precision and recall. The Neural Network was the most balanced model in terms of precision and recall, as evidenced by its greatest F1-Score for both Class 0 (Retention) (0.87) and Class 1 (Dropout) (0.63). Random Forest and Decision Tree both did well; however, for Class 1 (Dropout) predictions, Decision Tree marginally outperformed Random Forest, indicating that Decision Tree may be better at identifying dropout-related trends.

Table 1. Results of the experiment

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	73.98%	0.767	0.911	0.833	0.593	0.319	0.415

Decision Tree	80.58%	0.835	0.906	0.869	0.707	0.560	0.625
Random Forest	79.90%	0.837	0.891	0.863	0.681	0.572	0.622
SVM	79.48%	0.833	0.890	0.860	0.674	0.562	0.613
Neural Network	80.42%	0.840	0.895	0.867	0.692	0.582	0.632

Table 1 illustrates a comparison of five different machine learning models, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) and Neural network, for predicting student retention and student dropout rate. While the table includes performance metrics - Accuracy, Precision, Recall and F1 Score for each of Class 0 (Retention) and Class 1 (Dropout). The Neural Network had the highest accuracy and relatively good balance between precision and recall, therefore Neural Network was the best model overall. Decision Tree and Random Forest are closely matched - both performing quite well overall and particularly well with respect to predicting dropouts. Logistic Regression was overall not successful in identifying Class 1(Dropout) and had a significantly lower recall with respect to dropouts.

Neural Network stands out as the most accurate and balanced model overall, with high precision, recall, and F1-Score for both retention and dropout predictions. Its ability to appropriately identify both Class 0 (Retention) and Class 1 (Dropout) makes it the most capable option for predicting student retention and dropout. Decision Tree and Random Forest also performed reasonably well, with Decision Tree showing slightly better performance than Random Forest, particularly for Class 1 (Dropout) in terms of F1-Score. This indicates that Decision Tree may be better at dropout identification than Random Forest, but both models appear to have similar predictive capabilities.

Logistic Regression, while exhibiting a strong capacity to predict Class 0 (Retention), had quite serious shortcomings in predicting class 1 (Dropout). Therefore, it was not the right model for predicting the likelihood of students dropping out. Since predicting students dropping out is an important task, we can assess the Recall value for class 1, which means that Logistic Regression had failed to identify a range of students likely at risk of dropping out.

In conclusion, based on the evaluation results, Neural Network is the most appropriate model for the task, while the Decision Tree provides a second viable option for predicting dropout when dropout prediction is paramount. Future research could focus on hyper-tuning these models, or possibly hybrid approaches to enhance the predictive performance overall.

CONCLUSION

The results of the model evaluation showed Neural Network was superior to the other models at predicting retention and dropout rates, as Neural Network had the best accuracy, precision, recall, and F1-score for both Class 0 (Retention) and Class 1 (Dropout). Random Forest and Decision Tree both had comparable prediction capabilities overall, but Decision Tree narrowly improved on Random Forest in predicting dropout rates. Logistic Regression demonstrated significant limitations in predicting dropout rates, considering its low recall rate specifically for Class 1 (Dropout). The results of Logistic Regression suggest it may not be a suitable model for situations that do require dropout predictions.

Insights from the important factors affecting student retention and dropout rates indicated several important variables that were most critical. The factors that had the most significant effect on student retention were Academic Performance (GWA), Attendance Rate, and then Scholarship Status. Therefore, it appeared that students with a higher Academic Performance (higher grades to stay enrolled) and better Attendance Rate were also more likely to stay enrolled. In addition, Scholarship Status clearly had a significant effect on retention rate. We better retained students receiving scholarships or student aid compared to those who were not receiving

scholarships or student aid, which was observed within the models. In addition to the measures of Academic Performance, Attendance Rate, and Scholarship Status, the Behavioral and Engagement Factors (notably, use of sports and extracurricular activities, mental health, and behavioral issues) also had a significant influence on predicting retention. Students who were more engaged and had better mental health typically had a higher likelihood for retention; yet, students who experienced behavioral issues and less engaged tended to have higher chances to drop out.

Overall, this study demonstrates that academic performance, attendance, access to scholarships and engagement of students have the most impact on retention/dropout, when we apply the newest Neural Network model in evaluating the numerous interactions between these variables. It has also shown that the Neural Network is a better method to determine the predicted outcomes and make informed decisions to increase retention.

REFERENCES

1. A. Bombaes, J. Fuasan, & W. Garcia, "Exploring the factors in student's retention of e-learning mathematics: a case of grade 12 senior high school students at the university of perpetual help system-pueblo de panay campus", International Journal of Education Teaching and Social Sciences, vol. 1, no. 1, p. 1-7, 2021. https://doi.org/10.47747/ijets.v1i1.341
2. A. Cheong, P. Singh, N. Saat, & J. Hoon, "Retention amongst pre-university students at a foreign university branch campus in malaysia: an exploratory study", Journal of Education and Learning, vol. 10, no. 3, p. 39, 2021. https://doi.org/10.5539/jel.v10n3p39
3. A. Desfiandi and B. Soewito, "Student graduation time prediction using logistic regression, decision tree, support vector machine, and adaboost ensemble learning", Ijiscs (International Journal of Information System and Computer Science), vol. 7, no. 3, p. 195, 2023. https://doi.org/10.56327/ijiscs.v7i2.1579
4. A. Hadiyanoor, S. Cholifah, H. Junaidi, & I. Febrian, "Using c4.5 decision tree to determine the majors of students in sman 4 banjarmasin to reduce the cause of dropout from school", Iiai Letters on Informatics and Interdisciplinary Research, vol. 5, p. 1, 2024. https://doi.org/10.52731/liir.v005.209
5. A. Nabil, M. Seyam, & A. AbouElfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks", Ieee Access, vol. 9, p. 140731-140746, 2021. https://doi.org/10.1109/access.2021.3119596
6. B. Flores-Caballero, "Higher education: factors and strategies for student retention", Hets Online Journal, vol. 10, no. 2, p. 82-105, 2022. https://doi.org/10.55420/2693.9193.v10.n2.14
7. C. Li, N. Herbert, S. Yeom, & J. Montgomery, "Retention factors in stem education identified using learning analytics: a systematic review", Education Sciences, vol. 12, no. 11, p. 781, 2022. https://doi.org/10.3390/educsci12110781
8. C. Panda, K. Christopher, A. Paswan, D. Patel, & R. Sohane, "Students perception on enrolment factors in their retention in higher agricultural education", Current Journal of Applied Science and Technology, p. 107-113, 2020. https://doi.org/10.9734/cjast/2020/v39i630565
9. C. Wekullo, "Institution type, selectivity, and financial aid: an examination of institutional factors influencing first-time students retention in public universities", Social Education Research, p. 1-14, 2022. https://doi.org/10.37256/ser.4120231725
10. D. Ifenthaler and J. Yau, "Utilising learning analytics to support study success in higher education: a systematic review", Educational Technology Research and Development, vol. 68, no. 4, p. 1961-1990, 2020. https://doi.org/10.1007/s11423-020-09788-z
11. D. Rodgers-Tonge, M. Wray, & C. Baldwin, "Supportive programs and financial aid: measuring their impact on retention of blacks and latinx college students in the new england region", Journal of Business Diversity, vol. 23, no. 4, 2023. https://doi.org/10.33423/jbd.v23i4.6614
12. D. Shafiq, M. Marjani, R. Habeeb, & D. Asirvatham, "Student retention using educational data mining and predictive analytics: a systematic literature review", Ieee Access, vol. 10, p. 72480-72503, 2022. https://doi.org/10.1109/access.2022.3188767
13. E. Bambacus and A. Conley, "The impact of dosage on a mindfulness intervention with first-year college students", Journal of College Student Retention Research Theory & Practice, vol. 25, no. 4, p. 979-1000, 2021. https://doi.org/10.1177/15210251211041695

14. E. Sousa, B. Rosa, R. Mello, T. Falcão, B. Vesin, & D. Gašević, "Applications of learning analytics in high schools: a systematic literature review", Frontiers in Artificial Intelligence, vol. 4, 2021. https://doi.org/10.3389/frai.2021.737891
15. F. Alshareef, H. Alhakami, T. Alsubait, & A. Baz, "Educational data mining applications and techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1-8, 2020. https://doi.org/10.14569/ijacsa.2020.0110494
16. F. Tan, J. Lim, W. Chan, & M. Idris, "Computational intelligence in learning analytics: A mini review," *Asean Engineering Journal*, vol. 14, no. 4, pp. 135-151, 2024. https://doi.org/10.11113/aej.v14.21375
17. G. Gonçalves, F. Serra, J. Storópoli, I. Scafuto, & D. Rafael, "Undergraduate student retention activities: challenges and research agenda**",** Sage Open, vol. 14, no. 3, 2024. https://doi.org/10.1177/21582440241249334
18. G. Oswald, R. DuVivier, S. Wood, & T. Freeman, "Surviving and thriving at a uk university through a minority lens.", Journal of the Australian and New Zealand Student Services Association, vol. 29, no. 1, p. 35-51, 2021. https://doi.org/10.30688/janzssa.2021.1.05
19. G. Sani, F. Oladipo, E. Ogbuju, & F. Agbo, "Development of a predictive model of student attrition rate," *Journal of Applied Artificial Intelligence*, vol. 3, no. 2, pp. 1-12, 2022. https://doi.org/10.48185/jaai.v3i2.601
20. H. Aal, "Academic self-esteem and its relationship to practicing extracurricular activities among university students", Cypriot Journal of Educational Sciences, vol. 18, no. 1, p. 228-238, 2023. https://doi.org/10.18844/cjes.v18i1.8306
21. H. Al-Kadri, N. Nellitawati, S. Syahril, E. Ramli, J. Jasrial, L. Susantiet al., "Analyzing of extracurricular program management technical in junior high school",, 2020. https://doi.org/10.4108/eai.11-12-2019.2290899
22. H. Khalilia, T. Sammar, & Y. Sleet, "Predicting students performance based on their academic profile", مجلة جامعة فلسطين التقنية للأبحاث, vol. 8, no. 2, p. 23-39, 2020. https://doi.org/10.53671/pturj.v8i2.91
23. I. Salehin and D. Kang, "A review on dropout regularization approaches for deep neural networks within the scholarly domain", Electronics, vol. 12, no. 14, p. 3106, 2023. https://doi.org/10.3390/electronics12143106
24. J. Clement and P. Mwila, "Extracurricular activities: prospects and challenges among female students in secondary schools in chanika ward, tanzania", IJSSMR, vol. 01, no. 01, p. 14-30, 2023. https://doi.org/10.61421/ijssmer.2023.1102
25. J. Jamaluddin, S. Syam, S. Saleh, & N. Nasrullah, "The influence of extracurricular activities on character building of students of smpn 22 makassar", Jurnal Office, vol. 7, no. 1, p. 1, 2021. https://doi.org/10.26858/jo.v7i1.18989
26. J. Norvilitis, H. Reid, & K. O'Quin, "Amotivation: a key predictor of college gpa, college match, and first-year retention", International Journal of Educational Psychology, vol. 11, no. 3, p. 314-338, 2022. https://doi.org/10.17583/ijep.7309
27. J. Swacha and K. Muszyńska, "Predicting dropout in programming moocs through demographic insights", Electronics, vol. 12, no. 22, p. 4674, 2023. https://doi.org/10.3390/electronics12224674
28. J. Wong and T. Yip, "Measuring students' academic performance through educational data mining", International Journal of Information and Education Technology, vol. 10, no. 11, p. 797-804, 2020. https://doi.org/10.18178/ijiet.2020.10.11.1461
29. K. Talebi, Z. Torabi, & N. Daneshpour, "Predicting mooc dropout using ensemble models based on rnn and gru",, 2024. https://doi.org/10.21203/rs.3.rs-5243770/v1
30. L. Cagliero, L. Canale, L. Farinetti, E. Baralis, & E. Venuto, "Predicting student academic performance by means of associative classification", Applied Sciences, vol. 11, no. 4, p. 1420, 2021. https://doi.org/10.3390/app11041420
31. M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. Raza, M. Abidet al., "Predicting at-risk students at different percentages of course length for early intervention using machine learning models", Ieee Access, vol. 9, p. 7519-7539, 2021. https://doi.org/10.1109/access.2021.3049446
32. M. Amare and S. Šimonová, "Global challenges of students dropout: a prediction model development using machine learning algorithms on higher education datasets", SHS Web of Conferences, vol. 129, p. 09001, 2021. https://doi.org/10.1051/shsconf/202112909001

33. M. Elobaid, R. Elobaid, L. Romdhani, & A. Yehya, "Impact of the first-year seminar course on student gpa and retention rate across colleges in qatar university", International Journal of Learning Teaching and Educational Research, vol. 22, no. 5, p. 658-673, 2023. https://doi.org/10.26803/ijlter.22.5.34
34. M. Mihăescu and P. Popescu, "Review on publicly available datasets for educational data mining", Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, vol. 11, no. 3, 2021. https://doi.org/10.1002/widm.1403
35. M. Nadeem and S. Palaniappan, "Predictive model of postgraduate student's dropout and delay using machine learning algorithms", International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 2, p. 894-900, 2021. https://doi.org/10.30534/ijatcse/2021/591022021
36. M. Peralta and J. Vunueza-Martinez, "Application of academic analytical models in education management", Journal of Educational and Social Research, vol. 14, no. 6, p. 274, 2024. https://doi.org/10.36941/jesr-2024-0171
37. N. Samsudin, S. Shaharudin, N. Sulaiman, S. smail, N. Mohamed, & N. Husin, "Prediction of student's academic performance during online learning based on regression in support vector machine", International Journal of Information and Education Technology, vol. 12, no. 12, p. 1431-1435, 2022. https://doi.org/10.18178/ijiet.2022.12.12.1768
38. O. Rotar, "A missing theoretical element of online higher education student attrition, retention, and progress: a systematic literature review", Sn Social Sciences, vol. 2, no. 12, 2022. https://doi.org/10.1007/s43545-022-00550-1
39. S. Amjad, M. Younas, M. Anwar, Q. Shaheen, M. Shiraz, & A. Gani, "Data mining techniques to analyze the impact of social media on academic performance of high school students", Wireless Communications and Mobile Computing, vol. 2022, p. 1-11, 2022. https://doi.org/10.1155/2022/9299115
40. S. Ashraf, S. Saleem, T. Ahmed, Z. Aslam, & D. Muhammad, "Conversion of adverse data corpus to shrewd output using sampling metrics", Visual Computing for Industry Biomedicine and Art, vol. 3, no. 1, 2020. https://doi.org/10.1186/s42492-020-00055-9
41. S. Bulathwela, M. Pérez-Ortiz, E. Novak, E. Yılmaz, & J. Shawe-Taylor, "Peek: a large dataset of learner engagement with educational videos",, 2021. https://doi.org/10.48550/arxiv.2109.03154
42. S. Goundar, A. Deb, G. Lal, & M. Naseem, "Using online student interactions to predict performance in a first-year computing science course", Technology Pedagogy and Education, vol. 31, no. 4, p. 451-469, 2022. https://doi.org/10.1080/1475939x.2021.2021977
43. S. Lai, N. Shahri, M. Mohamad, H. Rahman, & A. Rambli, "Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data", Mathematics and Statistics, vol. 9, no. 3, p. 379-385, 2021. https://doi.org/10.13189/ms.2021.090320
44. S. Radovanović, B. Delibašić, & M. Suknović, "Predicting dropout in online learning environments", Computer Science and Information Systems, vol. 18, no. 3, p. 957-978, 2021. https://doi.org/10.2298/csis200920053r
45. T. Cardona, E. Cudney, R. Hoerl, & J. Snyder, "Data mining and machine learning retention models in higher education", Journal of College Student Retention Research Theory & Practice, vol. 25, no. 1, p. 51-75, 2020. https://doi.org/10.1177/1521025120964920
46. T. Panagiotakopoulos, S. Kotsiantis, G. Kostopoulos, O. Iatrellis, & A. Kameas, "Early dropout prediction in moocs through supervised learning and hyperparameter optimization", Electronics, vol. 10, no. 14, p. 1701, 2021. https://doi.org/10.3390/electronics10141701
47. U. ÖZKAN, "The effect of students' participation in extracurricular activities on academic achievement according to pisa-2015", İnönü Üniversitesi Eğitim Fakültesi Dergisi, vol. 21, no. 1, p. 254-269, 2020. https://doi.org/10.17679/inuefd.504780
48. X. Liu, T. Wang, D. Bressington, B. Easpaig, L. Wikander, & J. Tan, "Factors influencing retention among regional, rural and remote undergraduate nursing students in australia: a systematic review of current research evidence", International Journal of Environmental Research and Public Health, vol. 20, no. 5, p. 3983, 2023. https://doi.org/10.3390/ijerph20053983
49. Z. Sun, A. Harit, J. Yu, A. Cristea, & L. Shi, "A brief survey of deep learning approaches for learning analytics on moocs",, p. 28-37, 2021. https://doi.org/10.1007/978-3-030-80421-3_4