JOURNAL ON COMMUNICATIONS

ISSN:1000-436X

Scopus®

REGISTERED

www.jocs.review

A Comparative Study of Advanced ML-Based Forecasting Models for Higher Education Dropouts via Students

Sandeep Gupta[1] and Sumeet Mathur [2]

[1] SATI, Vidisha.
[2] University of Waikato NZ - Joint Institute at Zhejiang University, Hangzhou, China.

Abstract— Student dropout is an extreme issue because it negatively affects the former institution, family, and society at large, not to mention the individual student who dropped out. To handle this problem, several efforts to use ML to forecast students dropping out have been made. The aim of the present research is to estimate the student dropout rate by testing the capabilities of several ML models to interact with multifaceted patterns and non-balanced data. LightGMB scored the largest at 86.14, which is above the rates of other models such as Gradient Boosting (79.40) and multi-layer perceptron (78.52). The good performance of LightGBM means that it is able to offer justifiable and fair findings on the other categories of students like dropout, enrolment, and graduation. This software could assist schools with detecting high-risk students at an earlier stage and spending resources more effectively to assist them and use advanced algorithms. The findings explain why machine learning has the potential to revolutionize student retention plans with data-driven insights that would enhance decision-making. In general, the article, as well, can be placed in the already-too-long queue of articles that permit the application of predictive analytics in schools in order to enhance student academic performance and dropout rates.

Keywords—Student Dropout, Higher Education, Retention, Academic Performance, Machine Learning, classification.

1 Introduction

As a foundation for creativity, education plays an important role in shaping individuals and societies as well as in economic growth and social mobility[1]. Delivery of higher level of knowledge and skills as well as the success of the student in their educational programs is important to the role of the higher education institutions in this educational system[2]. One of the most essential pointers of institutional quality and effectiveness is student performance and the achievement of academic consistency among the various groups of students is a major objective of most universities and colleges around the world [3][4][5]. Student dropout has become one of the biggest issues facing an institution of higher education in the world despite emphasis on student success. The negative consequences of early student dropouts to the institutions include financial impacts on the institutions, decreased institutions reputation and social economic impacts on the affected students in the long term [6][7]. Despite years of research and intervention strategies since the 1970s, dropout rates are alarmingly high. To illustrate, in the United States, nearly forty percent of students do not graduate with an undergraduate degree within six years and the dropout rate of open access and online education systems such as MOOCs can reach 95%. These figures indicate the high level of urgency of developing more effective solutions to curb student attrition [8][9].

Consequently, the discourse on Student Dropout Prediction in Higher Education has seen a spurt of interest in recent years, aiming to discover answers to the major cause and effect factors of the loss of at-risk students and whether this can be identified early. Already, institutions have begun to value the importance of taking proactive measures, on the basis of timely and precise information, in order to be successful in boosting student retention [10][11]. Digitization of student records (e.g. academic transcripts, demographic profiles, and financial status) has brought about new possibilities in modelling and understanding dropout behavior [12][13]. Machine learning (ML) has become a useful part of predictive analytics to leverage this data to the fullest. The ML algorithms have the ability to analyze big and complicated data and expose previously untapped patterns and make predictions, thus potentially resulting in evidence-based decision-making [14][15]. Applying these algorithms to the dropout predictions challenge, educational institutions will be able to create early warning systems, which can address the needs of at-risk students and offer them interventions such as financial grants and academic counselling on time. The magnitude of the algorithmic, dataset, and dropout-rate-based class imbalance, that training influences the performance of such predictive models, is extremely large.

1.1 Motivation and Contribution

Student drop out is a principal issue in higher education institutions. There are high rates of dropouts, which lead to financial losses and damage the reputation of universities, which also has a negative impact on the future of students. Traditional methods to identify at-risk students are often slow and not very accurate. With more digital data available, such as academic records and demographic information, machine learning offers new ways to predict dropout risk. These models can analyze complex data and spot patterns that humans might miss. However, not all machine learning algorithms perform equally well. Their success depends on factors like data quality and how balanced the dropout rates are. This makes it important to compare different algorithms. By doing so, they can identify the most accurate and reliable methods for predicting dropout. This will help institutions to intervene early and assist students better. Finally, the research will improve student retention and the smartest use of education resources. The key contributions of this paper are:

- Created a predictive framework that uses machine learning algorithms to identify students who are at danger of dropping out.
- Data cleaning with missing value treatment, drop duplicates, feature extraction and normalization to feed the model were also robust.
- Examined the relationship between features and data through EDA tools such as KDE plots, bar charts, box plots and heatmaps in order to identify significant patterns.
- The resolution was to apply ADASYN resampling to balance the classes such that the models can act more fairly and give improved results.
- Tested a variety of models (Gradient Boosting, LightGBM and MLP Classifier) to compare their efficacy in prediction and generalization.
- Conducted wide-ranging model evaluation given classification reports, confusion matrices, ROC curves, precision-recall curves on test sets.

1.2 Significance and Novelty

This research is very applicable in addressing the grave issue of student dropout since it offers a predictive model that enhances early identification of at-risk students significantly. Its innovativeness lies in its ability to control successfully the imbalance of data and nonlinear interactions with features on the best basis of machine learning, namely LightGBM, which also has the added advantage of being very precise and guiding the balance across a number of groupings of student results. The research combines both advanced data analysis methods and research mechanisms with the conventional methods in a bid to enhance the level of reliability and generalization of forecasts. The suggested framework seems highly viable in the context of educational establishments because it allows introducing a data-driven intervention early in the process to decrease dropout and enhance retention. This contribution is particularly valuable to make positive judgments of and to allocate resources to the academic literature due to the combination of methodological rigor and practice in educational data mining.

1.3 Structure of the Paper

This research paper is structured as follows: The Literature Review summarizes related studies present in Section 2. Section 3 provide Methodology with data collection, preprocessing, and model training. Results and Discussion compare model performances present in Section 4. The Conclusion and Future Work section of the report offers suggestions for enhancements to Section 5.

2 Literature Review

Educational institutions throughout the globe are deeply troubled by the alarmingly high student dropout rates. The causes of student dropout have been the subject of several investigations, including the use of NN and ML for the purpose of prediction and analysis.

Pérez et al. (2025) applies traditional ML and DL models to dorecast student dropout in an Ecuadorian HEI using the CRISP-DM methodology. A comprehensive analysis of demographic, academic, and economic factors was conducted to develop an effective predictive framework. The evaluated models include LR, SVM, RF, XGBoost, Feedforward Neural Network, and TabNet. Various configurations were tested, including the application of PCA for dimensionality reduction and the SMOTE to address class imbalance. Experimental results reveal that PCA and SMOTE are unnecessary. RF outperformed the other models with an F1-score of 0.94, a ROC-AUC of 0.92, and an accuracy of 96.62% [16].

Wang (2024) proposed for predicting optimal features from the Massive Open Online Courses (MOOC) dataset. The student's data are taken from the MOOC dataset which undergoes pre-processing by using the Z-score normalization technique to reduce feature dominance. Then, the normalised features are further processed into feature selection with DTO method to select optimal features. After that, the selected features are further processed with SVM classifier to classify student dropout rate as dropout or persist. The proposed DTO method gives better results than existing Logistic Regression (LR) model in terms of acc (0.95), pre (0.96), rec (0.98) and F1score (0.97) respectively [17].

Aisyah et al. (2024) employs ML algorithms to forecast the graduation outcomes of undergraduate students, focusing on the effectiveness of various algorithms. They used the Random Forest algorithm and compared its performance with Decision Tree, Nave Bayes, ANN, and SVM, utilizing the Orange data mining tool. Their evaluation metrics included AUC, acc, F1score, pre, and rec, applied to a dataset of 387 students from the class of 2017 who graduated on time in 2021 and 2022. Features such as ID, major, age, sex, and first- and second-semester GPAs were taken into account. A 95% success rate was reached with the RF algorithm [18].

Deb et al. (2024) dataset comprises information on over 400 students who enrolled in the university between 2015 and 2020, including their academic records, demographic characteristics, and enrollment history. Analyze datasets utilizing a variety of ML approaches like SVM, RF, LR, etc. The algorithm performance can be evaluated according to the F1score, Rec, precision, and accuracy. They conclude that student dropout may be predicted with high accuracy, precision and recall using ML algorithms. The best-performing algorithm is Random Forest with a Prec of 0.78, Rec of 0.78, and F1score of 0.78. LR and KNN algorithms also perform reasonably well, with a precision of 0.75 and 0.76, respectively [19].

Akter et al. (2024) purposes to forecast the rate of university student dropouts in Bangladesh using a number of ML methods. It used seven standard classifiers SVM, KNN, XGBoost, DT, LR, NB and RF to construct a ML model that could forecast the student dropout rate. The dataset was comprised of the responses to 29 questions given by approximately 500 college students in

Bangladesh. The set contains information with a variety of attributes. After the data was preprocessed, it was divided into two sets: training and testing. Next, the classifiers underwent training and testing. It rigorously tested all seven classifiers to find the one that worked best with this dataset. For the purpose of forecasting attrition rates, their results show that XGBoost and RF achieve an impressive 98.06% accuracy rate [20].

Dewi et al. (2023) utilize LMS access log data, student statistics, and computed data in an effort to come up with an appropriate algorithm for early dropout prediction systems in online learning using ML. Among the four algorithms tested, NB has the best recall at 1 and LR with Lasso the best precision at 1. The SVM, which shares a value with LR with Lasso, has the best accuracy at 0.99 and F1score at 0.97. The early dropout prediction model may help teachers and school officials identify kids at danger of dropping out so they can intervene swiftly to improve their academic performance and decrease the number of students who drop out [21].

Despite the promising results achieved by recent studies in predicting student dropout and graduation outcomes using various ML and DL techniques, several research gaps remain, as summarized in Table 1. Most existing works rely on institution-specific or region-specific datasets with limited demographic diversity, which restricts the generalizability of the developed models. Additionally, many studies focus primarily on academic and demographic factors, overlooking other influential aspects such as psychological, behavioral, and socio-economic variables that could enhance predictive accuracy. Furthermore, while techniques like PCA and SMOTE are often tested, the dynamic integration of real-time learning analytics and adaptive feature selection remains underexplored. Therefore, future research should address these limitations by developing scalable, multi-institutional frameworks that incorporate broader data dimensions and deploy models in real-world educational settings for proactive intervention and validation

Table 1: Comparative Analysis of Recent Studies on Student Dropout and Graduation using machine learning

Author	Methods	Dataset	Key Findings	Limitations & Future Work
Pérez et al. (2025)	Logistic Regression, SVM, Random Forest, XGBoost, FNN, TabNet; PCA; SMOTE; CRISP-DM	Ecuadorian HEI student data (demographic, academic, economic)	Random Forest achieved highest accuracy (96.62%), F1-score (0.94), ROC-AUC (0.92); PCA and SMOTE found unnecessary.	No significant benefit from PCA or SMOTE; future work may explore interpretability and real-time predictions.
Wang (2024)	DTO (feature selection), Z-score normalization, SVM	MOOC dataset	DTO+SVM outperformed Logistic Regression; F1-score: 0.97, Recall: 0.98, Accuracy: 0.95.	Focus limited to MOOC; generalizability to traditional HEIs is unverified.
Aisyah et al. (2024)	RF, DT, NB, ANN, SVM using Orange	387 Indonesian students (2017–2022)	Random Forest performed best with 95% accuracy; considered first and second semester GPAs.	Limited dataset size; restricted to a single institution and program years.
Deb et al. (2024)	SVM, Random Forest, Logistic Regression, KNN	400 students from Bangladesh (2015–2020)	RF performed best (Precision, Recall, F1 = 0.78); LR and KNN also performed well.	Precision/Recall scores were moderate; broader data diversity and temporal validation needed.
Akter et al. (2024)	SVM, KNN, XGBoost, DT, LR, NB, RF	500 university students from Bangladesh	XGBoost and Random Forest achieved best performance (Accuracy = 98.06%).	Future work could examine model interpretability and longitudinal validation.
Dewi et al. (2023)	NB, LR with Lasso, SVM	LMS access logs and student statistical data	SVM and LR with Lasso showed highest accuracy (0.99) and F1 (0.97); Naive Bayes had highest recall (1.0).	Focus on online learning only; real-time deployment and scalability should be explored.

3 Methodology

The methodology for predicting student dropout involves a structured sequence of steps to ensure accurate and reliable outcomes. It begins with the collection of Student Dropout Dataset from the UCI repository, which included handling missing and duplicated values, dropping irrelevant columns, and performing EDA. Feature selection was conducted using the ANOVA method, followed by robust scaling to handle outliers and ADASYN oversampling to balance class distribution. The dataset was then split into training and testing sets using an 80:20 ratio. The following metrics were used for training and evaluation purposes: classification reports, confusion matrices, ROC curves, precision-recall curves, and f1score. The three models that were trained and evaluated were MLP Classifier, Gradient Boosting, and LightGBM. The best method for forecasting student academic results was determined by analyzing each model's performance, guaranteeing a thorough assessment of both training and test data. Figure 1 shows the implementation steps.

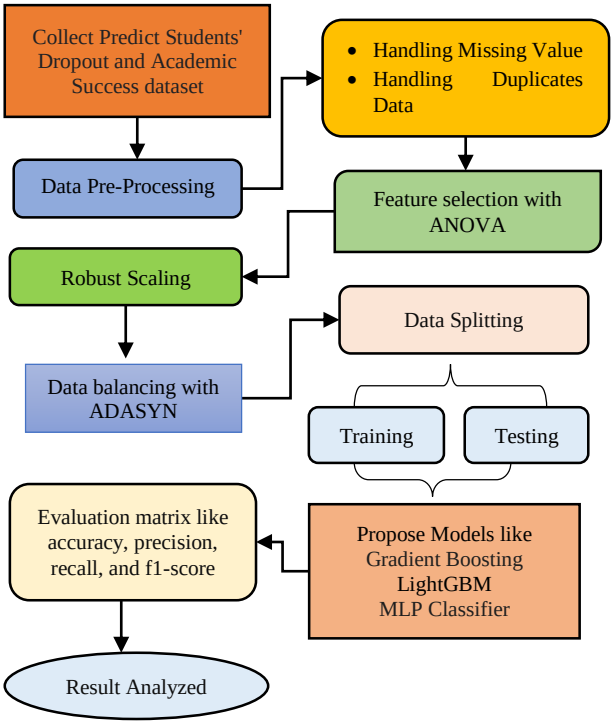


Figure 1: Proposed flowchart for Student Dropout Prediction

The following steps of the proposed methodology are briefly discussed below:

3.1 Data Collection

The Predict Students' Dropout and Academic Success dataset[1], y UCI ML Repository. It records 4,424 occurrences and 37 characteristics, including students' demographic, academic, and socioeconomic information known at enrolment, as well as their performance in the first and second semesters of the course. Early identification of at-risk children to direct timely interventions is made easier with this dataset, which is helpful for classification tasks in social science and education research, utilising real, categorical, and integer variables.

3.2 Exploratory Data Analysis

Before building a classification model, EDA is an important step. This enables us to select suitable machine learning algorithms by revealing hidden patterns in the data. Some of visualizations of dataset are given below:

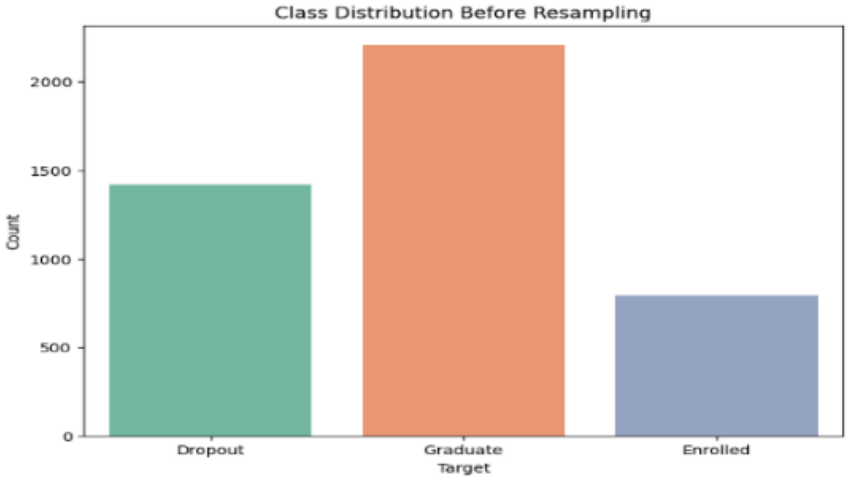


Figure 2: Count Plot for Distribution of Target

Figure 2 illustrates the count plot depicting the distribution of the target variable before any resampling was applied. It clearly shows an imbalance among the three categories: the majority of students fall into the Graduate class, followed by the Dropout

[1] https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

category, while the Enrolled group represents the smallest portion. This imbalance highlights the need for appropriate resampling techniques to ensure fair model training and accurate classification performance across all student outcomes.

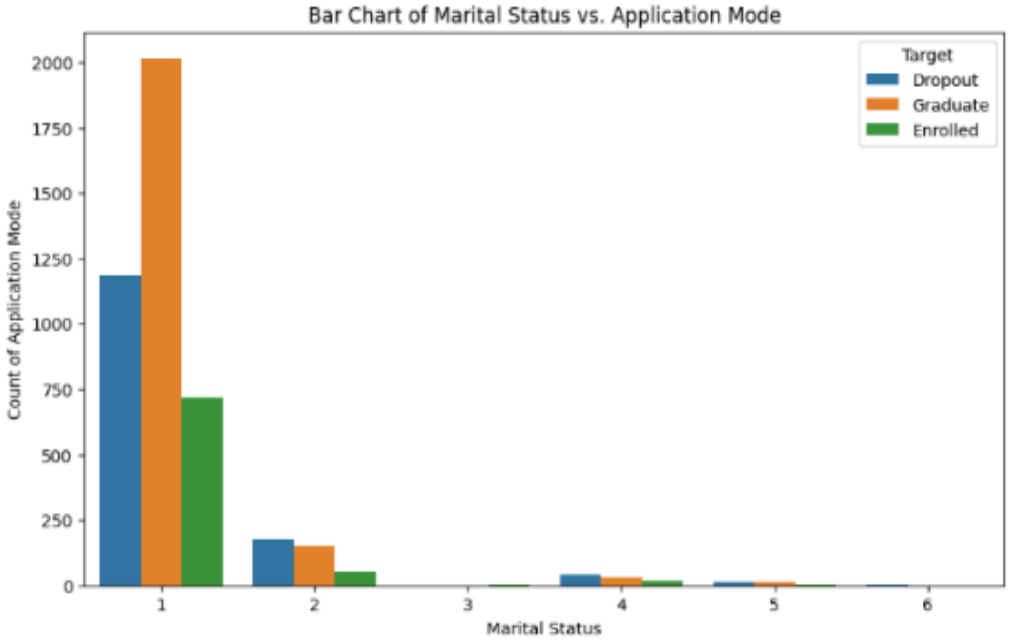


Figure 3: Bar Chart of Marital Status vs. Application Mode

Figure 3 displays the distribution of student outcomes (Dropout, Graduate, and Enrolled) across various Marital Status categories in a bar chart that also plots them against application mode. The majority of applications come from students with a marital status of 1, where most graduates, dropouts, and enrolled students are concentrated. As marital status increases from 2 to 6, the number of students in each category sharply decreases, indicating that single or unmarried students (likely represented by status 1) dominate the dataset, and Marital Status has a potential correlation with application and educational outcomes.

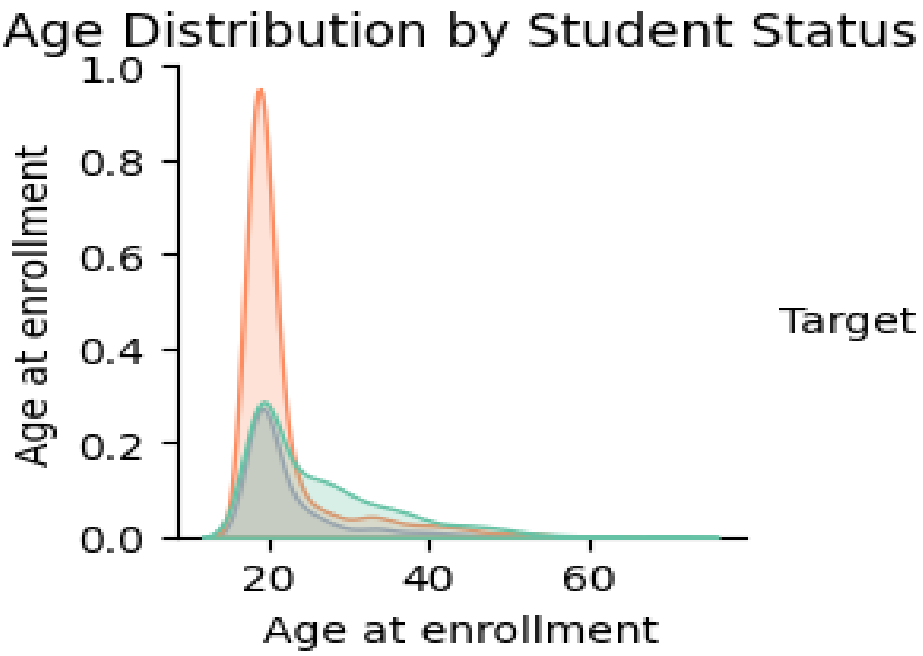


Figure 4: KDE Plot for Age Distribution by Student Status

Figure 4 presents the age distribution of students at enrollment across different student status categories. The plot indicates that the majority of students, regardless of whether it ultimately drop out, remain enrolled, or graduate, tend to enroll at a younger age, with a prominent peak around the late teens to early twenties. There is a noticeable long tail toward older ages, but these cases are less frequent. This visualization highlights the age concentration and slight variation in age profiles among the three groups, which can be an important factor in predicting student outcomes.

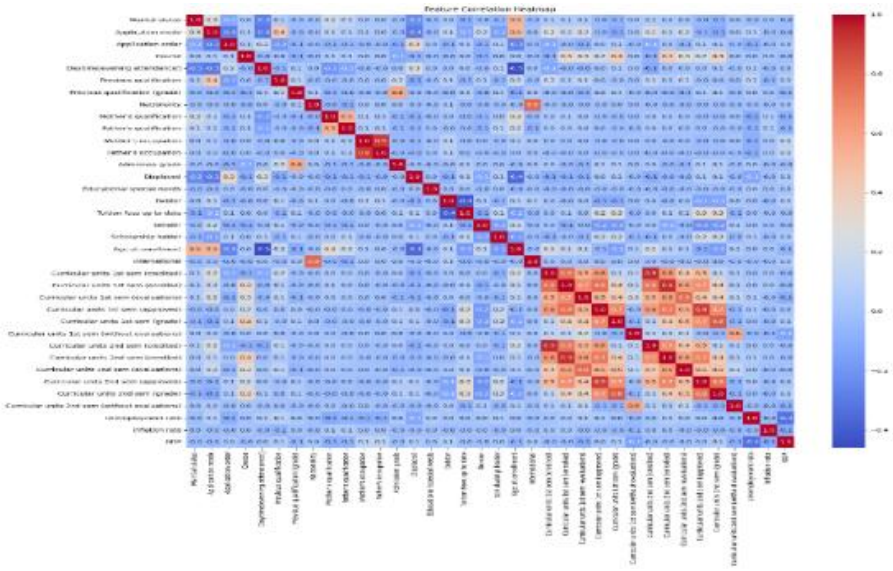


Figure 5: Correlation HeatMap of Features

The Correlation Heatmap visualizes relationships between features in the student dataset shows in Figure 5. Darker red squares indicate strong positive correlations, while blue squares indicate negative ones. The highest correlations are found among academic performance features, such as grades, approvals, and evaluations, across semesters. Most other features, such as demographics and socioeconomic factors, show weak or no correlation, suggesting limited direct influence on student performance or d ropout prediction.

3.3 Data Pre-Processing

Several stages of pre-processing improve the dataset's quality and structure, preparing it for use in ML models [22]. Data pre-processing involved handling missing and duplicate values, features selection, balancing classes with ADASYN, and scaling numeric data using Robust Scaler to prepare for accurate dropout prediction. The following steps of pre-processing are listed in below:

- **Handling Missing Values:** Missing values were then assessed using .isnull().sum(), and appropriate actions, like imputation or removal were applied to ensure the completeness and reliability of the data for subsequent processing.
- **Removing Duplicates:** duplicated rows in the dataset were identified and removed using the drop_duplicates() function, eliminating redundant information that could skew analysis. Additionally, irrelevant columns such as index numbers or ID fields that do not provide meaningful input for prediction were dropped to streamline the feature set and reduce noise in the model training process.

3.4 Feature Selection using ANOVA

Finding the most useful qualities for accurate categorization or prediction is the main goal of feature selection. Feature selection with ANOVA helps identify the most relevant features by comparing group means to see if a feature significantly affects the target variable. By calculating the F-ratio from the variance within and between groups, ANOVA removes features with no significant impact, reducing dimensionality and improving model performance. The bar chart shows the top 22 features ranked by ANOVA F-value scores.

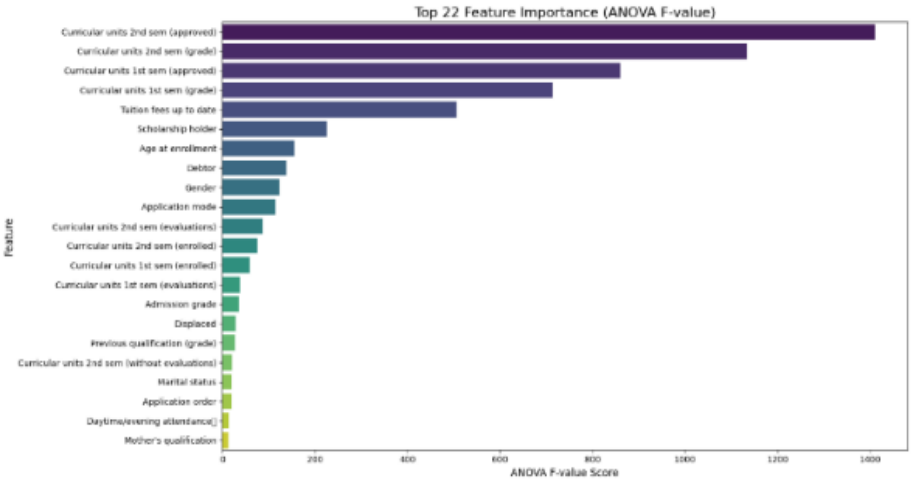


Figure 6: Top 22 Feature Importance Score

Figure 6 presents the top 22 feature importance scores based on ANOVA F-values, highlighting the most influential variables for the model. "CurricularUnits 2nd sem (approved)", "CurricularUnits 2nd sem (grade)", and "CurricularUnits 1st sem (approved)" are the most important elements, suggesting that the goal result is highly affected by students' academic achievement in both semesters. "Scholarship holder" and "tuition fees up to date" are two more noteworthy aspects that highlight the significance of financial assistance and status. Demographic and administrative factors like "Age at enrollment", "Gender", "Debtor", and "Application mode" also show moderate influence. Meanwhile, features like "Mother's qualification", "Marital status", and "Application order" have relatively low importance, suggesting they have minimal effect on the model's predictions.

3.5 Data Normalization with RobutSclaer

Normalization is an essential step for many ML methods since it guarantees that the numerical characteristics are of a consistent size. The applied the Robust Scaler, which scales features using statistics robust to outliers. Let xi be a feature, then the normalised feature x_i' is given by Equation (1):

$$x_i' = \frac{x_i - Q_2(x)}{Q_3(x) - Q_1(x)} \tag{1}$$

where $Q_1(x)$ and $Q_3(x)$ are the first and third quartiles of the feature x, and $Q_2(x)$ is the median.

3.6 Data Balancing using ADASYN

The dataset is resampled using ADASYN, in particular, to fix the class imbalance. ADASYN (Adaptive Synthetic Sampling) uses the density of instances that are hard to learn to create synthetic data for the minority class. It helps improve model performance by reducing class imbalance, allowing the classifier to better learn patterns of the minority class. Figure 7 illustrates the distribution of the balanced dataset.

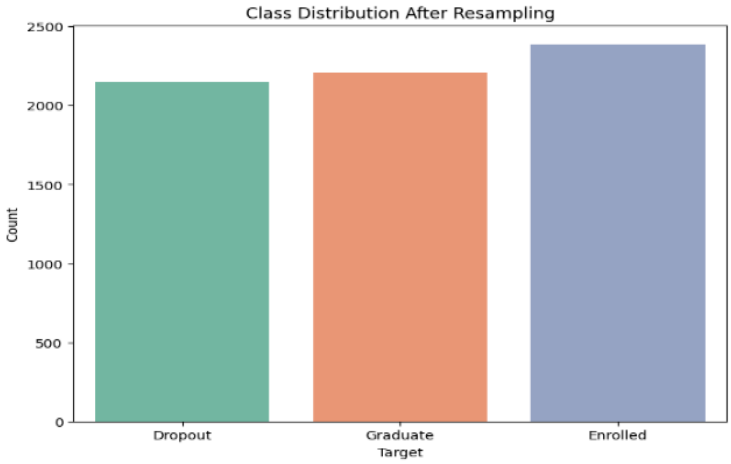


Figure 7: Count Plot for Balanced Class Distribution

After balancing the dataset using the ADASYN resampling approach, the class distribution is shown in Figure 7. After resampling, the distribution of the three target classes Dropout, Graduate, and Enrolled, became nearly equal, each with around 2200 to 2400 instances. When the data is evenly distributed, it makes it easier to train ML models on a more representative sample, which in turn improves the accuracy and fairness of forecasts for all student outcomes while decreasing bias.

3.7 Data Splitting

The resampled dataset is split into training and testing sets using the train_test_split function. In particular, tst_size=0.2 indicates that 80% of the data is used for model training and 20% is set aside for testing. The findings are repeatable since the random_state=42 setting makes sure that the split stays constant during several runs. Preventing overfitting is crucial for assessing the model's performance on new data. Classification using Machine learning models.

3.8 Classification With Machine Learning Models.

This study defines three models for categorization problems, each with its own set of parameters. The propose model include gradient boosting, LightGBM and MLP are discussed in below;

3.8.1 Gradient Boosting (GB) Model

The Gradient Boosted Regression Trees (GBRT) model, commonly known as Gradient Boosted Machine (GBM) or simply GBM, is a major player in the ML industry and a top performer in predictive analytics [15]. A specific kind of additive model, the Boosted Trees Model integrates prediction outcomes from a series of base models. A more formal way to express this category of models is as Equation (2):

$$g(x) = f0(x) + f1(x) + f2(x) + f3(x) + \cdots. \tag{2}$$

where the number of the particular classifiers fi is represented by the final classifier g. Model boosted trees use basic DT as their foundation classifiers. One large-scale technique for boosting prediction performance is model ensembling, which involves integrating numerous models. Gradient Boosting (GB) is generally a type of model in which values are adjusted, including the number of trees (e.g., 100--500), Learning Rate (e.g., 0.01-0.1), maximum tree depth (e.g., 3-8), and minimum samples per leaf/ split (e.g., 2-10). Selecting appropriate values of each of these values by means of selection procedures such as Grid Search or Random Search can provide a balance of model complexity and accuracy and avoid overfitting. Fine tuning the hyperparameters results in a robust and trustworthy prediction model where each successive tree essentially improves on the deficiencies of its predecessors.

3.8.2 LightGBM Model

LightGBM Framework is a stable approach to using gradient boosting by decision trees [23][11]. LightGBM can be used for gradient boosting since it employs tree-based learning techniques. Its decentralized and efficient design enables faster training and increased output. Variable bucketing is carried out via a histogram-based technique called LightGBM, which uses less memory while increasing training speed and accuracy. It can handle big and complicated datasets and operates faster when taught. There is support for learning using parallel processing and GPUs. In supervised learning contexts, the approach may be used to Equation (3) to infer information about a target Y given just X as input.

$$\hat{F} = argmin_f \, E_{y,X} \, L(y, f(x)) \tag{3}$$

To achieve this $\hat{F}$(x) the LightGBM approach minimises the anticipated value of a loss function L (y,f(x)) by using a supervised training set (X). LightGBM hyperparameter tuning is a process of optimizing hyperparameters like the number of trees, learning rate, max depth, and the number of leaves to achieve better accuracy and avoid overfitting. The objective is to discover the best mix of approaches, like grid or Random Search, to fully use LightGBM's efficient and parallel architecture for quicker and more accurate training.

3.8.3 MLP Classifier

A model for neural networks used for classification problems is the MLP Classifier, which stands for Multi-Layer Perceptron Classifier. A network architecture is comprised of linked nodes or neurones that form an input layer, a hidden layer or layers, and an output layer [24]. The model acquires sophisticated patterns within the data by tweaking weights in the course of backpropagation training. The classification type decides upon the activation functions used in the hidden layers, such as ReLU or tanh, and the output layer activation, it may be either softmax or sigmoid. MLP Classifier is effective for non-linear problems and works well when provided with properly scaled and pre-processed data.

$$\hat{Y} = \mathrm{f}(\mathrm{W}_2 \cdot \mathrm{f}(\mathrm{W}_1 \cdot \mathrm{X} + \mathrm{b}_1) + \mathrm{b}_2) \tag{4}$$

where W and b represent the weights and biases learned during training in Equation (4). Various hyperparameters, including hidden layer count, neurone per layer, learning rate, activation function, batch size, maximum iterations, optimisation solver (e.g., SGD or Adam), and optimisation solver selection, are crucial to an MLP Classifier's performance. Grid Search and Random Search are common methods for finding the optimal combination of hyperparameters, supposing the input characteristics have been pre-processed and scaled properly. This guarantees better convergence, better model generalisation, and resilient performance on unknown data.

3.9 Evaluation Metrics

Several indicators were used to evaluate the model's effectiveness in forecasting higher education student dropouts. The numerical expression of classification accuracy was achieved using a confusion matrix. Among the many popular ML methods, the confusion matrix compiles data on the real and anticipated classes produced by a classification algorithm. Two dimensions make up the confusion matrix: the actual classes and the expected classes. There is a projected class state for every row, and an actual class example for every column. Each row in the confusion matrix represents a different outcome: TP, TN, FP, and FN. Key metrics include Acc, which measures overall correctness, Pre and Rec, and f1score.

3.9.1 Accuracy (Acc)

It is a measurement of how closely the expected value resembles the real or hypothetical value. The ratio of accurate predictions to total occurrences is often used to calculate accuracy. Equation (5) displays the accuracy formula:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+Fp+TN+FN}} \tag{5}$$

3.9.2 Precision (Pre)

Precision is a measure of how many true values were accurately predicted out of all the expected values in the real class. Equation (6) displays the precision:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{6}$$

3.9.3 Recall (Rec)

Recall is the rate of correctly classified positive values. Recall determines the proportion of correctly classified true positives. Equation (7) shows the recall formula.

$$Recall = \frac{\text{TP}}{TP+FN} \tag{7}$$

3.9.4 F1-score

The F measure, also called the F1score, is the harmonic mean of Rec and Pre. Equation (8) shows the F1-score formula:

$$F1 - Score = \frac{2(Precision*Recall)}{\text{Precision+Recall}} \tag{8}$$

3.9.5 ROC

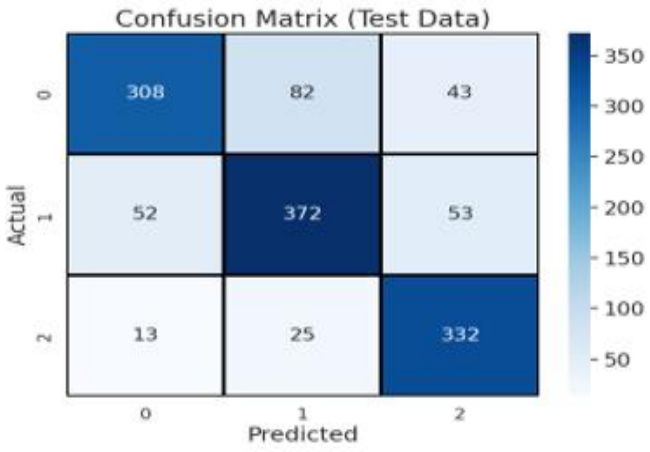
The ROC determines the likelihood that a sample is a member of a certain class. One-vs-one and one-vs-rest are the two methods that the ROC employs for multi-class situations. Lastly, a precision-recall curve graphically displays a classifier's performance and is a popular statistic for unbalanced datasets.

4 Result Analysis and Discussion

This section presents an experiment results of proposed model and system configuration. This article presents PC experiments that use a deep learning framework. The programming language used is Python 3.9 on a Windows 10 64-bit operating system. The CPU on this platform is an Intel(R) Core(TM) i7-8750H running at 2.20 GHz and 2.21 GHz. Graphics Processing Unit: CUDA 12.1, NVIDA GeForce GTX 1050 Ti. Table 2 shows the performance of models across the matrix. The LGBM model demonstrates superior results, with the highest Accuracy of 86.14%, Precision of 86.52%, Recall of 86.14%, and F1 score of 86.15%, indicating its strong capability to correctly identify students at risk of dropping out. In contrast, the GB model exhibits moderate performance, with scores around 79% for all metrics, while the MLP achieves comparable accuracy, precision, f1-score and recall, indicating an imbalance in its classification results. According to the findings, LGBM is the most effective model for predicting whether a student will drop out of the study.

Table 2: Performance of propose models for Student Dropout Prediction

Measures	GB	LGBM	MLP
Accuracy	79.40	86.14	78.52
Precision	79.54	86.52	78.63
Recall	79.40	86.14	78.52
F1-score	79.26	86.15	78.53



Classification Report (Test Set):

	precision	recall	f1-score	support
Dropout	0.83	0.71	0.76	433
Enrolled	0.76	0.78	0.77	477
Graduate	0.80	0.89	0.84	440
accuracy			0.79	1350
macro avg	0.80	0.79	0.79	1350
weighted avg	0.80	0.79	0.79	1350

Figure 8: Classification Report and Confusion Report for GB Model

The GB model's ability to forecast student outcomes, such as enrolment, graduation, and dropout, is shown in Figure 8. There were some mistakes in class predictions, but overall, the model got most of the instances right, according to the confusion matrix. For example, 308 were predicted correctly, while others were wrongly predicted as Enrolled or Graduate. Similarly, the model correctly predicted 372 and 332 Graduates. The classification report provides scores indicating the model's performance, with an overall Acc of 79%. The average Pre, Rec, and F1score are approximately 80%, demonstrating that the model yields good and balanced results across all categories.

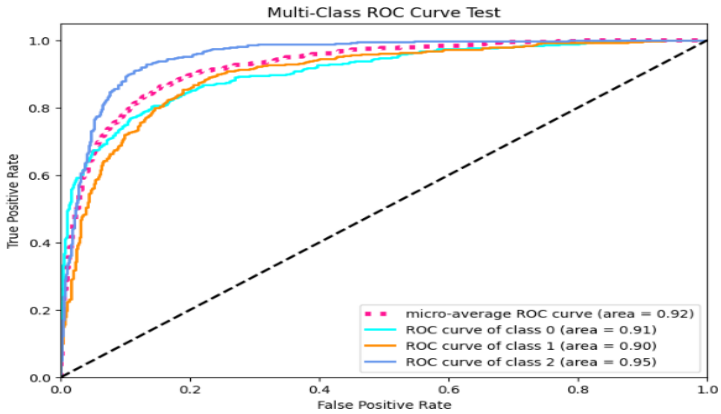


Figure 9: Multi-Class ROC Curve for GB Model

Figure 9 displays a Multi-Class ROC curve, which evaluates the performance of a classification model across three classes (class 0, class 1, and class 2). Each curve represents the trade-off among the TPR and FPR for one class. The AUC values indicate high model performance: 0.91 for class 0 (green), 0.90 for class 1 (blue), and 0.95 for class 2 (orange). The micro-average ROC curve (pink, AUC = 0.92) summarises overall performance across all classes. A curve closer to the top-left corner indicates better classification results.

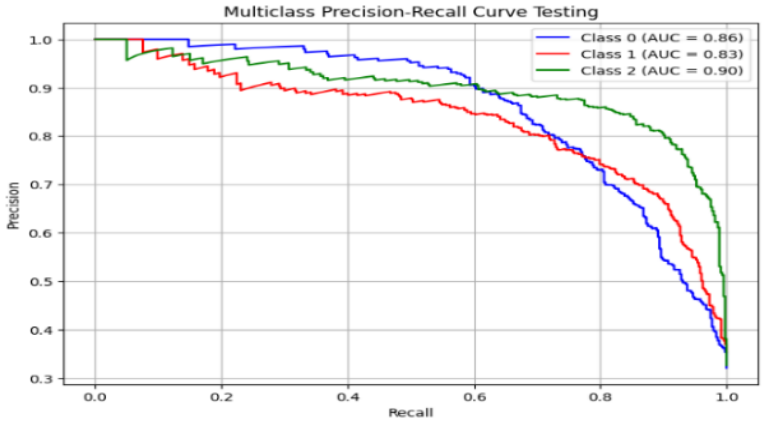
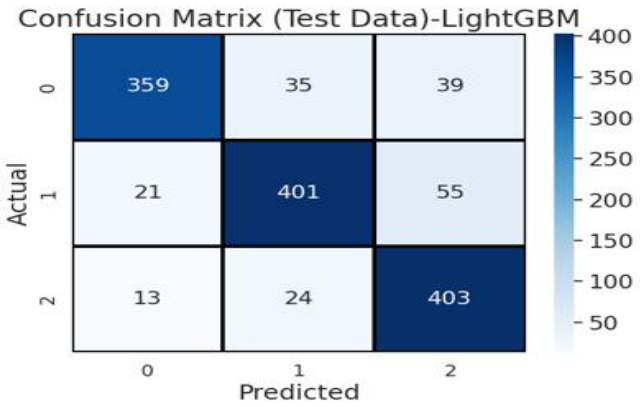


Figure 10: Multiclass Precision-Recall Curve for GB Model

A model's capacity to maintain a balance among recall and precision across three classes is seen in Figure 10. Class 2 (green) performs best with an AUC of 0.90, followed by Class 0 (blue, 0.86) and Class 1 (red, 0.83). Higher AUC values indicate better classification, especially for handling imbalanced data.



LightGBM Classification Report (Test Set):

	precision	recall	f1-score	support
Dropout	0.91	0.83	0.87	433
Enrolled	0.87	0.84	0.86	477
Graduate	0.81	0.92	0.86	440
accuracy			0.86	1350
macro avg	0.87	0.86	0.86	1350
weighted avg	0.87	0.86	0.86	1350

Figure 11: Classification Report and Confusion Report for LGBM

Figure 11 displays a confusion matrix and a classification report for a LightGBM model trained on test data. Each of the three classes (0, 1, and 2) in the Confusion Matrix displays the number of TP, TN, FP, and FN predictions, which together constitute the model's performance. As an example, 359 class 0 occurrences were accurately predicted, however 35 were incorrectly categorized as class 1 and 39 as class 2. The LightGBM Classification Report that comes with it gives the metrics for the 3 classes: "Dropout" (class0), "Enrolled" (class1), and "Graduate" (class2). It displays class-specific measures like support, f1-score, and accuracy, as well as aggregate metrics like overall accuracy, macro average, and weighted average; all of these point to a well-rounded performance with an accuracy level of 0.86%.

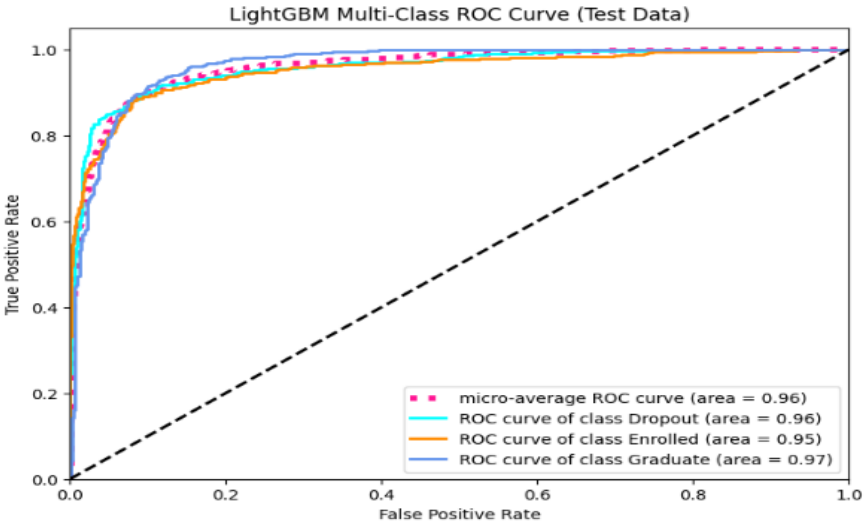


Figure 12: Multi-class ROC Curve for LightGBM Model

Figure 12 demonstrates excellent model performance, with all class AUC scores exceeding 0.95. The Graduate class has the highest AUC (0.97), followed by the Dropout class (0.96) and the Enrolled class (0.95). The micro-average AUC is also 0.96, indicating that the model is highly effective in distinguishing between classes with very low false positive rates.

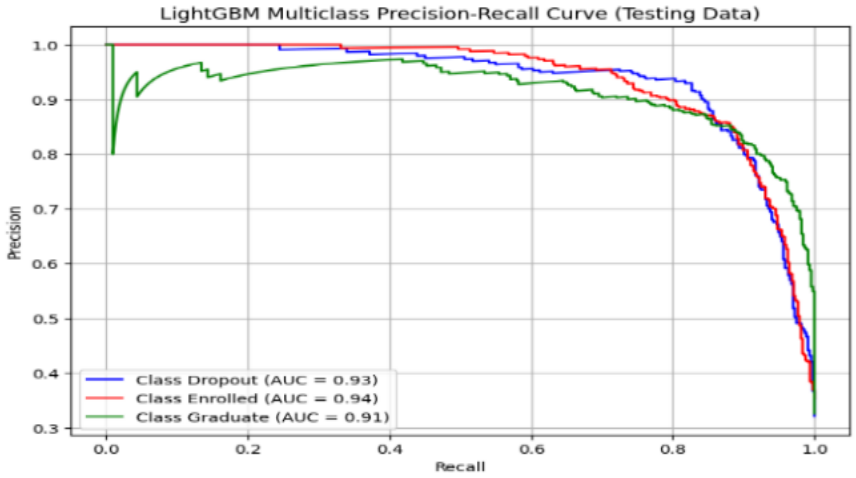
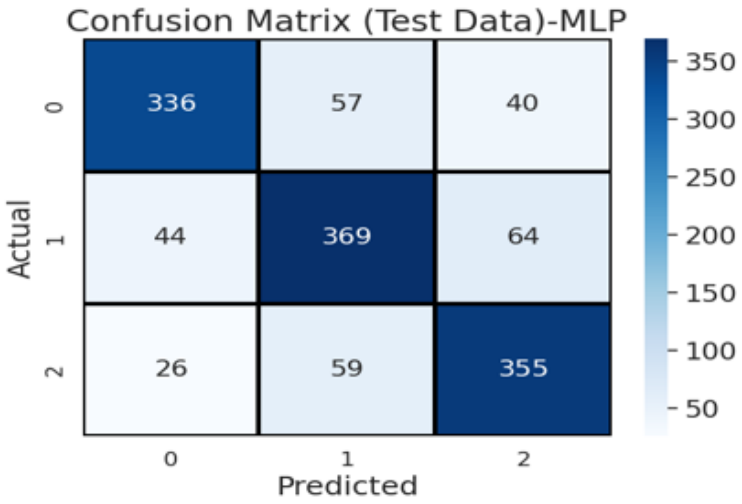


Figure 13: Multiclass Precision-Recall Curve for LightGBM Model

The LightGBM multiclass Precision-Recall curve in Figure 13 shows strong performance across all classes. The AUC scores are high: 0.94 for Enrolled, 0.93 for Dropout, and 0.91 for Graduate, indicating the model maintains a good balance among Pre and Rec. This means it effectively identifies each class with minimal FP and FN.



--- tst Set Classification Report ---

	precision	recall	f1-score	support
Dropout	0.83	0.78	0.80	433
Enrolled	0.76	0.77	0.77	477
Graduate	0.77	0.81	0.79	440
accuracy			0.79	1350
macro avg	0.79	0.79	0.79	1350
weighted avg	0.79	0.79	0.79	1350

Figure 14: Classifier Classification Report and Confusion Report for MLP

Figure 14 presents the confusion matrix and classification report for the Multi-Layer Perceptron (MLP) model tested on student dropout data. The confusion matrix indicates that the MLP accurately predicted 336 dropouts, 369 enrolled students, and 355 graduates, with some misclassifications among these classes. The classification report indicates balanced performance across the three categories, with precision values of 0.83 for dropout, 0.76 for enrolled, and 0.77 for graduate students. The MLP model predicted student status across enrolment, dropout, and graduation outcomes with a dependable and balanced approach, as shown by its 79% accuracy on the test set and 0.79 for the weighted averages of Pre, Rec, and F1score.

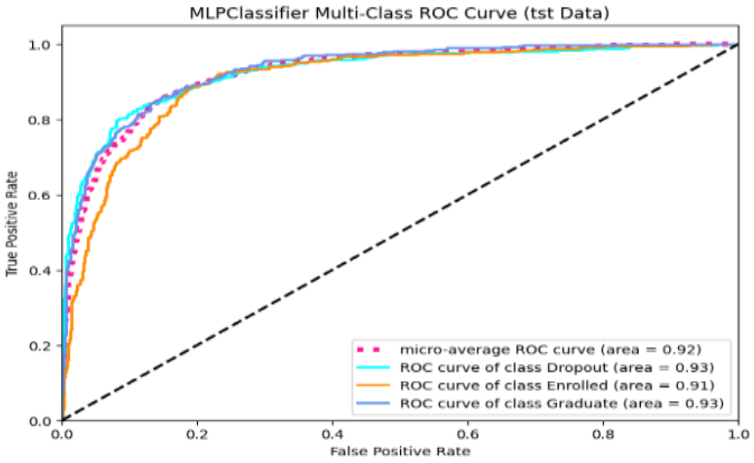


Figure 15: Multi-class ROC curve for MLP Model

Figure 15 shows how well the MLP Classifier performed on test data with many classes. Each curve represents one class: Dropout, Enrolled, and Graduate, with respective AUC values of 0.92, 0.93, 0.91, and 0.93, indicating high classification performance across all categories. Additionally, the micro-average ROC curve (which displays the overall performance of the model) reaches an AUC of 0.92. A low FPR and a high TPR, as shown by the curves being near to the top-left corner, suggest that the model efficiently differentiates between classes.

4.1 Comparative Analysis and Discussion

The comparative analysis of different models for students dropout prediction are present in Table 3. The level of accuracy is lower in Traditional models such as KNN, DT, and NB, in this comparison, which is indicative of their inability to identify the complex patterns present in the data. Conversely, more recent models like Gradient Boosting, LightGBM, and MLP indicate

better accuracy implying a higher level of generalization and predictive outcomes. The rise in accuracy using the proposed models shows their effectiveness and certainty in recognizing at-risk students of dropping out.

Table 3: Comparison betcbween the base and proposed model performance for students dropout prediction

Measures	Accuracy	Precision	Recall	F1-Score
DT[25]	70.1	70	70	70
KNN[26]	66	64	63	65
NB[27]	0.77	0.72	0.93	0.82
GB	79.40	79.54	79.40	79.26
LGBM	86.14	86.52	86.14	86.15
MLP	78.52	78.63	78.52	78.53

The advantages of the proposed model, particularly the LightGBM, over the traditional models employed including DT, KNN, and Naive Bayes, are several. It scores higher in Acc, Pre, Rec and F1score therefore making it a more precise measure of student dropout. LightGBM is easy to work with large and complex information, and achieves faster preparation, which is particularly useful in practice. It performs well in all of its classes, dropout, enrolled and graduate categories with consistent and balanced results. In contrast to the more traditional models that have problems in learning when the data is unbalanced, the used oversampling algorithm, ADASYN, simply jumps the data into a much-balanced learning space. The scaling is also strong with outliers. The improvements will assist you in knowing the students at risk at an earlier time so that you may intervene. LightGBM can be likened to Gradient Boosting and MLP since it is quicker, bigger and generally more effective. This puts it in a good place as an educational institution seeking to enhance student outcomes and reduce the extent of dropout through the use of data to make decisions.

5 Conclusion and Future Work

Student dropout is a very serious problem that is being witnessed in Higher Education Institutions (HEIs) and which impacts academic planning and student success. This work provides an efficient means of predicting student dropout with the help of intricate ML models. In terms of the accuracy, GB, LGBM and MLP LGBM had the highest accuracy score of 86.14 per cent respectively, then GB and MLP, 79.40 per cent and, 78.52 per cent respectively. With these findings, one can observe how LightGBM can handle more complex data and make appropriate predictions. All in all, the suggested plan will enhance the early detection of at-risk people, thus enabling the educational facility to effectively mitigate the dropout rates through appropriate early intervention. This shows that the schools can take advantage of the proposed method to sift the at-risk students and pull them at an earlier stage and, therefore, enhance success and retention rates. Nonetheless, the study has its own limitations such as the fact that it used one source of information and this may restrict the applicability of the model to other educational environments. It also lacks behavioural, emotional, and psychological data that can influence student performance. It should be scalable to other institutions and add more data in the future to include things like attendance, participation and real-time academic performance that may result in greater model generalizability, accuracy and applicability in real-world academic systems.

References

[1] Adarsh Reddy Bilipelli, “Application of AI and Data Analysis for Classification of Student Success in Large-Scale Educational Dataset,” *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 428–441, Nov. 2024, doi: 10.48175/IJARSCT-22564.

[2] G. M. and H. Kali, “Exploring Big Data Role in Modern Business Strategies: A Survey with Techniques and Tools,” *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 3, pp. 1–11, 2023.

[3] M. Vaarma and H. Li, “Predicting student dropouts with machine learning: An empirical study in Finnish higher education,” *Technol. Soc.*, vol. 76, Mar. 2024, doi: 10.1016/j.techsoc.2024.102474.

[4] V. S. Thokala, “Integrating Machine Learning into Web Applications for Personalized Content Delivery using Python,” *Int. J. Curr. Eng. Technol.*, vol. 11, no. 06, 2021, doi: https://doi.org/10.14741/ijcet/v.11.6.9.

[5] Suhag Pandya, “Innovative blockchain solutions for enhanced security and verifiability of academic credentials,” *Int. J. Sci. Res. Arch.*, vol. 6, no. 1, pp. 347–357, Jun. 2022, doi: 10.30574/ijsra.2022.6.1.0225.

[6] J. Martínez and D. Castillo, “Prediction of student dropout using Artificial Intelligence algorithms,” *Procedia Comput. Sci.*, vol. 251, pp. 764–770, 2024, doi: 10.1016/j.procs.2024.11.182.

[7] N. Malali, “Exploring Artificial Intelligence Models for Early Warning Systems with Systemic Risk Analysis in Finance,” in *2025 International Conference on Advanced Computing Technologies (ICoACT)*, 2025, pp. 1–6. doi: 10.1109/ICoACT63339.2025.11005357.

[8] S. Kim, E. Choi, Y.-K. Jun, and S. Lee, “Student Dropout Prediction for University with High Precision and Recall,” *Appl. Sci.*, vol. 13, no. 10, p. 6275, May 2023, doi: 10.3390/app13106275.

[9] N. V. M. B. S. Singamsetty, “Enhancing Student Engagement and Outcomes through an Innovative Pedagogy for Teaching Big Data Analytics in Undergraduate Level,” *Int. J. Comput. Math. IDEAS*, vol. 16, no. 1, pp. 2000–2011, 2024.

[10] M. A. Hassan, A. H. Muse, and S. Nadarajah, “Predicting Student Dropout Rates Using Supervised Machine Learning: Insights from the 2022 National Education Accessibility Survey in Somaliland,” *Appl. Sci.*, vol. 14, no. 17, Aug. 2024, doi: 10.3390/app14177593.

[11] N. Prajapati, “The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications,” *ESP J. Eng. Technol. Adv.*, vol. 5, no. 2, pp. 16–22, 2025, doi: 10.56472/25832646/JETA-V5I2P103.

[12] C. H. Cho, Y. W. Yu, and H. G. Kim, “A Study on Dropout Prediction for University Students Using Machine Learning,” *Appl. Sci.*, vol. 13, no. 21, 2023, doi: 10.3390/app132112004.

[13] V. Varma, “Secure Cloud Computing with Machine Learning and Data Analytics for Business Optimization,” *ESP J. Eng. Technol. Adv.*, vol. 4, no. 3, 2024, doi: 10.56472/25832646/JETA-V4I3P119.

[14] H. S. Won, M. J. Kim, D. Kim, H. S. Kim, and K. M. Kim, “University Student Dropout Prediction Using Pretrained Language Models,” *Appl. Sci.*,

vol. 13, no. 12, 2023, doi: 10.3390/app13127073.

[15] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 04, pp. 3557–3564, 2025.

[16] M. Pérez, D. Navarrete, M. Baldeon-Calisto, Y. Guerrero, and A. Sarmiento, "Unlocking Student Success: Applying Machine Learning for Predicting Student Dropout in Higher Education," in *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, 2025, pp. 1–6. doi: 10.1109/ISDFS65363.2025.11012013.

[17] M. Wang, "Academic Warning for College Students for Predicting Student Dropout Rate using Dipper Throated Optimization Algorithm," in *2024 First International Conference on Software, Systems and Information Technology (SSITCON)*, 2024, pp. 1–5. doi: 10.1109/SSITCON62437.2024.10797187.

[18] E. S. Aisyah, D. Manongga, A. Iriani, and S. Santoso, "Comparative Analysis of Machine Learning Algorithms for Predicting Undergraduate Academic Performance," in *2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT)*, 2024, pp. 1–6. doi: 10.1109/ICCIT62134.2024.10701234.

[19] S. Deb, M. S. R. Sammy, A. N. Tusher, M. R. S. Sakib, M. F. Hasan, and A. I. Aunik, "Predicting Student Dropout: A Machine Learning Approach," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–7. doi: 10.1109/ICCCNT61001.2024.10726161.

[20] T. Akter, U. Ayman, N. R. Chakraborty, D. A. Islam, A. Mazumder, and M. H. I. Bijoy, "Dropout Prediction of University Students in Bangladesh using Machine Learning," in *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–7. doi: 10.1109/COMPAS60761.2024.10797033.

[21] M. A. Dewi, F. I. Kurniadi, D. F. Murad, S. G. Rabiha, and A. Romli, "Machine Learning Algorithms for Early Predicting Dropout Student Online Learning," in *2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED)*, IEEE, Nov. 2023, pp. 1–4. doi: 10.1109/ICCED60214.2023.10425359.

[22] N. Malali, "AI-Powered Data Preprocessing and Transformation Platform for Autonomous Data Cleaning, Advanced Fea," 202521035175, 2025

[23] Y. Zhou, W. Wang, K. Wang, and J. Song, "Application of LightGBM Algorithm in the Initial Design of a Library in the Cold Area of China Based on Comprehensive Performance," *Buildings*, 2022, doi: 10.3390/buildings12091309.

[24] S. Nokhwal, P. Chilakalapudi, P. Donekal, S. Nokhwal, S. Pahune, and A. Chaudhary, "Accelerating Neural Network Training: A Brief Review," *ACM Int. Conf. Proceeding Ser.*, pp. 31–35, 2024, doi: 10.1145/3665065.3665071.

[25] S. A. Sulak and N. Koklu, "Predicting Student Dropout Using Machine Learning Algorithms," *Intell. Methods Eng. Sci.*, vol. 3, no. 3, Jan. 2024, doi: 10.58190/imiens.2024.103.

[26] S. Hoca and N. Dimililer, "A Machine Learning Framework for Student Retention Policy Development: A Case Study," *Appl. Sci.*, vol. 15, no. 6, pp. 1–30, 2025, doi: 10.3390/app15062989.

[27] J. Kabathova and M. Drlik, "Towards predicting student's dropout in university courses using different machine learning techniques," *Appl. Sci.*, vol. 11, no. 7, 2021, doi: 10.3390/app11073130.