IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

Application of AI and Data Analysis for Classification of Student Success in Large-Scale Educational Dataset

Adarsh Reddy Bilipelli
Independent Researcher

Abstract: *A major obstacle in higher education is the early detection of students who are at risk of dropping out, which has major consequences for retention rates and students' general performance in the classroom. New possibilities for data-driven interventions are opening up as large-scale educational datasets like OULAD become more widely available. Traditional prediction methods, however, are frequently rendered ineffective by educational data due to its intrinsic complexity, heterogeneity, and imbalance. This research presents a scalable and powerful machine learning framework for early student success prediction that makes use of AI and sophisticated data analysis methods. After cleaning, imputation of missing values, normalization, one-hot encoding, and feature engineering using recursive feature elimination (RFE), the OULAD dataset was ready for preprocessing. After determining which variables were most predictive, split the dataset in half to use for training and testing. A total of three models, namely LGBM, ANN, and FCM-RF (Fuzzy C-Means with Random Forest), were trained and assessed using critical metrics. Outperforming ANN (accuracy: 78% and FCM-RF accuracy: 88.33%), experimental results show that LGBM obtained higher predictive performance with a precision of 94.40%, recall of 93.21%, and F1-score of 93.86%. The findings highlight the effectiveness of ensemble-based methods, particularly LGBM, in accurately identifying at-risk students, enabling timely and targeted interventions to enhance institutional outcomes.*

Keywords: Early prediction, machine learning, educational data mining, OULAD Dataset, LGBM Model, academic performance, student retention

I. INTRODUCTION

Digital learning platforms and advancements in educational technology over the past few years have allowed for the collection of enormous datasets that record a variety of student academic involvement, performance, and behaviour [1]. These data offer unprecedented opportunities to understand and enhance learning processes by identifying patterns that were previously difficult to detect[2]. However, the increasing volume, heterogeneity, and complexity of educational data pose in extracting meaningful and actionable insights[3]. Timely interventions can immediately increase retention rates, equality, and overall academic success; but, identifying at-risk students early on continues to be a key concern in higher education.

The educational institutions all over the world are starting to implement open-ended data-driven decision making in their continual improvement process [4][5]. A comparative report on the education system in some European states, such as the United Kingdom, Germany, Poland, Lithuania, and the Netherlands, has shown that data analytics is emerging to support and help in advancing instructional improvements, institutional accountability, and student progress [6]. Data-driven decision-making has been defined as the process of gathering and analyzing the digital footprints of learners in an attempt to inform pedagogical and policy decision-making processes [7][8]. The lack of collaboration among stakeholders, inadequate professional training in data analytics, and difficulty in obtaining high-quality data are some of the ongoing issues that persist despite these advancements [9].

The methods of data analysis and AI emerged as promising avenues for discovering answers to [10]. Modern ML temporal patterns (RTPs) trained using a set of foundational models, including a cutting-edge deep learning model,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

three traditional ML models (k-Nearest Neighbors, support vector machines, and logistic regression), and others Traditional ML approaches that depend on statistical methods are outdone by LSTM in this regard [11] shine a light on complicated and non-linear interactions [12]. By building predictive models capable of classifying students based on their likelihood of success, educators and administrators can implement proactive and personalized interventions. Such data-driven approaches not only enhance individual student outcomes but also improve institutional efficiency and accountability a robust, scalable, and practical framework for early prediction of student success by leveraging AI and advanced data analysis techniques[13][14]. Using the anonymized Open University Learning Analytics Dataset (OULAD), this study thoroughly prepares the data, engineers features using recursive feature elimination (RFE), and compares the top machine learning models available. The suggested method aims to find the most important factors that affect how well students do in school, compare how well different models can predict outcomes, and give useful information to help with timely and targeted intervention.

A. Motivation with Contribution

The significant problem of early identification of students at risk of underperformance or withdrawal persists in educational institutions. Higher rates of student retention, better academic performance, and overall institutional effectiveness can be achieved by timely interventions grounded in accurate predictions. However, the complicated interplay of factors impacting student progress is typically missed by traditional evaluation systems, which are reactive and subjective. The availability of rich educational datasets, such as the OULAD, alongside advances in AI and data analysis techniques, presents an opportunity to build proactive, data-driven early warning systems. This research is motivated by the need for a robust and scalable predictive framework that can process high-dimensional educational data, identify key determinants of success, and deliver reliable predictions to support evidence-based decision-making in academic settings. The main contributions of early prediction of student success in education environment are as follows:

- A comprehensive methodological pipeline for early prediction of student success is proposed, integrating rigorous data pre-processing, normalization, and feature engineering to ensure data quality and relevance.
- The process of enhancing academic outcome models by the identification of the most significant factors through the application of Recursive Feature Elimination (RFE).
- Implementation and evaluation of three cutting-edge ML techniques Applying ANN, LGBM, and FCM-RF—a hybrid of Fuzzy C-Means and Random Forest—to the OULAD dataset.
- Experimental assessment of the model's performance using a classification matrix and the measures of recall, accuracy, precision, and F1-score.
- Demonstration of a robust, scalable, and data-driven approach that provides actionable insights for educators to implement timely and targeted interventions for at-risk students.

B. Significance and Novelity

The current research is important because it introduces a strong, weather-resistant, and implementable ML architecture that is explicitly built to anticipate early victor of student performance (a critical issue in the contemporary learning landscape). Its originality consists in creating a full pipeline which combines a rigorous preprocessing of the data, a feature engineering via Recursive Feature Elimination (RFE), and an extensive assessment of various ML models on two real-life educational texts (OULAD). In contrast to a number of previous works, which are limited to study one specific algorithm or ignore the issues of real implementation, the present work gives a comparative analysis of more advanced models such as Light Gradient Boosting Machine (LGBM), ANN, and Fuzzy C-Means with Random Forest (FCM-RF) and reveals the LGBM model as the best predictor. Visual exploratory data analysis that is incorporated helps to improve the comprehension of the most essential factors that affect student outcomes, and thus, allows educators and administrations to take real-life decisions based on its interpretation. Furthermore, the integration of ensemble techniques, neural networks, and clustering-based hybrid methods offers unique perspectives on algorithm

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 6, November 2024 Impact Factor: 7.57

effectiveness, addressing the critical need for adaptable and interpretable early warning systems in dynamic educational settings.

C. Structure of paper

The paper's outline is as follows: Section II presents a literature analysis on the topic of early student success prediction, and Section III lays out the suggested technique in great depth, detailing the system design phases such as data preprocessing, feature engineering, and model training. The experimental results of the suggested models are detailed and analysed in Section IV, along with a comparison of their performance. Section V concludes the work and discusses potential future directions for research.

II. LITERATURE REVIEW

This section delves into the literature review's discussion on how to use data-driven techniques and AI to make early predictions about students' achievement. Those covered later on are summarised in Table I:

Pande (2023) Possessing a high level of education opens several doors for one's future. An individual's and a society's level of achievement are both enhanced by good education. Using criteria to evaluate students' progress and accomplishments is a problem in India's educational system. Assessing the growth and development of students in the classroom calls for the right set of tools. Because there are so many factors that influence a student's performance—their age, their parents' occupation and education level, their health, etc.—analyse and present these variables to you. Students who are having difficulty can be identified early on using the visualization technique, allowing them to concentrate on improving. Multiple ML methods, such as KNN, LR, and SVM, are employed to forecast students' performance. On the given dataset, the SVM model with a linear kernel achieved the highest accuracy rate of 84.37%, surpassing all others. [15].

Ridwan et al. (2023) ML is an AI technique that can predict students' academic performance by tracking the variables that have an impact on their grades over time. Develop a system that can precisely foretell the enrolment levels of all educational institutions. To enhance student state prediction using the ANN, RF, Tree, and NB models as opposed to the more traditional ML methods. The KCV method was used to optimise the dataset processing efficiency. This method is ideal for dealing with imbalanced datasets like the one that was initially used to fix it. With the highest accuracy score (92%), this approach significantly outperformed the others [16].

Yağcı (2022) Educational data mining has several practical applications, one of which is the prediction of students' academic achievement using their midterm exam scores. This method forecasts how well students do on their final exams based on their performance on the midterms using machine learning algorithms. I anticipated students' performance on the final test by utilising various machine learning methods, including random forests, nearest neighbour, support vector machines, logistic regression, Naïve Bayes, and k-nearest neighbour. One thousand eight hundred fifty-four students enrolled in Turkish Language-I at public universities had their grade reports included in the collection. Based on the results, the suggested model achieved a classification accuracy of 70–75%. There were just three factors that were considered for setting expectations: midterm performance, knowledge level, and participation in decision-making [4].

Alalawi, Chiong and Athauda (2021) Educators may offer kids the right solutions at the right time with early identification in this environment, increasing the likelihood that they succeed. An early warning system that identifies struggling students and communicates their progress to teachers and students alike can be built using ML algorithms. Investigate the efficacy of several ML algorithms in detecting beginning-of-semester underachievers in a particular undergraduate course, leading to a multiclass classification issue. With 84% accuracy, it can isolate borderline students and those with poor performance on the first exam of the semester. a tool for predicting students' success in the classroom that enables teachers to develop ML models and spot low-achieving pupils before the beginning of the school year [17].

Drousiotis, Shi and Maskell (2021) a dependable model for making predictions using OULAD-derived de-identified demographic and behavioural data. The results showed that the proposed model was better at distinguishing between students who withdrew, failed, passed, or received a distinction than the two groups of Non-completers and Completers.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 6, November 2024 Impact Factor: 7.57

The algorithm achieved an accuracy of 80% or higher when it came to predicting whether students withdraw, fail, or obtain a distinction. Could use this information to create tailored therapies with higher precision. Notably, model forecasts the end outcome right from the start of a course, in contrast to other research that has used a similar approach [18].

Gull et al. (2020) Academics can greatly enhance their learning strategies and focus on a variety of instructional approaches to guarantee a fruitful learning experience when they obtain early signals about students' progress. By utilizing ML, educators may anticipate which students may face difficulties in specific areas of their coursework and adjust individual lesson plans appropriately. LR, linear discriminant analysis, KNN, regression and classification trees, Gaussian Naive Bayes, and support vector machines were utilized for the model's training. For the next term, it was utilized to predict the marks of incoming freshmen taking the identical class. Clinical trials indicated that linear discrimination analysis yielded the best results, with an anticipated accuracy of 90.74% [19].

Recent studies have looked into how AI and data-driven methods can be used to guess early on how well someone does in school. The results show that these methods are much more accurate and can make predictions much earlier. A lot of people have used ensemble learning methods like RF, XGBoost, and hybrid stacking models to fix unequal class sizes and make sense of complicated connections in student data. In addition, clustering-based methods have been proposed to detect outliers and atypical learning behaviors within student populations. Several studies have demonstrated the strong predictive power of ANN and LR models, whereas others have introduced optimization-based algorithms combined with ML to further improve early warning capabilities. Despite encouraging results, persistent challenges include handling heterogeneous educational contexts, enhancing model transparency and interpretability, addressing highly skewed datasets, and ensuring scalability for real-world implementation. Moreover, integrating temporal patterns of student activity and adopting explainable AI techniques remain promising directions for building more trustworthy and actionable prediction systems

Table 1: Comparative Analysis Of Recent Studies On Early Prediction Of Student Success Using Machine Learning Techniques.

Author	Dataset	Methodology	Key Findings	Advantages	Limitations	Future Work
Pande (2023)	Custom dataset with student demographics and performance factors	Visualization of factors (age, parents' education/occupation, health, etc.); ML (KNN, LR, SVM)	SVM with linear kernel achieved 84.37% accuracy in predicting student performance	Early identification of weak students through visualization	Limited to one dataset and factors; kernel choice not justified	Explore more diverse datasets and kernel options for SVM
Ridwan et al. (2023)	University student dataset (imbalanced, processed via K-fold Cross Validation)	ML models: ANN, RF, Decision Tree, Naive Bayes; K-fold Cross Validation for imbalance handling	RF achieved the highest accuracy (92%)	Robust handling of imbalanced data; comparison of multiple models	Focused only on university-level data; no deep analysis of feature impact	Extend to other educational levels and study feature importance
Yağcı (2022)	Results from the first two exams taken by 1,854 students in the Turkish language	ML models like RF, KNN, SVM, LR, and Naïve Bayes are used to guess final grades based on	Predicted final test marks with a 70%-75% success rate.	Minimal features (only midterm and demographics) used for prediction	Moderate accuracy; no intervention suggestions	Incorporate more behavioral data to improve predictions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

	course	student data and test scores				
Alalawi et al. (2021)	Undergraduate course data (first assessment scores)	ML models for multiclass classification to detect under-performing & borderline students early	Early detection of under-performing students, borderline at 84%	Enables early intervention in semester	Focus on one course; no detail on feature selection	Test on multiple courses & expand feature set for generalization
Drousiotis et al. (2021)	Open University Learning Analytics Dataset (OULAD)	Behavioral + demographic features; ML for multiclass classification (Withdrawn, Fail, Pass, Distinction)	Prediction accuracy ≥80% for all classes, at very beginning of course	Fine-grained prediction; allows targeted interventions early on	Limited to OULAD; impact of interventions not assessed	Study real impact of interventions & tests on different datasets
Gull et al. (2020)	Historical student grades in undergraduate course	ML models: Logistic Regression, LDA, KNN, CART, Gaussian NB, SVM	LDA achieved highest accuracy (90.74%) for predicting next term grades	Identifies weaknesses in learning early; compares several models	Focused on single course; no student-specific interventions	Apply model across disciplines & integrate with learning analytics platforms

III. METHODOLOGY

This study utilizes a thorough methodology to predict students' achievement utilizing AI and data analysis tools. The OULAD dataset undergoes rigorous data preprocessing including data cleaning, missing value imputation, and one-hot encoding to ensure data quality Figure 1. After preprocessing, feature scales are normalized using data normalization. Then, to find the best predictors of student performance, feature engineering with Recursive Feature Elimination (RFE) is employed. After that, the processed dataset is subjected to feature selection techniques for the purpose of optimizing the model input variables. Next, training and testing data sets are made from the data. To help confirm the model, this useful. Look at LGBM, ANN, and FCM-RF (Fuzzy C-Means with Random Forest) to see how well they work and how they compare to each other. Most of the time, accuracy, precision, recall, and F1-score are used to measure how well a model does in a full classification grid. This systematic approach ensures robust model development and reliable prediction of student academic outcomes through comparative analysis of multiple AI techniques.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

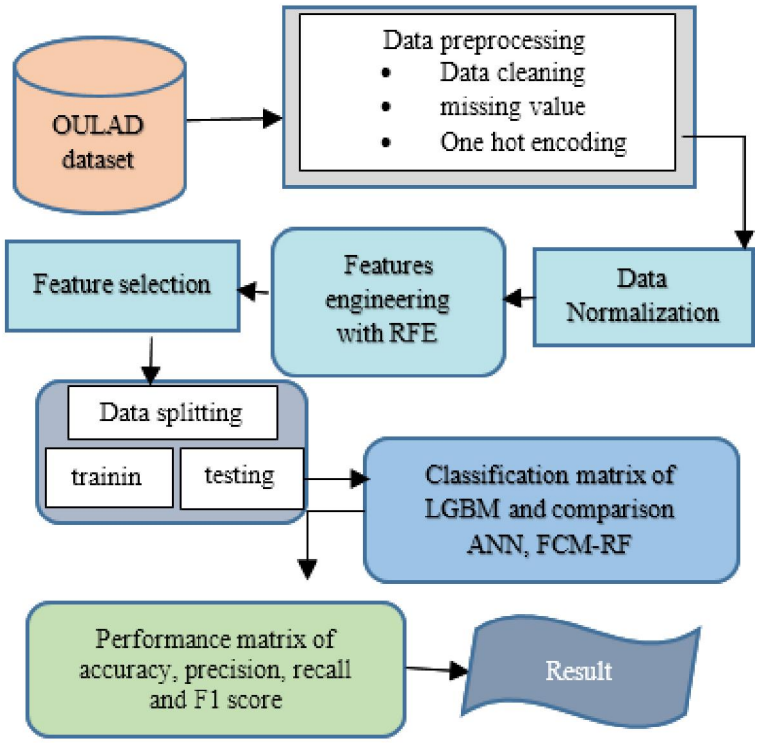

Figure 1: Flowchart for Early Prediction of Student Success using Machine Learning

A. Dataset Description

This study made use of the anonymous Open University Learning Analytics Dataset (OULAD). It draws on seven selected courses offered by the UK's Open University, which caters mostly to undergraduates who take their courses off-campus— There are 32,593 student records in the databases, covering 22 classes offered between February and October. Out of the total number of pupils, 17,208 were marked as "fail" and 15,385 as "pass." Drawing on OULAD's records Students' demographics, registration details, test scores, and comprehensive records of their interactions with the VLE (including when and how they accessed resources and completed tasks) are all part of the collection. Based on their academic achievement, learner outcomes are divided into four classes: distinction, pass, fail, and withdrew. Student Info, student Assessment, assessment, student VLE, student Registration, VLE, and courses are the seven CSV files that comprise the dataset. some of the visualizations are given below:

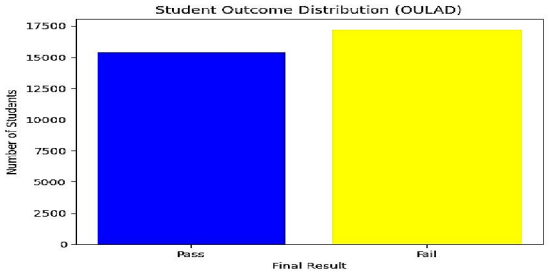

Figure 2: Student Outcome Distribution in OULAD Dataset

Its distribution of student outcomes in the OULAD dataset, comparing the number of students who passed and failed. Approximately 15,385 students passed while 17,208 students failed, indicating a higher failure rate in Figure 2. This

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

emphasises the need for early prediction models in identifying pupils who may struggle academically and in enhancing their academic achievement rates.

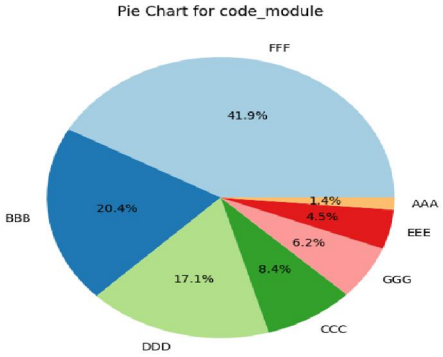

Figure 3: Pie Chart for Code Module Distribution in OULAD Dataset

This pie chart illustrates the distribution of students across different course modules in the OULAD dataset in Figure 3. The largest segment represents module FFF with 41.9% of students, followed by BBB (20.4%), DDD (17.1%), CCC (8.4%), GGG (6.2%), EEE (4.5%), and AAA (1.4%). This distribution provides crucial baseline information for early prediction modeling of student success across various academic disciplines.

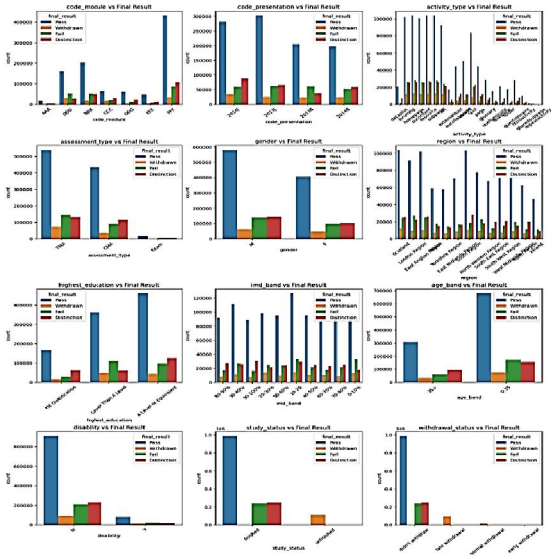

Figure 4: Distribution Analysis of Student Demographic and Academic Feature in OULAD Dataset

Its distribution of various student characteristics and academic performance indicators in the OULAD Dataset in Figure 4. The bar charts systematically examine demographic variables including gender, age bands, region, highest education level, disability status, and socioeconomic factors, alongside academic metrics such as assessment types, study behaviors, and final results. Essential insights into the multifaceted elements influencing early prediction of student achievement in online learning environments are provided by each subplot, which uses colour-coded categories to distinguish between distinct classification outcomes.

B. Data Preprocessing

Data preprocessing involves removing irrelevant or noisy records, filling in missing values, and applying one-hot encoding to categorical attributes in order to make better predictions. Normalize the feature scales to ensure consistency.After finding relevant predictors using Recursive Feature Elimination (RFE), feature engineering entails

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

selecting features. Finally, in order to guarantee that students can be reliably predicted for early success, the dataset is divided into training and testing sets. The procedures are outlined in brief below:

- **Data Cleaning:** No structure and lots of noise characterize most real-world data sets [20]. By removing noise and filling in missing numbers, the data cleaning process attempts to address data anomalies.
- **Missing value:** Academic real-world datasets often face the problem of missing data, which might be caused by typos. Identified missing data in the dataset and removed whole entries to fix the problem. Eliminating the features enhances the model's anonymity and guarantees that training is conducted using only complete and reliable data.
- **One hot encoding:** One-hot encoding was used to make sure that the ML model saw each category as separate from the others. For each separate group, this way of encoding categorical data makes a new binary feature (0 or 1). The gender trait has two groups: male and female.

C. Data Normalization

Data normalization is a pre-processing method that uniformly adjusts numerical features in a dataset without influencing their interrelationships or variations in value. Data is "normalized" when it is scaled or averaged to a common range or format. used Min-Max normalization approaches to scale the features to a range between 0 and 1, as described in [21]. This helped to prevent numerical instability and reduce bias towards variables with larger values. Equation (1) guided the data normalization process.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$

where X_{max} is the highest possible value, X_{min} is the lowest possible value, and X_{norm}This is the value that has been normalised.

Feature Engineering for RFE Technique

Engineering features for early success prediction using the RFE method on the OULAD dataset. Method for estimating the "final_result" target class using RFE [22] to hone the feature selection process even further. Using model performance as a metric, RFE iteratively removes the least significant features, leaving only the most relevant predictors. This method is based on wrappers.

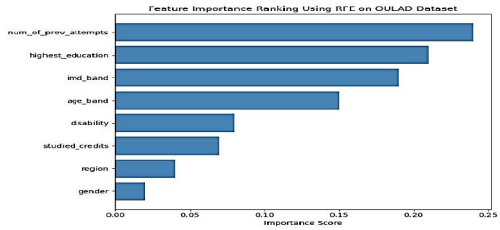

Figure 5: Feature Importance Ranking Using RFE On OULAD Dataset

As shown in Figure 5 the most important features identified by RFE include num_of_prev_attempts, highest education, imd_band, and age band, indicating their strong influence on student outcomes. Less impactful features, such as region and gender, exhibit lower importance scores and contribute minimally to the prediction task. This ranking of features not only improves model efficiency by reducing dimensionality but also enhances interpretability.

E. Feature Selection

Feature selection is a crucial step because it involves picking out the most pertinent features from the provided dataset. Beginning with all features, this method progressively removes those that do not significantly affect the model's performance [23]. This approach streamlines the model while simultaneously enhancing performance and reducing the risk of overfitting.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

F. Data Splitting

There is an 80:20 split between the dataset's training and testing sets. The training set was exclusively utilized for training and hyperparameter tweaking. The purpose of the test set is to evaluate the model's performance and generalizability.

G. Classification of LGBM Model in Student Success in Early Prediction

LightGBM provides better computing efficiency and accuracy through its use of decision trees and quick gradient boosting [24]. It reduces feature dimension and computes information gain by excluding instances with small gradients and focusing on those with large ones using exclusive functional grouping and histogram-based methods [25]. Using a tree-leaf-wise method with a maximum depth limit also improves LightGBM performance. Because of these and other improvements, LightGBM outperforms competing algorithms in terms of accuracy and computing efficiency.

LightGBM introduces a fresh perspective to the gradient boosting community with its innovative tree growth strategy that grows leaves at a faster rate than the current level-wise methods.The following aims to minimize loss in the following Equation (2), and this plan is the best way to make that happen:

$$Minimize\ E_{x,y} = \left[L\left(y, F(x)\right)\right] \quad (2)$$

For each iteration t, the model computes the gradients as follows in Equation (3) below:

$$g_i^{(t)} = \frac{\delta L(y_i, f_{t-1}(x_i))}{\delta F_{t-1}(xi)} \quad (3)$$

LightGBM uses these gradients to formulate a leaf-wise expansion strategy and an expansion approach builds upon the delta-loss leaf. This is the method in which regularization term λ is included in Equation (4) to avoid overfitting governed by the following calculation:

$$Gain = \frac{1}{|I_L|+\lambda}\left(\sum_{i\in I_L} g_i^{(t)}\right)^2 + \frac{1}{|I_R|+\lambda}\left(\sum_{i\in I_R} g_i^{(t)}\right)^2 - \frac{1}{|I|+\lambda}\left(\sum_{i\in I} g_i^{(t)}\right)^2 (4)$$

This λ is a kind of complexity penalty, so that the model does not become overoptimized to the training data specifics, a requirement that is instrumental to generalization and prediction performance on future data.

H. Performance Matrix

The developed prediction model of student success can be tested against performance measures of standard classifiers, such as accuracy, precision, recall, and the F1-score. These measures show information about general accuracy of predictions, potential to predict successful students and precision/recall tradeoff. These measures help assess the effectiveness and reliability of the predictive model in identifying student performance early Additionally, the confusion matrix is employed to summarize prediction outcomes below:

- **True positive (TP):** reflects the number of pupils who were accurately projected as becoming successful.
- **True negative (TN):** represents students' accurate predictions of failure.
- **False positive (FP):** involves pupils who are wrongly assumed to have achieved.
- **False negative (FN):** shows that pupils made inaccurate predictions about how successful.

Accuracy

Equation. (5) shows that can determine the accuracy by dividing the sum of all correctly classified students in the dataset by the total number of students in the dataset [26].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (5)$$

Precision

The precision metric, as described in Equation (6), displays the proportion of students identified as belonging to the positive class or at risk:

$$Precision = \frac{TP}{TP+FR} \times 100 (6)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

Recall

One way to evaluate the proposed model's efficacy in predicting which pupils struggle academically is to look at its recall. It indicates the frequency with which the model properly identifies a student as at-risk when they actually fall into this group. An explanation of this is given in Equation (7):

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (7)$$

F1 Score

The following equations determine the F-measure, which is a measure of both recall and precision: (8):

$$F1 - score = \frac{2 \times recall \times precision}{recall + precision} \quad (8)$$

ROC Curve

ROC is a great method for seeing how well classifiers are doing their job. ROC is applied to the study of signal detection and to the interpretation of radar images. The ROC plots the compromise between the TP and FP rates graphically. Evaluations and the capacity of the model to precisely.

IV. RESULT AND DISCUSSION

This section presents the experimental results for early student success prediction using ML approaches on the Open University Learning Analytics Dataset (OULAD). Important measures for multiclass classification problems, such as F1-score, recall, accuracy, and precision, are used to assess the model's performance. The code is based on Python and its necessary libraries, including scikit-learn, XGBoost, pandas, NumPy, seaborn, and matplotlib. Executed it in a Google Colab Jupiter Notebook. The training of models was done on a computer with an NVIDIA RTX 3070 graphics card and a 32-GB memory to get optimal results in the experiments. The analysis suggested a performance Light Gradient Boosting Machine (LGBM), and the results of predictions with specific findings regarding the mode used to obtain the results are presented in the following way to prove the efficiency of the provided model to precondition the opportunity to create interventions in time and achieve better results for students.

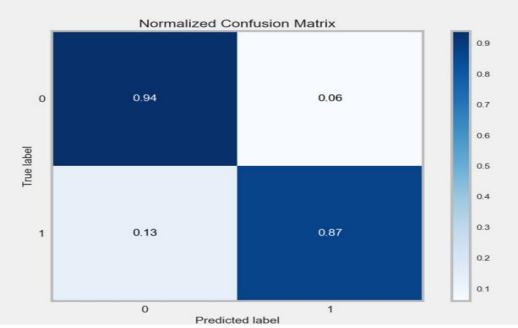

Figure 6: Confusion Matrix of LGBM Model

The LGBM model, which uses the early prediction model to categorise pupils' development, has a normalised confusion matrix, as shown in Figure 6. This model gets 94% accuracy to differentiate between the unsuccessful students (class 0) and 87% accuracy to also predict successful students (class 1). The combination of the matrix identifies low misclassification rates of 6% and 13% respectively, which express strong predictive applicability regarding early intervention approaches to an educational context.

Table 2: Proposed models Performance on early detection of student success on OU-LAD dataset

Measure	LGBM
Accuracy	92.23
Precision	94.40
Recall	93.21
F1-score	93.86
ROC AUC	96.24

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

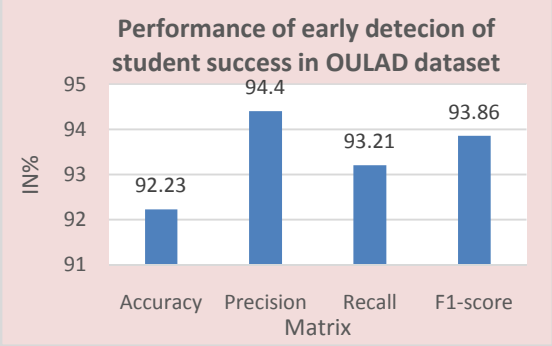

Figure 7: Bar Graph Performance Metrics of LGBM Model

Table II and Figure 7 show the results of the suggested Light Gradient Boosting Machine (LGBM) model in anticipating student achievement early through the OULAD. LGBM model has shown a high predictive performance, accuracy 92.23%, precision 94.40%, recall 93.21%, and F1-score 93.86%. These measures show that the model has the ability to predict at-risk students well together with reducing false flags. This high level of precision means good reliability in identifying successful students, and the high level of recall means the model would identify most at-risk learners to enable early interventions. The advantages of LGBM in Brooks' experiments are reflected in the balanced F1-score which shows that model is robust and can be applied to real-world educational environments, providing solution to scalable, data-driven early warning systems that can help to improve student and institution effectiveness.

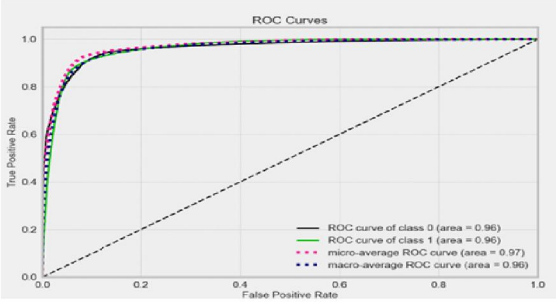

Figure 8: ROC Curve for LGBM Model

The ROC curve analysis in Figure 8 was used to review the discriminative ability of an early student success prediction model in binary classification tasks. The model shows excellent results as it reached 0.96 AUC per each separate classification (successful and unsuccessful student) as well as macro-average levels of 0.97. The model's high true positive rates and low false positive rates, shown by curves that are close to the ideal corner, make it reliable for use in the early education intervention system.

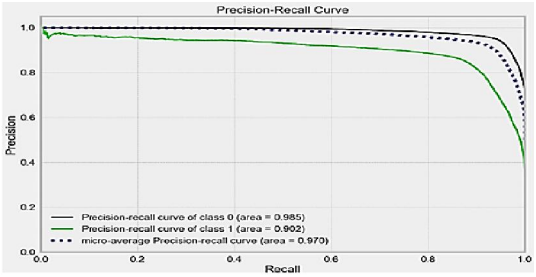

Figure 9: Precision-Recall Curve for LGBM Model

This precision-recall illustrates the performance of the predictive model to early student success classification where there is optimal precision-recall trade-off at various thresholds, as shown in Figure 9. The model presents a high area under curve of 0.985 in that persons who are not successful students (class 0), 0.902 on those who are successful and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 6, November 2024
Impact Factor: 7.57

0.970 macro-average. The curves are highly precise in different recall rates verifying the effectiveness of the model on making highly precise early predictive academic intervention inference.

A. Discussion

To predict early student achievement using the OULAD dataset, three ML models were compared and contrasted, as shown in Table III. The models were evaluated according to significant categorization criteria. The biggest overall performance of the Light Gradient Boosting Machine (LGBM) was found to be 92.23% of accuracy, 94.40% of precision, 93.21% of recall, and 93.86 of F1-score, which indicated that this type of machine had higher efficiency in predicting the outcomes of the students. The ANN fared better in comparison with 78% level of accuracy, 81 percent precision, 72% recall and 76% F1-score. The hybrid Fuzzy C-Means with Random Forest (FCM-RF) was always graded 88.33% on all 4 criteria, which makes it better than ANN but not better than LGBM These outcomes leave no doubt that in the early prediction tasks based on structural educational data, the tree-based ensemble approach, namely, LGBM is better to use than both neural network-based and hybrid approaches. It is, likely, due to the fact that LGBM is a valid candidate in making early warning systems in the education environment since it performs well with its excellent interpretability, capability to handle complex feature interactions, and unbalanced outcome.

Table 3: Comparison Between All Proposed Models and Existing Models For Early Prediction Of Student Success In Education Environment

Measure	LGBM	ANN[27]	FCM-RF[28]
Accuracy	92.23	78	88.33
Precision	94.40	81	88.33
Recall	93.21	72	88.33
F1 score	93.86	76	88.33

The proposed model of Light Gradient Boosting Machine (LGBM) imagination delivers well during the early prediction of student achievement and yields an accuracy of 92.23%, precision of 94.40%, recall of 93.21%, and the F1-score of 93.86%. LGBM is very accurate in predicting children at risk as it uses its highly efficient tree-based ensemble learning method to point out complex connections and patterns in the learning data set. It is particularly advantageous when applied to structured learning analytics owing to resistance to overfitting and ability to deal with inputs which are unbalanced. The two problems that continue to exist despite its excellent predictive power are an adaptation to changing student behavior with time and ensuring real-time application in education. By all means, the LGBM application provides a reliable and scalable method to develop early warning systems that give rise to timely action and increased student performance.

B. Application of AI for Student Success Prediction

The fields of data mining and learning analytics have poured resources into artificial intelligence research with the hope of predicting how well pupils do in school. Education institutions gather vast amounts of data from students through learning management systems (LMSs), student information systems (SISs), and other online platforms. This data includes student demographics, engagement levels, assignment and assessment records, and interaction details. These large and diverse datasets are hard to analyze manually; AI-based tools offer effective and scalable ways to develop patterns that could predict future performance of students.

Predictive modelling driven by AI allows us to early identify students who are most likely to have academic challenges. DT, LR, SVM, RF, and gradient boosting are among the machine learning methods that are used on the preprocessed dataset. Such models are fit on past data to gain understanding of factors related to the input data like previous grades, participation, and time-on-task and the outcome variable of student success [29]. The models would then be able to identify success rates to a great extent, by categorizing the students into various risk levels after being trained.

The AI-powered approach to prediction also enhances a better perceiving of the at-risk students at the early stages and delivers comprehensible information about the most important student outcome variables. The values of SHAP (Shapley Additive Explanations) and other model explainability and feature importance analysis tools can help

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 6, November 2024 Impact Factor: 7.57

educators determine which student behaviors and attributes are most predictive of future performance. With this knowledge, special and more personal approaches of treatment that are meant to increase learning results and experiences can be achieved.

V. CONCLUSION AND FUTURE WORK

A ML framework that utilizes AI and advanced data analysis to predict student success at an early stage. Leveraging large-scale educational datasets using the OULAD, the approach accurately identifies at-risk students to enable proactive support. A thorough data preprocessing step is part of the suggested methodology. Next, features are selected using RFE. Finally, three models—LGBM, ANN, and FCM-RF—are evaluated. For early success prediction utilizing the OULAD dataset, this study presented a strong machine learning architecture utilizing AI and advanced data analysis. Achieving the maximum accuracy of 92.23%, LGBM topped the analyzed models, surpassing ANN at 78% and FCM-RF at 88.33%. In order to successfully identify kids at risk and promote timely interventions, the results demonstrate the potential of ensemble-based techniques. Future work focuses on integrating temporal models such as LSTMs or Transformers to capture student behavior over time and enhance prediction accuracy. Additionally, incorporating explainable AI (XAI) techniques improves model interpretability, and validating the framework across diverse datasets ensures broader applicability in real-world educational settings.

REFERENCES

[1] A. F. Gkontzis, S. Kotsiantis, R. Tsoni, and V. S. Verykios, “An effective LA approach to predict student achievement,” ACM Int. Conf. Proceeding Ser., pp. 76–81, 2018, doi: 10.1145/3291533.3291551.
[2] E. Alyahyan and D. Düştegör, “Predicting academic success in higher education: literature review and best practices,” 2020. doi: 10.1186/s41239-020-0177-7.
[3] A. A. Khan et al., “Instructional Leadership and Students' Academic Performance: Mediating Effects of Teacher’s Organizational Commitment,” Int. J. Learn. Teach. Educ. Res., vol. 19, no. 10, pp. 233–247, Oct. 2020, doi: 10.26803/ijlter.19.10.13.
[4] M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learn. Environ., vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
[5] V. Shah, “Scalable Data Center Networking: Evaluating VXLAN EVPN as A Next-Generation Overlay Solution,” Asian J. Comput. Sci. Eng., vol. 8, no. 04, 2023, doi: https://doi.org/10.22377/ajcse.v8i04.237.
[6] B. Albreiki, N. Zaki, and H. Alashwal, “A Systematic Literature Review of Student’ Performance Prediction Using Machine Learning Techniques,” Educ. Sci., vol. 11, no. 9, 2021, doi: 10.3390/educsci11090552.
[7] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, “Student Performance Prediction Model based on Supervised Machine Learning Algorithms,” IOP Conf. Ser. Mater. Sci. Eng., vol. 928, no. 3, 2020, doi: 10.1088/1757-899X/928/3/032019.
[8] V. S. Thokala, “A Comparative Study of Data Integrity and Redundancy in Distributed Databases for Web Applications,” Int. J. Res. Anal. Rev., vol. 8, no. 04, pp. 383–390, 2021.
[9] L. Aguagallo, F. Salazar-Fierro, J. García-Santillán, M. Posso-Yépez, P. Landeta-López, and I. García-Santillán, “Analysis of Student Performance Applying Data Mining Techniques in a Virtual Learning Environment,” Int. J. Emerg. Technol. Learn., vol. 18, no. 11, pp. 175–195, Jun. 2023, doi: 10.3991/ijet.v18i11.37309.
[10] H. Zeineddine, U. Braendle, and A. Farah, “Enhancing prediction of student success: Automated machine learning approach,” Comput. Electr. Eng., vol. 89, p. 106903, 2021, doi: https://doi.org/10.1016/j.compeleceng.2020.106903.
[11] C. Kaensar and W. Wongnin, “Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning,” Eurasia J. Math. Sci. Technol. Educ., 2023, doi: 10.29333/ejmste/13863.
[12] Y. Mao, R. Zhi, F. Khoshnevisan, T. W. Price, T. Barnes, and M. Chi, “One minute is enough: Early prediction of student success and event-level difficulty during a novice programming task,” EDM 2019 - Proc. 12th Int. Conf. Educ. Data Min., no. Edm, pp. 119–128, 2019.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 6, November 2024 Impact Factor: 7.57

[13] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," in Proceedings of the 2010 InSITE Conference, 2010. doi: 10.28945/1281.
[14] K. Murugandi and R. Seetharaman, "A Study of Supplier Relationship Management in Global Procurement : Balancing Cost Efficiency and Ethical Sourcing Practices," pp. 724–733, 2022, doi: 10.48175/IJARSCT-7744B.
[15] S. M. Pande, "Machine Learning Models for Student Performance Prediction," in 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), 2023, pp. 27–32. doi: 10.1109/ICIDCA56705.2023.10099503.
[16] M. Ridwan, T. Wahyono, A. Iriani, R. Darmastuti, Sukisno, and A. H. Nugroho, "K-Fold Cross-Validation Analysis for Student Status Prediction Using Classification Models," in 2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2023, pp. 247–252. doi: 10.1109/ICITISEE58992.2023.10405343.
[17] K. Alalawi, R. Chiong, and R. Athauda, "Early Detection of Under-Performing Students Using Machine Learning Algorithms," in 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), 2021, pp. 1–8. doi: 10.1109/CITISIA53721.2021.9719896.
[18] E. Drousiotis, L. Shi, and S. Maskell, "Early Predictor for Student Success Based on Behavioural and Demographical Indicators," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12677 LNCS, pp. 161–172, 2021, doi: 10.1007/978-3-030-80421-3_19.
[19] H. Gull, M. Saqib, S. Z. Iqbal, and S. Saeed, "Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning," in 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1–4. doi: 10.1109/INOCON50539.2020.9298266.
[20] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a Web-Based Prediction System for Students' Academic Performance," Data, vol. 7, no. 2, 2022, doi: 10.3390/data7020021.
[21] T.-T. Huynh-Cam, L.-S. Chen, and K.-V. Huynh, "Learning Performance of International Students and Students with Disabilities: Early Prediction and Feature Selection through Educational Data Mining," Big Data Cogn. Comput., vol. 6, no. 3, 2022, doi: 10.3390/bdcc6030094.
[22] S. M. F. D. Syed Mustapha, "Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods," Appl. Syst. Innov., vol. 6, no. 5, 2023, doi: 10.3390/asi6050086.
[23] M. A. A. Al-Masoudy and A. Al-Azawei, "Proposing a Feature Selection Approach to Predict Learners' Performance in Virtual Learning Environments (VLEs)," Int. J. Emerg. Technol. Learn., 2023, doi: 10.3991/ijet.v18i11.35405.
[24] E. Alhazmi and A. Sheneamer, "Early Predicting of Students' Performance in Higher Education," IEEE Access, 2023, doi: 10.1109/ACCESS.2023.3250702.
[25] F. Afrin, M. Hamilton, and C. Thevathyan, "On the Explanation of AI-Based Student Success Prediction," 2022, pp. 252–258. doi: 10.1007/978-3-031-08754-7_34.
[26] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review," Appl. Comput. Intell. Soft Comput., vol. 2022, 2022, doi: 10.1155/2022/8924028.
[27] M. Hooda, C. Rana, O. Dahiya, J. P. Shet, and B. K. Singh, "Integrating la and EDM for Improving Students' Success in Higher Education Using FCN Algorithm," Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/7690103.
[28] M. A. Baig, S. A. Shaikh, K. K. Khatri, M. A. Shaikh, M. Z. Khan, and M. A. Rauf, "Prediction of Students Performance Level Using Integrated Approach of ML Algorithms," Int. J. Emerg. Technol. Learn., 2023, doi: 10.3991/ijet.v18i01.35339.
[29] T. T. Mai, M. Bezbradica, and M. Crane, "Learning behaviours data in programming education: Community analysis and outcome prediction with cleaned data," Futur. Gener. Comput. Syst., vol. 127, pp. 42–55, Feb. 2022, doi: 10.1016/j.future.2021.08.026.