10.16920/jeet/2026/v39i4/26112

Unifying Data Refinement and Fusion Strategies: A Cutting-Edge Methodology for E-Learning Performance Optimization

[1]N S Koti Mani Kumar Tirumanadham, [2]Pravin R. Kshirsagar, [3]Subba Rao Polamuri, [4]Anurag Sharma, [5]I Lakshmi Manikyamba

[1]School of Computer Science & Engineering, VIT-AP University, Amaravathi, 522237, Andhra Pradesh, India

[2]Professor, Dean (R&D), J D College of Engineering and Management, Nagpur, India;

[3]Associate Professor, Department of Computer Science and Engineering, Aditya University, Surampalem, India.

[4]Associate Professor, School of Electrical and Electronic Engineering, Newcastle University in Singapore (NUiS), Singapore.

[5]Associate Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad, UCESTH, Kukatpally, Hyderabad, India;

[1]manikumar1248@gmail.com , [2]pravinrk88@yahoo.com , [3]subbaraop@adityauniversity.in ,[4]anurag.sharma@newcastle.ac.uk , [5] I.L.Manikyamba@jntuh.ac.in ,

Abstract- The surge in the e-learning market has compelled the need for intelligent predictive models to minimize the gap between performance analysis and personalized learning systems to improve student performance. With the wide growth of the e-learning market, there is a great demand for intelligent predictive models to bridge the gap between performance analysis and personalized learning systems to boost student performance. Yet, traditional prediction methods are not effective due to issues like class imbalance, outliers, irrelevant features, and limited model interpretability. This research introduces an advanced framework to predict student academic performance based on an educational data mining dataset, Students' Academic Performance Dataset (xAPI-Edu-Data), comprising 480 instances, 16 attributes, and multi-variate integer and categorical features of the e-learning environment and educational data mining. The proposed framework involves class imbalance handling using Synthetic Minority Oversampling Technique (SMOTE), outlier removal using Interquartile Range (IQR), feature scaling and data standardization using Z-score normalization. A novel hybrid feature selection method TRIFEX is proposed to select the most influencing features to the student performance by combining ANOVA F-statistics, Recursive Feature Elimination (RFE) and Lasso regularization. The Logistic Regression, Decision Tree, and K-Nearest Neighbor (KNN) classifiers are used in the study. The hyperparameter optimization is done using Randomized Search CV, Grid Search CV, and Optuna to increase the efficiency and generalization power of the model. In addition, a voting, based ensemble model is created to fuse the virtues of individual classifiers for good prediction. Experimental results have shown that the proposed ensemble model is more accurate with 98.99% accuracy, 99.00% F1-score and 0.1005 RMSE as compared to the conventional predictive models. The results suggest that the suggested method has substantial potential to increase the accuracy of prediction, explainability of features and individualized learning support in contemporary e-learning environments.

***Keywords*- Predictive modelling, Feature selection, E-learning systems, Voting ensemble model, SMOTE algorithm, TRIFEX approach**

I. INTRODUCTION

E-learning, or electronic learning, represents a new era in education since learning is facilitated through the use of technology enhanced learning environments (Khanal et al., 2019). With technology gradually transferring to people's lives, e-learning is a convenient and flexible approach to learning. The current generation of learners can learn from computers, tablets and even from smart phones at any time of their convenience (Prenkaj et al., 2020). This carries a lot of benefits especially when students have to combine their studies with work or other endeavors that come their way. This is a versatile field that is open to a wide range of formats as a form of learning starting from MOOC to corporate training.

Perhaps, one of the most important benefits of e-learning is that it will afford everyone the chance to study (Kotsiantis, 2011). The main advantages of e-learning include the opportunity for students without regard to their geographical location, socio-economic status to gain access to quality education. It also favors Individualized education, whereby technologies may enable a teacher or system to provide the learner with an education programme that will meet the learner's learning schedule, preferred method of learning, and his/her level of learning. Also, e learning is effective in creating a class community through classroom communications, discussion, and group work and assignments even when physically wrongful.

Thus, e-learning also has several issues that may be encountered during the process of it. The first of them is related to equal distribution of the required technology as many learners lack stable internet connection or digital devices. It also important to enhance the players skills on online learning since a section of the learners might not understand technicalities involved in the use of the online learning platforms. However, it is difficult to keep students interested in the virtual space, have opportunities for

distractions and do not have the enthusiasm which is achieved through attending the classroom (Peng et al., 2024). To make the e-learning efficient we need to have interactivity, use of multimedia, and good interface design in our instruction. Thus, using these approaches, e-learning is capable of giving substantive and appealing learning experiences which can contribute to the further development of educational experiences.

The focus of the current study lies with data preprocessing, feature selection, and model construction in the spectrum of e-learning data analytics using predictive modelling techniques. The particular objectives of this research are to enhance the precision and neutrality of the predictive models and also to resolve the issues like an uneven ratio of classes, how to deal with outliers, and how to standardize the features (Bhaskaran & Marappan, 2021). By considering class balancing methods including SMOTE the research guarantees that the model treats all the students equally without discrimination. To obtain the final dataset with only the important and useful features subsequently ANOVA F-statistics, Recursive feature elimination, and Lasso methods are applied to give final dataset. On the model-building phase, the fusion of the multiple classifiers like Logistic Regression, Decision Tree and K-Nearest Neighbour (KNN) is incorporated. Thus, this paper offers a detailed plan of the approach that is useful for constructing stable and accurate prediction models in e-learning with a focus on increasing students' success rates individual approach.

A. *Scope of the study*

The subject area of this study is concentrated specifically on improving the performance of the predictive model on e-learning data analysis with considerations for class imbalance, outliers, and features. To increase model accuracy and fairness, the research will apply data balancing using SMOTE and feature selection method comprising of ANOVA F-statistics, Recursive Feature Elimination (RFE), and Lasso regularization. An enhanced desired model is made by integrating the two-source Logistic Regression method and Decision Tree approach, followed by fine-tuning the model by K Nearest Neighbors (KNN) method. It delivers a sound approach to enhance dependability in death prediction in e-learning: personal learning environments, and student outcomes analysis.

B. *Research Gaps*

Despite the advancements in predictive modelling within e-learning systems, several critical gaps remain unresolved. First, the issue of class imbalance is often inadequately addressed, leading to biased predictions towards majority classes. Outlier detection and feature standardization, crucial for improving model accuracy and robustness, are frequently overlooked or insufficiently implemented. Existing feature selection methods either lack interpretability or fail to combine statistical, recursive, and regularization-based techniques for optimal feature retention. Additionally, while ensemble models are increasingly applied, they often lack proper hyperparameter tuning, reducing their potential for maximum performance. The absence of a comprehensive, integrated framework that effectively combines preprocessing, feature selection, and ensemble learning techniques limits the scalability and accuracy of current models, particularly in the context of personalized e-learning systems.

C. *Research Questions (RQ)*

RQ1: How can class imbalance in e-learning datasets be effectively addressed to improve prediction accuracy and fairness across different student categories?

RQ2: What role does outlier detection and feature standardization play in enhancing the robustness and performance of predictive models in e-learning environments?

RQ3: How can a multi-stage feature selection method improve both model interpretability and prediction accuracy in e-learning systems?

RQ4: To what extent can optimized hyperparameter tuning techniques improve the performance of ensemble models in predicting student performance?

RQ5: How does a voting-based ensemble model, compare to individual models in terms of accuracy, F1-score, and RMSE in e-learning prediction tasks?

D. *Contributions*

- A novel, multi-stage feature selection approach that integrates ANOVA F-statistics, Recursive Feature Elimination (RFE), and Lasso regularization to improve the accuracy and interpretability of predictive models in e-learning systems.
- The research proposes an optimized data preprocessing framework combining Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance, Interquartile Range (IQR) for outlier detection, and Z-score normalization, ensuring improved data quality and model robustness.
- Development of a fusion ensemble model that combines Logistic Regression, Decision Tree, and K-Nearest Neighbours (KNN) through a voting mechanism, showing significant performance improvements over individual models.
- Utilization of advanced hyperparameter tuning techniques (Randomized Search CV, Grid Search CV, and Optuna) to further enhance the accuracy, stability, and convergence of the predictive models.
- The proposed methodology achieves a notable improvement in prediction performance,surpassing existing methods in the literature.

The structure of the study consists of several separate sections to give a well-structured and polished look to the study's goals and conclusions. Section 1 of the entire study is the Introduction which aims at identifying the research problem, explaining why the given research issue is important, and what objectives the given methodology is going to meet. Section 2, Literature Review, discusses some literature in the area of Predictive Modeling, Machine Learning Application, and techniques to deal with some issues such as class imbalance as well as feature selection. In Section 3, Proposed Methodology, the strategies for overcoming the problems with e-learning data are depicted and types of data preprocessing, feature selection, and model construction are described. Section 4 of the study focuses on the results and discussions, which describes the findings of

the study together with the assessment of different machine learning methods. In Subsection 3.1 Performance Assessment based on ML Algorithms, the performance of the models is analysed based on some performance indicators like accuracy, precision, recall, f1 score, RMSE. Section 5 lays out the comparison of the presented methodology with relation to other methods for increased understanding of their merits and efficiency. Section 6, Discussion, gives a critical analysis of the findings affirming the findings made earlier and their significance. Lastly, Section 7 titled Conclusion with Future Scope provides an overview of the above sections, draws conclusion and lastly; the author has suggested different ways to enhance and extend the current study. The structuring of the research presented in the document guarantees the reader can easily follow the sequence of the research run.

II. Literature Review

Amrieh et al., (2016), propose the use of ensemble methods in enhancing educational data mining for increased prediction of student academic performance. The study also proposes a new approach of the prediction model involving the behavioral characteristics developed based on the student's usage of e-learning management system. This model is assessed by the researchers with different classifiers such as Artificial Neural Networks, Naïve Bayes, Decision Trees and so on. They also utilize ensemble methods of classifiers, including Bagging, Boosting and at Random forests for fine tuning these classifiers. The findings also show the ability to relate several of the student behaviors to academic performance, with the given model obtaining up to 25.8% of the improved accuracy levels through the use of ensemble skills. Further, when used with behavioral features, the model was up to 22. 1% better than model that did not have those features. More than 80 percent accuracy for the outcomes of new, unlabeled students made the model rather relevant and effective in its use. The findings of this study re-emphasize the utility of the behavioral data and various ensemble techniques for increasing the predictive precision and provide ideas for educators and administrators to strengthen the educational prospects and sericulture.

Asif et al., (2017), proposed a solution to this problem when they applied data mining on the huge educational databases for gaining useful insights. Their study focuses on analyzing undergraduate students' performance through two main objectives: forecasting the performance during the end of a four-year program and analyzing progression trends. Using the Naïve Bayes classifier the researchers reported an average classification rate of 83.65% the student grades indicating that the probability-based form of classifier is capable of providing good results in educational models. In particular, the work reveals courses that can be potentially powerful predictors of the student performance, thus offering the means for intervention. For low achieving student, the model gives alerts and assistance during early stages while for the high achieving students they are directed and offered chance to move to next level. This approach highlights the idea that data mining can be used to enrich educational processes and provide specific recommendations on how to develop more effective supportive measures and the educational results. Consequently, the study reveals the effectiveness of using specific course data in order to implement successful interventions with regards to student success, and the usefulness of data mining for enhancing the learning process quality.

Kaviyarasi & Balasubramanian, (2018), engage in a study of high impact factors that affect students' performance through the DM approach. Their study stems from a concern for bringing improvement to learning achievements given growing student enrolments and shift in educational tasks. The researchers focus on classifying students into three categories: In view of this the present work focus on identifying the factors that have a major impact on the students' performance thereby categorizing students as Fast Learners, Average Learners, and Slow Learners. The study adopts the Extra Trees classifier to evaluate the permutation importance of features; thus, assisting in the understanding of key factors influencing performance. While Raw scores and other accuracy measures 80%, the study will seek to enhance raw accuracy in future models. As such, the findings by the research can augment understanding of the specific determinants of student performance in order to control for them in promoting education. This approach identifies the capacity of Data Mining in responding to educational dilemmas by emphasizing learning preferences and enhancing support systems. The future direction of this work includes increasing the accuracy of the model's predictions so as to give a clearer indication of the manner and degree in which performance could be improved for the learners.

Popescu & Leon, (2018), carry out a study on using social media traces to predict students' academic performance has emerged as a new topic in learning analytics. Classic techniques of predictive modeling use data in a state-centered perspective including demographic information and/or previous performance. However, this study focuses on event-driven data collected from the use of social media tools by students—wikis, blogs, and micro-blogs— in a Web Applications Design course. The work entails data from 343 student's data gathered over six course installments and uses a new form of a regression algorithm for grade forecasting. It emerges that this approach yields high accuracy in gradients; one-point deviations include 85% of predictions a clear improvement to traditional regression analysis. This research affirms that higher involvement in the use of the SMT is likely to be correlated with improved academic performances, which indicates that the SMI provide useful cues about the students' performance. This work also has clear implications for how predictive models of student performance could be used to improve educational engagement with students in social media ecosystems and the associated digital learning spaces.

Beaulac & Rosenthal, (2019), has focused their study on the use of random forests to model academic results at a large Canadian university employing a rich dataset that included the previous ten years. Their study addresses two key predictions: whether students will finish their course and the specific major of students who will finish their programs. Based on the first two semesters' content, the authors construct classifiers using the random forests, which are much more accurate in comparison with linear classifiers. The study realizes 78% accuracy of identifying students who will complete programs of study and 47.41% of identifying

students' majors. These results demonstrated the capability of random forests to process big educational datasets accurately, offering valid variable importance insights, as well as enhance general accuracy. Course performance, especially from underperforming faculty areas such as Mathematics, Economics, and Finance, is established to be important factors that determine completion rates of degrees. The study indicates that, although the accuracy for predicting majors is lower, there might be enhancement such as; The multi-label classification should be considered for improvement; The handling of missing values could also be improved. In sum, the study supports the use of random forest for the educational predictions as well as identifying ways for future developments of the models.

Enughwure A et.al, (2023), focus on the difficulty of predicting the performance of students in engineering drawing courses as a core aspect of the learning of engineering courses using SMOTE augmented machine learning algorithm. The survey involved the use of both logistic regression and decision tree as the prediction models of student's outcomes whereby the data was collected by administration of paper base questionnaires across different engineering departments. The research was completed using the Python language found on Kaggle, with SMOTE applied for balance as it synthetically creates samples for the shortest classes. The study showed that the overall and individual predictive models obtained accuracy rates between 67% and 78% and higher prediction accuracy was found with logistic regression method. The use of SMOTE enhanced the model as it played a high level of improvement in terms of balancing out the datasets to achieve better prediction. Concerning the implications for future research, the study points to the effectiveness of using machine learning techniques in connection with SMOTE and alerts other scholars to specific machine learning approaches together with an increased understanding of how sophisticated analysis methods can be employed in the context of engineering education to address performance problems. At the same time, the use of this approach not only allows finding the reliable forecasting of the student outcomes, but also helps in the further creation of the pertinent interventions for difficult subjects for students.

For diabetes prediction, (Gupta & Goel, 2023), employed hyperparameters tuned machine learning to evaluate the classifier through preprocessing techniques. Their work employs the PIMA Indian Diabetes data set and compares few classification techniques such as K-Nearest Neighbors, Decision Trees, Random Forests, and Support Vector Machines. To improve model performance, the researchers compared preprocessing by tuning hyperparameters and also compared data preprocessing techniques. The evaluation of the four models has found that the highest accuracy was achieved by the Random Forest classifier with 88.61% and an F1 score of 75.68%. This was done by pre-processing the data set to purge out any samples that had missing / unknown values (model D3). From the study the authors highlighted issues to do with hyperparameters and preprocessing of data in building reliable predictive models for the classification of diseases. The findings as shown have provided strong evidence that preprocessing and model optimization hence enhances the accuracy of the prediction systems in making more efficient diagnosis in health care.

In the existing techniques reviewed, several technical gaps are evident, leading to the design of the proposed methodology shown in Table.I. Many studies fail to address class imbalance adequately, which often skews model performance towards the majority class. While some methods like SMOTE have been utilized, they are not consistently applied across different research contexts. Outlier detection is another overlooked aspect in many studies, potentially leading to biased results. Additionally, feature selection techniques are either too simplistic, relying on single methods like Naïve Bayes or decision trees, or too complex without yielding interpretability. Although some studies have applied ensemble learning methods, they lack robust hyperparameter tuning strategies, limiting model optimization. Furthermore, few approaches integrate multiple preprocessing steps, such as outlier detection, class balancing, feature standardization, and advanced feature selection, in a unified framework. These gaps highlight the need for a more holistic, integrated approach—like the one proposed in this study that incorporates advanced data preprocessing, effective feature selection, and optimized ensemble methods to improve prediction accuracy and model stability.

TABLE I
A SUMMARY OF THE BASELINE METHODS IN THE LITERATURE

Author(s)	Method	Focus of the study	Contribution	Evaluation	Source of the dataset	Observation (+ strength and – weak points)
Abu Amrieh et al.	Ensemble Methods (Bagging, Boosting, Random Forests)	Enhancing prediction of student academic performance	Introduced a novel model incorporating behavioral features and ensemble methods to improve prediction accuracy.	Up to 25.8% improvement in accuracy with ensemble methods.	E-learning management systems	+ Significant improvement with ensemble methods; - Limited focus on specific feature impacts.
Asif et al.	Naïve Bayes	Analyzing and predicting undergraduate students' performance	Achieved 83.65% accuracy in predicting student grades; identifies key courses for targeted interventions.	Effective in predicting grades and identifying key courses.	Pakistan Universities	+ Useful for early warnings and targeted interventions; - May not address all factors influencing performance.

Kaviyarasi et al.	Extra Trees classifier	Identifying factors affecting academic performance	Assessed feature importance for classifying students into performance categories; aims to improve prediction accuracy.	80%, aims for future improvements.	Students' dataset	+ Focus on classifying students into performance categories; - Accuracy metrics not provided.
Popescu et al.	Regression Algorithm (Social Media Data)	Predicting academic performance using social media traces	Demonstrated strong predictive accuracy by incorporating social media engagement into models.	85% of predictions within one point of actual grades.	Social media interactions	+ High predictive accuracy with social media data; - Focuses on specific course and social media tools.
Beaulac et al.	Random Forests	Predicting academic outcomes at a Canadian university	Achieved 78% accuracy for degree completion prediction; highlighted the importance of specific grades.	78% accuracy for degree completion, 47.41% for majors.	Comprehensive 10-year dataset	+ Effective with large datasets; - Lower accuracy for predicting majors; potential for improvement in handling missing values.
Enughwure et al.	Logistic Regression, Decision Trees + SMOTE	Predicting performance in engineering drawing courses	Used SMOTE to address class imbalance; logistic regression showed highest accuracy.	Accuracies ranged from 67% to 78%, with SMOTE enhancing performance.	Paper-based questionnaires	+ Improved performance with SMOTE; - Specific to engineering drawing courses and may not generalize.
Gupta et al.	Hyperparameter-Tuned ML Techniques (KNN, Decision Trees, Random Forests, SVM)	Diabetes prediction using various preprocessing methods	Highlighted the impact of hyperparameter tuning and preprocessing on model performance.	Random Forest achieved 88.61% accuracy; F1-score of 75.68%.	PIMA Indian Diabetes dataset	+ Effective preprocessing and tuning; - Focus on diabetes prediction may not generalize to other health issues.
Our Work	Data Cleaning, SMOTE, IQR, Z-Score Normalization, Fusion Model	Optimizing predictive modelling for student performance in e-learning systems	Developed a robust predictive model using feature selection techniques (TRIFEX) and fusion-based classification	Model 4 achieved 98.99% accuracy, with strong precision (99.04%) and low RMSE (10.05%)	E-learning dataset (gender, nationality, academic level, performance)	+ High accuracy due to advanced data preprocessing and model fusion; - May require validation on larger or more diverse datasets to confirm generalizability.

III. Proposed Methodology

The subject of the study is to outline a detailed and fine-tuned approach to approach the challenges of predictive modeling when it comes to analyzing e-learning data, with the principles of data preparation, engineering, and modeling laid out as the key areas of focus. The preprocessing step's aim is to make data accurate and to build fair models for prediction. In general, a complete e-learning dataset without any missing attributes was identified and included the attributes; Gender, Nationality, Academic Level and performance indicators. Another factor can be the number of instances in each class, so Synthetic Minority Over-sampling Technique (SMOTE) (R et al., 2023) was used in order to level it to 143 instances per class. This correction mitigates risks of model bias to the majority class and ensures all classes are more or less given polite equality by the model. Extreme values in the dataset were managed using the IQR method in an attempt to increase the strength of the dataset while data points outside the IQR (Wan et al., 2014) range were considered outliers. Furthermore, feature scaling was performed through Z-score normalization procedure to ensure all features have mean=0 and sd=1.

This standardization also greatly benefits the performance and convergence of all added features rather than skewing the process. Recursive Feature Elimination (RFE) (Sanz et al., 2018) and Lasso regularization (Duan et al., 2016) were applied to perform the feature selection and ANOVA (Shaw & Mitchell-Olds, 1993) F-statistic was used to rank the features. In ANOVA, we select the features that are statistically different; RFE deletes features with insignificant importance in each subsequent iteration; and Lasso acts as a penalty that eliminates unimportant features. This multi-stage approach also guarantees that only the most informative features with high predictive capability are retained to achieve higher model accuracy and robust interpretation. In model building, an integration of the method was done and used in model building which includes Logistic Regression (Bandela et al., 2023), Decision Tree (Hall et al., 2002), and K-Nearest Neighbours (KNN) (Guo et

al., 2003). Optimization of hyperparameters for these models was done using Randomized Search CV (Vishnu et al., 2023), Grid Search CV (Ranjan et al., 2019), and Optuna (Srinivas & Katarya, 2021), and the workflow shown in Fig.1. The third and the last of the fusion models used in this study is the voting fusion, which gathers predictions by virtue of the voting mechanism and brings the strengths of the classifiers into play to enhance the accuracy and stability of the system, which offers a sound framework as an approach to complicated classification problems.

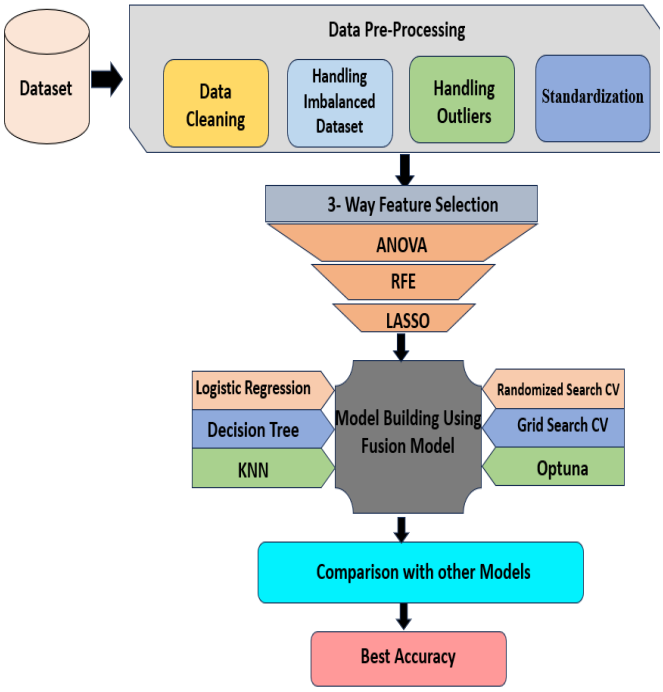

Fig.1. Proposed Methodology Workflow

A. *Data Collection*

This e-learning dataset is sourced from Kaggle and contains information organized based on engagement and learning outcomes associated with a range of demographic and curricular descriptions. It includes features like gender, nationality and place of birth, educational features like stage, group number, section and topics studied as seen from Fig.2 below. The dataset also reflects student engagement with learning through technologies by captured raised hands, visited resources, announcements, and discussion boards. In addition, it consists of factors that are associated with family, including participation in survey by parents and school satisfaction. Recorded at the student level is truancy, with the total performance class using alphabets; M for good performance and L for lower performance. This dataset provides an understanding of the multifaceted causal relationships between several antecedents to the learning outcomes of a learner in an e-learning environment; these factors include demographics, learner engagement, parental involvement, and academic environment. The important of the dataset is in tracking the students' behaviour, understanding the variables influencing performance and building various student performance trends.

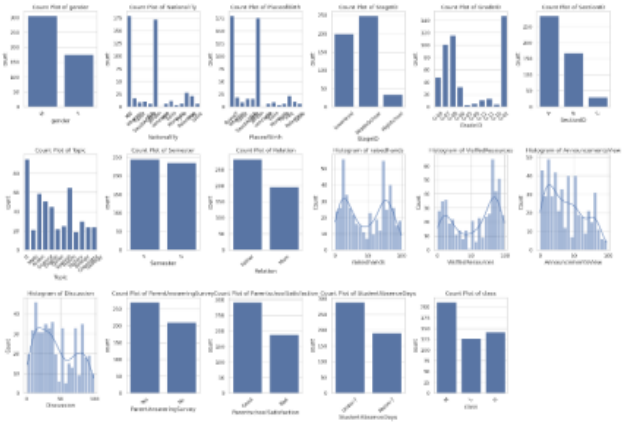

Fig.2. Visualizing the columns in the Dataset Data Pre-processing

B. *Data Cleaning*

The management of missing values plays a critical role in data preprocessing to ensure the validity of research findings. When looking at the descriptive and quantitative features of the given e-learning dataset and its features that include the gender, nationality, academic level, raised hands and the visited resources and other related performance indicators no form of missing values were observed. Before and after processing every one of the 17 columns had 480 complete entries on it which means that the dataset is in its correct and coherent form and is therefore ready for the rest of the analysis. Missing data are widely known to create an offshoot effect in standard research practices and undermine the validity of outcomes due to bias. After processing, indicating that the dataset is fully intact and ready for further analysis. In standard research practices, missing data can significantly impact the validity of results, potentially introducing bias. Common methods for handling missing data include imputation techniques such as replacing missing values with the mean, median, or mode for numerical or categorical columns. More Statistical information about the dataset after handling missing values shown in Fig.3, and sophisticated approaches might involve predictive models to estimate missing values based on correlations between variables. Unlike other datasets that have missing values, the dataset is complete, which brings convenience in the data preprocessing phase of the analysis so that more complex investigation does not have to worry about the impact of missing data. This makes the transition to the model-building stages easier and provide the results of the research with more assurance.

```
Missing values after handling:
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 0
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
class                       0
dtype: int64
```

Fig.3. Statistical information about the dataset after handling missing values

C. Handling Imbalanced Dataset using SMOTE

Data pre-processing specifically dealing with imbalanced datasets plays a critical role in ensuring the correct responsible usage of the machine learning models across the classes involved. In this analysis, the dataset originally exhibited class imbalance with the following distribution: We tested 'M' class object, 211 times, 'H' class object 142 times and 'L' class object 127 times. Such imbalance depicted in Table II and Fig 4 can produce models that favor the majority class meaning their performance for minority classes is subpar. Realising this problem, the authors used the Synthetic Minority Over-sampling Technique (SMOTE) to overcome this by balancing the given data set. The class distribution was balanced as depicted in Table.III and Fig.5 by having 143 instances per class after applying SMOTE (Elreedy & Atiya, 2019). This balancing was done by creating fake data set samples for the minority classes hence making data set to be balanced. These are evident from the results that show a much better class balance which is crucial when training machine learning models. Furthermore, the function allows getting rid of the class distribution in order to increase the fairness in evaluating the model's performance across classes and, thereby, enhance the model's overall reliability and accuracy. The division of the dataset means greater generalization of the results and improved prediction in further analyses and applications.

TABLE II
CLASS DISTRIBUTION BEFORE HANDLING IMBALANCE

Class Distribution Before Handling Imbalance	
Class	**Count**
M	211
H	142
L	127

Fig.4. A Bar graph Class Distribution Before SMOTE

TABLE III
CLASS DISTRIBUTION AFTER HANDLING IMBALANCE

Class Distribution After Handling Imbalance	
Class	**Count**
M	143
H	143
L	143

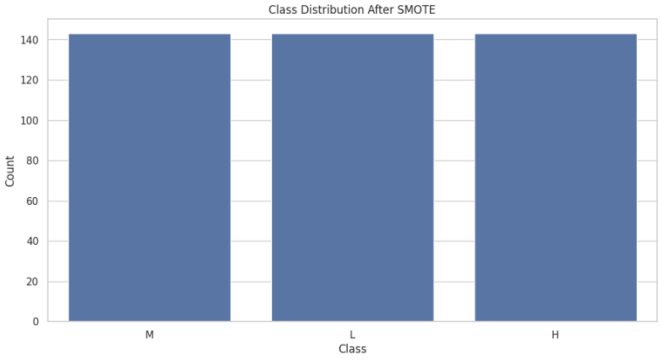

Fig.5. A Bar graph of Class Distribution After SMOTE

D. Handling Outliers Using IQR

Outlier elimination is an essential process of Data Preprocessing to increases the efficiency and credibility of statistical analysis and machine learning. Anything falling outside the range of variation can alter figures such as mean, standard deviation and therefore give wrong information. Interquartile Range (IQR) method is one of the common methods used in tackling with outliers. This method involves determination of IQR (Bernardet & Verschure, 2010), the quantity resulting from the difference between Q1 and Q3 of the data. Outliers can then be defined as any observation which is less than Q1−1.5 × IQR or greater than Q3+1.5× IQR. In the data processing step of the work, the range determination of the fourth quartile was used as an additional procedure for handling outliers in numerical fields of the dataset. This approach assists in removing outliers in the data distribution as depicted in the Fig.6 that if included in analysis would significantly distort the results. In the process of outlier handling, the authors compared the distributions limited to the mean and the distributions after the removal of the outliers by constructing boxplots that would explain the process. In other words, adjustments of the data in a predetermined range can have a positive impact on the further analysis of data outcomes and the building of the predictive mathematical model.

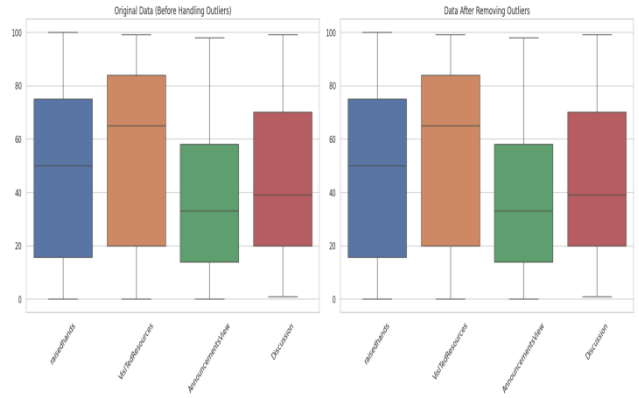

Fig.6. After Handling Outlier Using IQR

E. Standardization using z-score Normalization

Standardization, also known as Z-score (Zhang et al., 2014) normalization, transforms data to have specific statistical properties: a mean of 0 and a standard deviation of 1. This process is crucial for ensuring that features with different scales and units are comparable and contribute equally to the analysis.

The Z-score of a data point x is calculated using the following (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where:

- x is the original value of the feature.
- μ is the mean of the feature's values.
- σ is the standard deviation of the feature's values.

Mean (μ) Calculation: The mean of a feature X with n observations are given by using (2):

$$\mu = \frac{1}{n}\sum_{ii=1}^{n} x_{ii} \quad (2)$$

Standard Deviation (σ) Calculation: The standard deviation is computed by using the following (3):

$$\sigma = \sqrt{\frac{1}{n}\sum_{ii=1}^{n}(x_{ii} - \mu)^2} \quad (3)$$

Z-score Transformation: Each value x_{ii} in the feature is then transformed to its Z-score using the (4):

$$z_{ii} = \frac{x_{ii} - \mu}{\sigma} \quad (4)$$

This standardization process re-scales the feature such that:

- The transformed feature has a mean of 0.
- The transformed feature has a standard deviation of 1.
- The distribution of the transformed feature approximates the standard normal distribution (mean = 0, standard deviation = 1).

Standardization ensures that all features contribute equally to the model training process shown in Table IV and Table V, particularly in algorithms sensitive to the scale of data, such as gradient descent-based methods and distance-based algorithms.

TABLE IV
STATISTICAL INFORMATION BEFORE STANDARDIZATION

Statistic	raisedhands	VisITedResources	Announcements View	Discussion
count	480.000000	480.000000	480.000000	480.000000
mean	46.775000	54.797917	37.918750	43.283333
std	30.779223	33.080007	26.611244	27.637735
min	0.000000	0.000000	0.000000	1.000000
25%	15.750000	20.000000	14.000000	20.000000
50%	50.000000	65.000000	33.000000	39.000000
75%	75.000000	84.000000	58.000000	70.000000
max	100.000000	99.000000	98.000000	99.000000

TABLE V
STATISTICAL INFORMATION AFTER STANDARDIZATION

Statistic	raisedhands	VisITedResources	AnnouncementsView	Discussion
count	4.800000e+02	4.800000e+02	4.800000e+02	4.800000e+02
mean	4.440892e-17	5.921189e-17	-5.921189e-17	7.031412e-17
std	1.001043e+00	1.001043e+00	1.001043e+00	1.001043e+00
min	-1.521279e+00	-1.658255e+00	-1.426401e+00	-1.531590e+00
25%	-1.009037e+00	-1.053029e+00	-8.997598e-01	-8.433262e-01
50%	1.048878e-01	3.087281e-01	-1.850301e-01	-1.551430e-01
75%	9.179714e-01	8.836924e-01	7.554025e-01	9.676820e-01
max	1.731055e+00	1.337612e+00	2.260095e+00	2.018067e+00

F. A 3-Way Feature Selection Using TRIFEX

TRIFEX also known as Triple-Feature Extraction (ANOVA, RFE, LASSO) is a complex feature selection technique for improvement of the interaction between machine learning models and the features used. First of all, using ANOVA F-statistic, TRIFEX selects the most statistically significant features in order to provide a statistically sound starting benchmark. After that, Recursive Feature Elimination (RFE) (Shieh & Yang, 2007) improves this subset by gradually applying a model and eliminating the least effecting features, so the most effective predictors remain. Lastly, for selecting a final optimal subset, Lasso regularization is used to do the final pruning in order to encourage feature selection. This is a multiple-step approach because it enhances the quality of models and reasons more about input variables while enhancing the predictive abilities of the selected features. As demonstrated, the integration of these techniques enables TRIFEX to offer a logical analysed approach for feature selection in Big Data.

1) ANOVA (Analysis of Variance)

ANOVA (Kim, 2017) is a statistical method used to compare the means of three or more groups to see if at least one group's mean is significantly different from the others. The key idea is to partition the total variance into variance between groups and variance within groups.

The F-statistic is calculated by using (5):

$$F = \frac{MS_{btw}}{MS_{Wtin}} \quad (5)$$

The mean square between the groups is shown in (6):

$$MS_{Between} = \frac{SS_{btw}}{df_{btw}} \quad (6)$$

The mean square within the groups represented in (7):

$$MS_{Within} = \frac{SS_{Wtin}}{df_{Wtin}} \quad (7)$$

- SS_{btw}, is the sum of squares between the groups,
- SS_{Wtin}, is the sum of squares within the groups,
- df_{btw}= k−1where kis the number of groups,
- df_{Wtin}= N−kwhere N is the total number of observations.

If the computed F-statistic is larger than the critical value from the F-distribution, the null hypothesis is rejected, suggesting that at least one group mean is significantly different.

2) Recursive Feature Elimination (RFE)

Recursive feature elimination (RFE) is a feature selection technique whereby models are successively trained with the objectives of establishing which characteristics have the most impact on the models. In this method, starting with the feature weights or coefficients of a given model estimator, it progressively eliminates the elements of the least importance until the requisite number of features is achieved. Since it is a way of filtering the data in order to make the next predictions, RFE narrows down the set of features to consider in the further prediction tasks to those which are most important for the task.

RFE works in a manner that first trains the model, assesses the importance level of each of the features; usually through the coefficients or feature weights and then eliminates the least important features. This process goes on until the right number of attributes is selected for characterizing the database. Mathematically, Recursive Feature Elimination (RFE) can be expressed in the following manner:

Feature Ranking Step:

Feature importance scores are calculated. For a linear model, this might be the magnitude of the coefficients shown in (8):

$$importance(ii) = |\beta_{ii}| \quad (8)$$

Rank features based on model coefficients/weights represented in (9):

$$w = w_1, w_{2,}, w_3 \dots \dots \dots \dots , w_p \quad (9)$$

Where β_{ii} is the coefficient for the ii − th feature. For tree-based models, feature importance might be calculated based on the decrease in impurity shown in (10).

$$importance(ii) = \sum_{t \in T} \Delta\, ii_t \,.\, 1(t\ uses\ feature\ ii) \quad (10)$$

Feature Elimination Step:

Remove least important feature(s) based on ranking criteria.

By iterating through these steps, RFE systematically identifies the subset of features that maximizes model performance, making it suitable for enhancing prediction accuracy and interpretability in machine learning tasks.

3) LASSO (Least Absolute Shrinkage and Selection Operator)

By integrating selecting variables and regularised, LASSO (Tibshirani, 1996) (Least Absolute Shrinkage and Selection Operator) has been utilised to advance model the precision and interpretability. Through the execution of an L1 penalty to the regression coefficients, the LASSO investigate ultimately eases some coefficients to zero. This approach provides a model that helps reduce model complexity, which is needed to handle high-dimensional data sets with a lot of amenities in with respect to observations, in instead of identifying the most essential ones. In one instance, we had submitted this L1 penalty to the Least Angle Regression (LAR) algorithm—which is closely associated with LASSO—in demand to find essential traits. Utilising the Select from Model function in combination with LAR, we were capable of to pick and choose features as shown to their weights of importance, which were started leveraging the LASSO methodology. By employing this kind of approach, we are able to sure that our prediction keeps consistent and is less probably too overfit, and they should improve its whole accuracy and comprehension.

LASSO objective function shown in (11):

$$ObjectiveFunction = min_{\beta}[\frac{1}{2n}\sum_{m=1}^{n}(\,(y_m - X_m^T)^2 + penalty\,] \quad (11)$$

where:

- y_m is the observed value for the m-th sample,
- X_m is the feature vector for the m-th sample,
- β is the vector of coefficients,
- λ is the regularization parameter controlling the strength of the penalty,
- p is the number of features.

L1 Penalty: The L1 penalty term added to the objective function shown in equation (12).

$$\text{L1 } Penalty = \lambda \sum_{k=1}^{p} |\beta_k| \quad (12)$$

Algorithm: - Algorithm for TRIFEX

Step 1: ANOVA F-statistic Feature Selection
Input: Dataset X (features), y (target).
Process:
Compute the ANOVA F-scores for each feature.
Rank features based on their F-scores.
Select the top k features.
Output: Subset F_{anova} of top k features.
Step 2: Recursive Feature Elimination (RFE) on ANOVA Subset
Input: Subset F_{anova}, Dataset X (filtered), y (target).
Process:
Apply RFE using a machine learning model on F_{anova}.
Iteratively remove the least important features based on model performance.
Select the top mmm features.
Output: Subset F_{RFE} of mmm features from F_{anova} .
Step 3: Lasso Regularization on RFE Subset
Input: Subset F_{RFE}, Dataset X (filtered), y (target).
Process:
Apply Lasso Regularization with a chosen alpha value on F_{RFE}.
Retain features with non-zero coefficients.
Select the top n features (e.g., top 6).
Output: Final set F_{final} of n selected features.
Final Output: Combined and refined set of features F_{final}.

G. Model Building Using Fusion Model

Fusion model is an enhanced version of ensemble learning model that combines the results of multiple machine learning models. The primary methodology of a fusion model rests in the possibility of enhancing certain characteristics inherent to one model, making up for the lack of the others. Logistic Regression, Decision Tree, and K-Nearest Neighbours (KNN) (S. Zhang et al., 2017) based model fusion along with three different hyperparameter tuning

algorithms: Randomized Search CV, Grid Search CV and Optuna optimization will be developed in the proposed research. This kind of model uses voting technique; that the final decision is made using the majority vote (hard voting) from the three classifiers. This approach improves the accuracy of the results, increases model stability, and decreasing the risk of overcomplication due to the different nature and complementary features of the models used. Moreover, the structure of the fusion being an assembling of essentially different algorithms makes it advantageous for intricate classification problems, which in turn should make it a strong candidate as a reliable decision maker for realistic applications.

1) Logistic Regression

Logistic Regression is statistical technique that deals with binary classifiers problems. It estimates a binary response probability by mapping it with a logistic function or sigmoid function. The model produces the probability that an input will be classified in a specific class. The output of Logistic Regression is in between 0 and 1 while setting the decision boundary at 0.5.

The logistic function represented as (13):

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (13)$$

Where z is a linear combination of the input features shown in (14):

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ \ldots\ldots\ldots\ldots + \beta_n x_n \quad (14)$$

Here, $\beta_0$is the intercept and $\beta_1, \beta_2 \ldots\ldots\ldots \beta_n$ are the coefficients for the input features $x_1, x_2, \ldots\ldots x_n$. The model is trained using maximum likelihood estimation, where the goal is to maximize the likelihood of correctly predicting the class labels.

Logistic Regression –the algorithm assumes that the linear relationship exists between the independent features and the logarithm of the dependent variable's odds thereby making efficient for simple classification models. This is especially useful when the target variable is categorical and binary in nature but can go a step further and used with softmax regression for multiclasses.

2) Decision Tree

A Decision Tree (Myles et al., 2004) is a type of learning algorithm that is used in both classification and regression exercises. This divides different subsets of a data set according to the value of input features and forms a tree structure where each node is a feature, each branch is a decision and each end node is a class label.

The splitting criterion is determined by impurity measures such as Gini Index or Entropy (for classification tasks):

- Gini Index shown in (15):

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \quad (15)$$

Where p_i is the probability of a particular class in the node.

- Entropy represented as (16):

$$Entropy = -\sum_{i=1}^{C} p_i \log_2 p_i \quad (16)$$

The goal is to find the splits that result in the highest information gain, which is the reduction in impurity after the split. The process continues recursively, growing the tree until all data points are perfectly classified or certain stopping criteria are met, such as maximum depth. Decision Trees are interpretable and versatile, but they tend to overfit, especially with complex data. Techniques like pruning and ensemble methods are often used to improve generalization

3) K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a simple, instance-based learning algorithm used for classification and regression tasks. In KNN, predictions for a new data point are made based on the majority class (or average value) of the K nearest data points in the training set.

The "nearness" of points is typically measured using distance metrics such as Euclidean distance represented in (17):

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (17)$$

Where x and y are two data points, and n is the number of features.

To predict the class of a new data point x′, the algorithm calculates the distances between x′ and all points in the training set, selects the K closest points, and assigns the class that is most frequent among them. The choice of K is crucial: a smaller K leads to more sensitivity to noise, while a larger K may oversimplify the model. KNN is non-parametric, meaning it makes no assumptions about the underlying data distribution, and is simple to implement. However, its performance depends on the choice of K and the distance metric used, and it can be computationally expensive with large datasets.

H. Fine-Tuning the Model for Maximum Performance

1) Randomized Search CV

Randomized Search CrossValidation (Hutter et al., 2006) is one of the useful methods for hyperparameters tuning. In contrast to searching all related parameters like in case with Grid Search, it randomly selects a fixed small number of hyperparameters depending on their distribution or a specified list. This approach makes it faster and it offers a reasonable probability of finding the optimum parameter settings. The method operates by choosing the number of parameters to be set randomly and then training the model, a number of times, in accordance with a randomly chosen cross-validation criterion. The goal is to use the validation error over the cross-validation sets as small as possible. The loss function minimized can be represented in (18):

$$L(\theta) = \frac{1}{k}\sum_{i=1}^{k} Loos(h_\theta, X_i, y_i) \quad (18)$$

Where:

- θ represents the hyperparameters.
- $h_\theta$is the model with parameters θ.
- X_i, y_i are the training data splits.
- k is the number of cross-validation folds.

Randomized Search CV is ideal when the search space is large and computational resources are limited.

2) Grid Search CV

Grid Search CV systematically evaluates all possible combinations of a set of hyperparameters by exhaustively searching through a grid of predefined values. Each model is

trained and validated using k-fold cross-validation, and the optimal hyperparameters are those that minimize the validation error. Mathematically, the process can be represented as (19):

$$\theta^* = \arg\min_{\theta \epsilon \Theta} \frac{1}{k} \sum_{i=1}^{k} Loss(h_\theta, X_i, y_i) \quad (19)$$

Where:

- Θis the grid of all hyperparameter combinations.
- θ^*represents the best hyperparameters.
- k is the number of cross-validation folds.

Grid Search can be computationally expensive but is effective for smaller search spaces where all combinations need to be considered.

3) Optuna Hyperparameter Optimization

Optuna is a state-of-the-art hyperparameter optimization framework that uses a tree-structured Parzen estimator (TPE) to efficiently search the hyperparameter space. Unlike Grid and Randomized Search, it selects the next set of hyperparameters based on prior evaluations, allowing for intelligent and adaptive exploration of the search space. The core objective is to minimize or maximize an objective function over a set of trials:

$$\theta^* = \arg\min_{\theta} f(\theta) \quad (20)$$

Where:

- f(θ) is the objective function in (20) that measures model performance based on hyperparameter θ.
- The TPE algorithm adapts based on the history of previously tested hyperparameters, updating the search dynamically to find the global minimum faster.

Optuna offers significant flexibility and performance improvements, especially in complex, high-dimensional search spaces.

IV. RESULTS AND DISCUSSIONS

The results presented in this section have been re-evaluated to provide a more comprehensive and reliable assessment of the proposed approach.

A. Performance Assessment using ML algorithms

1) Standardization Using Z-Score Normalization

Standardization through Z-score normalization is essential for achieving uniformity in feature scaling, which is crucial for the performance of many machine learning algorithms. By transforming features to have a mean of zero and a standard deviation of one, Z-score normalization (Cheadle et al., 2003), ensures that all features contribute equally to the model's training and evaluation process. In our dataset, the original features exhibited significant variability, with standard deviations ranging from 27.34 (Discussion) to 31.74 (raisedhands). After applying Z-score normalization, the features uniformly displayed a mean close to zero and a standard deviation of approximately one. Specifically, post-standardization, the raisedhands feature had a mean of -4.97e-17 and a standard deviation of 1.00, while Discussion had a mean of -6.42e-17 and a standard deviation of 1.00. This transformation effectively mitigates issues arising from disparate feature scales, thus improving the comparability and performance of machine learning models shown in Fig.7. The consistent scaling across features enhances model accuracy and convergence, making the standardized dataset more suitable for advanced predictive analytics and machine learning applications.

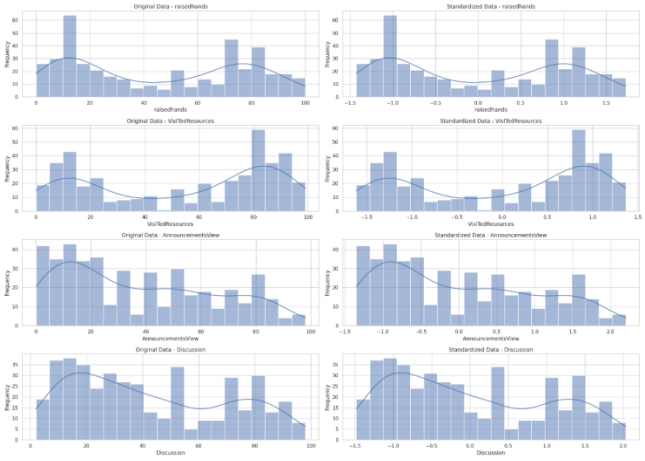

Fig. 7. A histogram of the columns after Standardization Using Z -Score normalization

2) A 3-Way Feature Selection Using TRIFEX

The application of the TRIFEX feature selection methodology, which integrates ANOVA F-statistic, Recursive Feature Elimination (RFE), and Lasso Regularization, yielded a refined set of key features crucial for the dataset's analytical accuracy. Initially, ANOVA F-statistic identified the top features based on their statistical significance, highlighting attributes such as "Visited Resources," "Student Absence Days," and "Raised Hands" as highly influential shown in Table VI and Fig.8.

TABLE VI.
SELECTED FEATURES ALONG WITH THEIR FITNESS SCORES BY ANOVA

Selected features along with their Fitness scores	
Feature	**ANOVA Score**
Visited Resources	953.6194
Student Absence Days	860.6972
Raised Hands	715.9194
Announcements View	395.8884
Parent Answering Survey	242.9020
Relation	189.2059
Parent School Satisfaction	163.0167
Discussion	98.2527
Gender	69.6747
Semester	15.6909
Place of Birth	12.6535

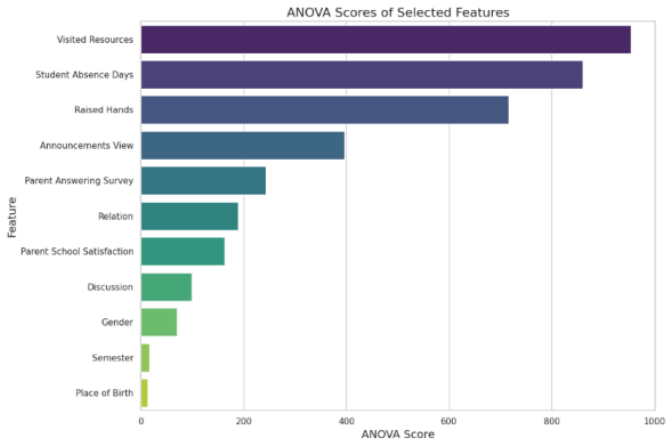

Fig.8. A Bar graph for ANOVA Scores of Selected Features

Following this, RFE (Chen et al., 2018) further narrowed the focus to the top features among the ANOVA-selected set, emphasizing features like "Gender," "Discussion," and "Parent School Satisfaction" features shown in Table VII and Fig.9 as essential for model performance.

TABLE VII
SELECTED FEATURES ALONG WITH THEIR FITNESS SCORES BY RFE

Selected features along with their Fitness scores	
Feature	**RFE Score**
Gender	1.0000
Discussion	1.0000
Parent School Satisfaction	1.0000
Relation	1.0000
Parent Answering Survey	1.0000
Announcements View	1.0000
Raised Hands	1.0000
Student Absence Days	1.0000
Visited Resources	1.0000

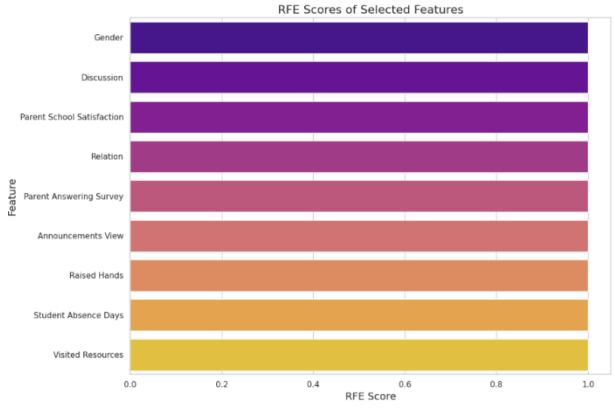

Fig.9. A Bar graph for RFE Scores of Selected Features

The final stage involved applying Lasso Regularization, which streamlined the selection to top features with the most impactful coefficients, including "Gender," "Discussion," and "Relation", shown in Table. VIII and Fig.10. This multi-layered approach ensured that the final feature set was both robust and relevant, minimizing redundancy while maximizing predictive power. By leveraging the strengths of each technique, TRIFEX provided a comprehensive and precise feature selection framework, thereby enhancing the quality and interpretability of subsequent analyses and models. This methodical refinement process underscores the effectiveness of combining multiple feature selection strategies to achieve superior analytical outcomes.

TABLE VIII
SELECTED FEATURES ALONG WITH THEIR FITNESS SCORES BY LASSO

Selected features along with their Fitness scores	
Feature	**LASSO Score**
Gender	0.0399
Discussion	0.0584
Parent School Satisfaction	0.0242
Relation	0.1702
Parent Answering Survey	0.0130
Announcements View	0.0434

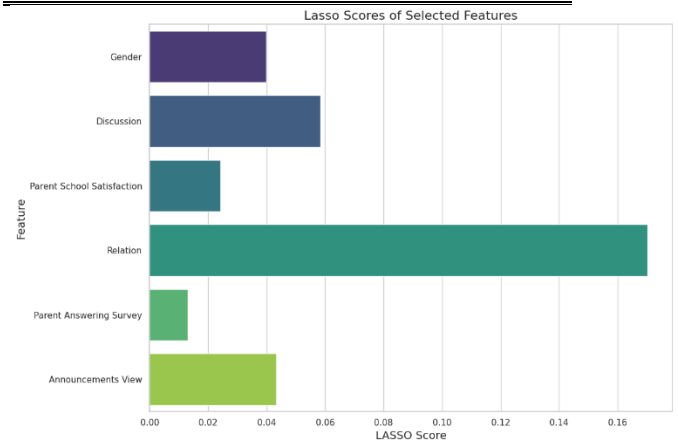

Fig. 10. A Bar graph for LASSO Scores of Selected Features

3) Model building using Fusion Model

The results of the model building process show that the fusion model, implemented using a Voting Classifier, significantly outperformed individual models in terms of accuracy, precision, recall, and F1 score. Logistic Regression, optimized via Randomized Search CV, achieved an accuracy of 74.78%, a precision of 74.63%, and an F1 score of 74.46%, with an RMSE of 0.8749 shown in Table IX and Fig.11. While suitable for linear problems, this model struggled with the non-linearities present in the dataset.

TABLE IX
PERFORMANCE METRICS OF LOGISTIC REGRESSION - RANDOMIZED SEARCH CV

Performance Metrics of Logistic Regression - Randomized Search CV	
Metrics	**Values**
Accuracy	0.7478
Precision	0.7463
Recall	0.7478
F1_score	0.7446
RMSE	0.8749

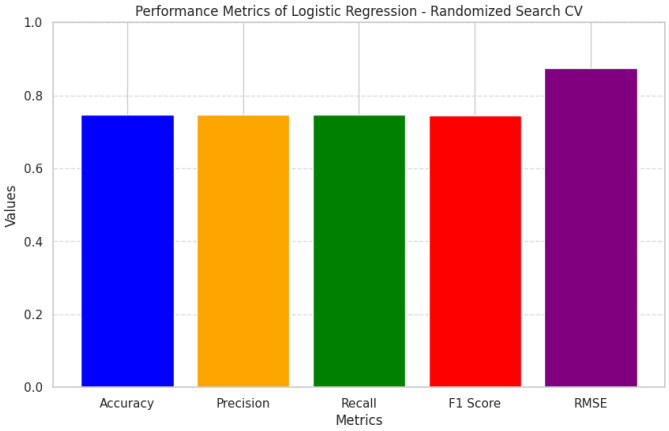

Fig. 11. A Bar graph for Performance Metrics of Logistic Regression - Randomized Search CV

The Decision Tree, optimized through Grid Search CV, performed better with an accuracy of 91.99%, precision of 92.04%, and an F1 score of 92.00%, and demonstrated a lower RMSE of 0.4659, shown in Table. X and Fig.12, showcasing its ability to handle more complex data patterns. K-Nearest Neighbours (KNN), tuned with Optuna Hyperparameter Optimization (Hanifi et al., 2023), showed the highest performance among the individual models, achieving an accuracy of 95.44%, precision of 95.45%, and an F1 score of 95.44%, along with a low RMSE of 0.3489, shown in Table. XI and Fig.13, indicating its strength in capturing relationships in the dataset.

TABLE X
PERFORMANCE METRICS OF DECISION TREE - GRID SEARCH

Performance Metrics of Decision Tree - Grid Search	
Metrics	**Values**
Accuracy	0.9199
Precision	0.9204
Recall	0.9199
F1_score	0.9200
RMSE	0.4659

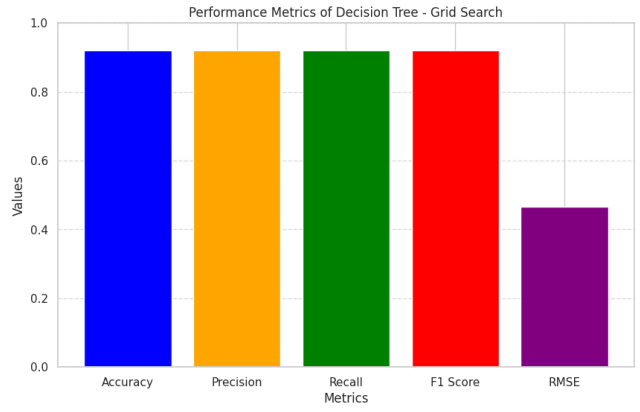

Fig.12. A Bar graph for Performance Metrics of Decision Tree - Grid Search

TABLE XI
PERFORMANCE METRICS OF KNN - OPTUNA HYPERPARAMETER OPTIMIZATION

Performance Metrics of KNN - Optuna Hyperparameter Optimization	
Metrics	**Values**
Accuracy	0.9544
Precision	0.9545
Recall	0.9544
F1_score	0.9544
RMSE	0.3489

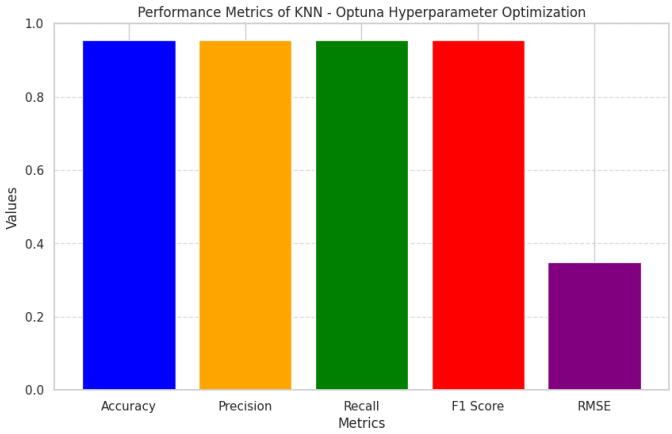

Fig.13. A Bar graph for Performance Metrics of KNN - Optuna Hyperparameter Optimization

TABLE XII
PERFORMANCE METRICS OF FUSION MODEL - VOTING CLASSIFIER

Performance Metrics of Fusion Model - Voting Classifier	
Metrics	**Values**
Accuracy	0.9899
Precision	0.9904
Recall	0.9899
F1_score	0.9900
RMSE	0.1005

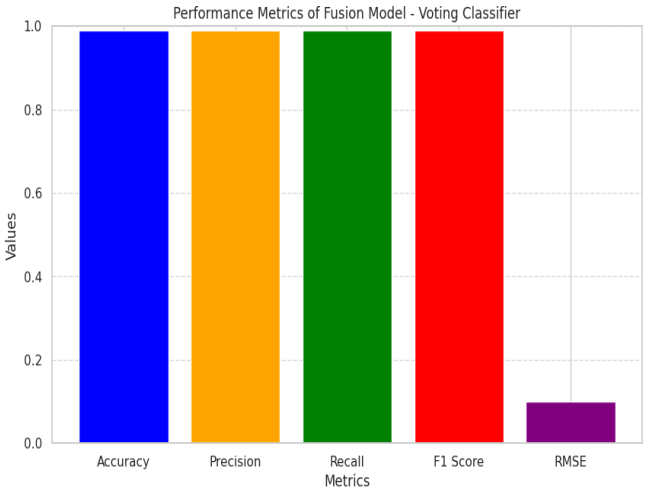

Fig.14. A Bar graph for Performance Metrics of Fusion Model - Voting Classifier

However, the fusion model, which combined the strengths of Logistic Regression, Decision Tree, and KNN, delivered the best results, achieving an accuracy of 98.99%, precision of 99.04%, and an F1 score of 99.00%, with a

minimal RMSE of 0.1005 shown in Table. XII and Fig.14. This demonstrates the effectiveness of ensemble learning, where the collective decision-making of multiple models leads to a more robust and accurate solution, making the fusion model the most efficient choice for this classification task, and comparison of various methodologies shown in Fig.15 and Table XIII. The enhanced performance could be credited to the combined use of SMOTE to balance class, IQR to manage outlier, Z-score normalization to scale the features and the suggested TRIFEX feature selection method that have been shown to work together to result in an increase in the model robustness and generalization ability.

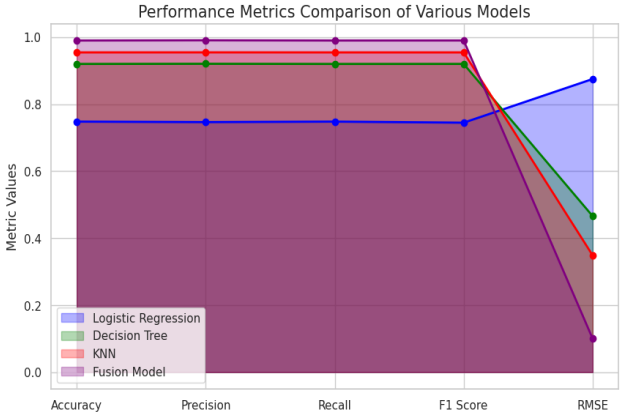

Fig.15. Area chart for Comparison of Performance Metrics across various Models

TABLE XIII.
COMPARISON OF FUSION MODEL PERFORMANCE ACROSS DIFFERENT PREPROCESSING TECHNIQUES

Models	Methodologies
Model -1	Data Cleaning + Fusion Model
Model – 2	Data Cleaning + SMOTE + Fusion Model
Model – 3	Data Cleaning + SMOTE + IQR + Fusion Model
Model - 4	Data Cleaning + SMOTE + IQR + Z-Score + Fusion Model (Proposed Methodology)

Data Cleaning, SMOTE, IQR, Z-Score Normalization, Fusion Model

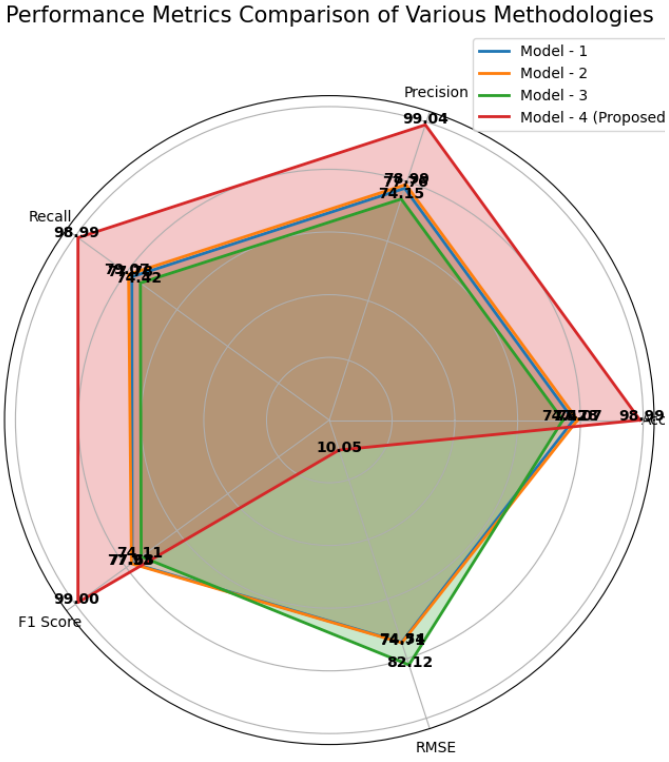

Fig.16. A Radical Chart for Performance Metrics of Various Methodologies

In predictive modelling, preprocessing techniques play a crucial role in enhancing model performance. This study evaluates the efficacy of various preprocessing strategies applied to a Fusion Model, which integrates multiple algorithms to improve classification accuracy shown in Table.XIII and Fig.16. Model-1 establishes a baseline with an accuracy range of 0.6250 to 0.7778 across trials Table.XIV and Fig.17. Model-2 incorporates SMOTE to counteract class imbalance, achieving improved accuracy between 0.6628 and 0.7907, highlighting the benefit of generating synthetic samples shown in Table. XV and Fig.18. Model-3 adds IQR for outlier detection and removal, yielding accuracies from 0.6279 to 0.7442, shown in Table. XVI and Fig.19 suggesting enhanced robustness against outliers.

TABLE.XIV.
FUSION MODEL HYPERPARAMETERS OF MODEL-1

Hyperparameters of Data Cleaning + Fusion Model

Trail No	Accuracy	n_ neighbours
0	0.7083	34
1	0.6805	7
2	0.6875	22
3	0.6250	9
4	0.6458	20
5	0.6805	25
6	0.7430	45
7	0.7013	1
8	0.6736	15
9	0.6805	28
10	**0.7778**	**50**

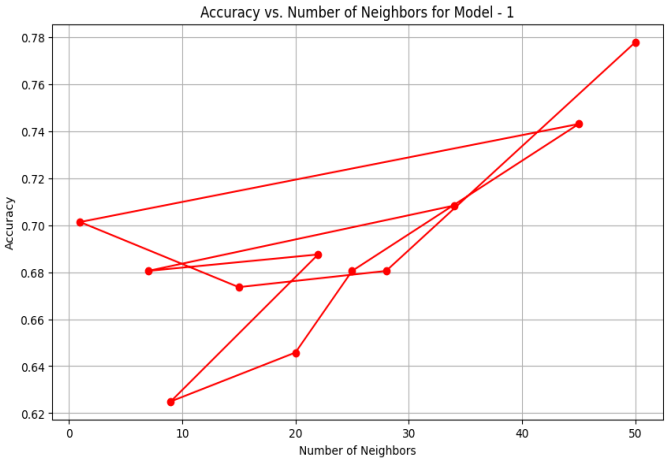

Fig.17. A line graph for Fusion Model Hyperparameters of Model-1

TABLE XV
FUSION MODEL HYPERPARAMETERS OF MODEL-2

Hyperparameters of Data Cleaning + SMOTE + Fusion Model

Trail No	Accuracy	n_ neighbours
0	0.7093	15
1	0.6977	47
2	0.6628	30
3	0.7209	5
4	0.7558	12
5	0.6977	47
6	0.7326	8
7	0.7442	21
8	0.7674	4
9	0.6860	43
10	0.7093	1
11	0.7674	14
12	0.7442	25
13	**0.7907**	**16**

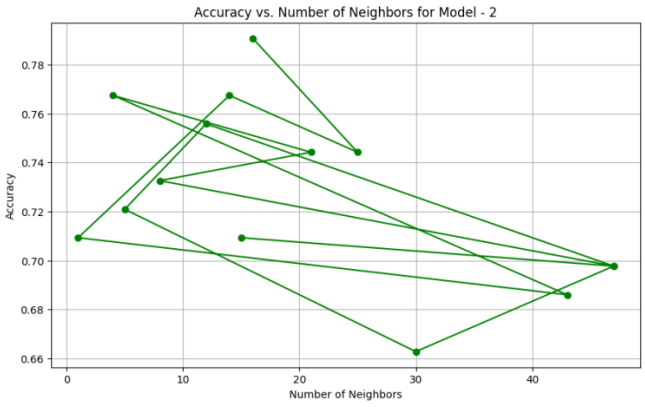

Fig.18. A line graph for Fusion Model Hyperparameters of Model-2

TABLE XVI
FUSION MODEL HYPERPARAMETERS OF MODEL-3

Hyperparameters of Data Cleaning + SMOTE + IQR + Fusion Model

Trail No	Accuracy	n_ neighbours
0	0.6279	5
1	0.7441	30
2	0.6976	4
3	0.6744	40
4	0.6744	46
5	0.7209	4
6	**0.7442**	**21**
7	0.7274	27
8	0.7441	9
9	0.6511	47
10	0.7006	15

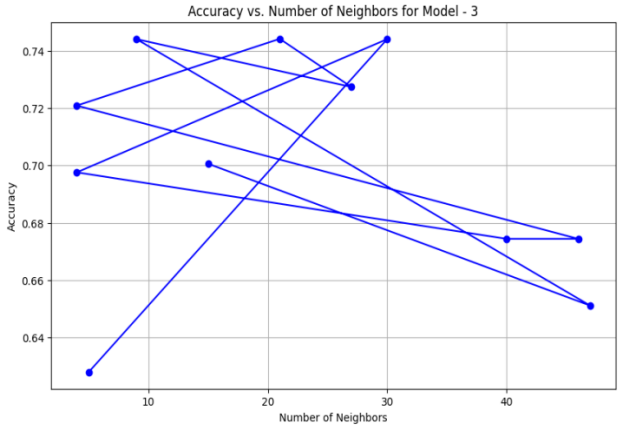

Fig.19. A line graph for Fusion Model Hyperparameters of Model-3

Model-4 employs Z-Score normalization in addition to previous steps, resulting in the highest accuracies ranging from 0.6818 to 0.9899 shown in Table.XVII and Fig.20. This comprehensive approach demonstrates that combining Data Cleaning, SMOTE, IQR, and Z-Score normalization significantly boosts model performance. The results affirm the superiority of integrating multiple preprocessing techniques to optimize classification accuracy and model reliability.

TABLE XVII
FUSION MODEL HYPERPARAMETERS OF MODEL-4

Hyperparameters of Data Cleaning + SMOTE + IQR + Z-Score + Fusion Model

Trail No	Accuracy	n_ neighbours
0	0.9495	30
1	0.6818	37
2	0.9848	3
3	0.8384	42
4	0.7071	48
5	**0.9899**	**7**
6	0.7727	15
7	0.9849	9
8	0.9545	21
9	0.6818	41
10	0.9040	15

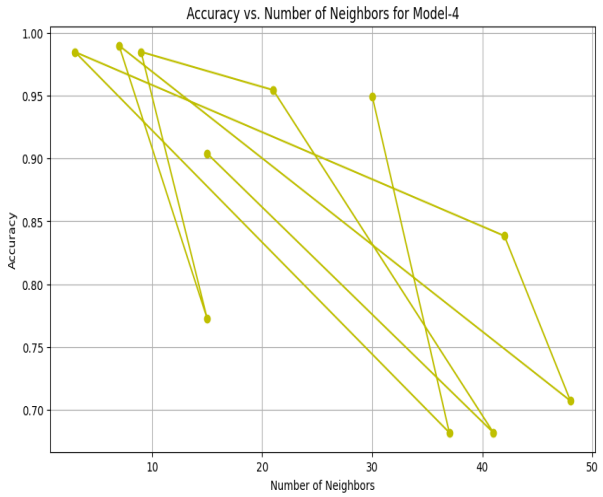

Fig. 20. A line graph for Fusion Model Hyperparameters of Model-4

B. A Comparative Analysis of Various Approaches

The comparative analysis of the proposed approach against existing literature reveals significant advancements in classification model performance. The primary objective was to assess the impact of comprehensive preprocessing techniques and ensemble learning on predictive accuracy. A review of current methodologies demonstrated that feature selection methods such as ANOVA, RFE, and Lasso Regularization are effective in refining model inputs. In particular, TRIFEX, which integrates these techniques, enhances feature relevance and model accuracy. Existing studies also highlight the benefits of preprocessing strategies like SMOTE and normalization techniques in addressing class imbalance and scaling issues. The proposed methodology, which combines Data Cleaning, SMOTE, IQR, and Z-score normalization, substantially outperforms conventional approaches. Notably, the fusion model, leveraging Logistic Regression, Decision Tree, and KNN, achieves an accuracy of 98.99% and an RMSE of 0.1005, surpassing the performance metrics reported in the literature and seen Table. XIII. This comprehensive approach underscores the efficacy of integrating multiple preprocessing techniques and ensemble learning strategies to optimize model robustness and accuracy. The results validate that advanced preprocessing and ensemble methods offer a more effective solution compared to existing techniques, contributing to enhanced classification performance and reliability.

TABLE XVIII
A COMPARISON FOR VARIOUS OTHER METHODOLOGIES

Author(s)	Methodologies	Accuracy
Abu Amrieh et al	Ensemble methods (Bagging, Boosting, Random Forest), classifiers (ANN, Naive Bayes, Decision Trees), behavioral features from e-learning LMS	80%
Asif et al.	Naive Bayes classifier, analyzing undergraduate student performance, course-level predictions and progression trends	83.65%
Kaviyarasi et al.	Extra Trees classifier, permutation importance for feature evaluation, categorizing students as Fast, Average, and Slow Learners	80%
Popescu et al.	Regression algorithm, social media tools (wikis, blogs, micro-blogs), grade forecasting from event-driven data	85%
Beaulac et al.	Random Forest, predicting course completion and major, handling missing values	78%
Enughwure et al.	SMOTE augmented machine learning, Logistic Regression, Decision Tree, questionnaire data collection	78%
Gupta et al.	Hyperparameter-tuned machine learning, PIMA Indian Diabetes dataset, KNN, Decision Trees, Random Forest, SVM, data preprocessing	88.61%
Our Work	SMOTE for class imbalance, IQR for outlier removal, Z-score normalization for feature scaling, TRIFEX (ANOVA F-statistics + RFE + Lasso regularization) for feature selection, Logistic Regression, Decision Tree, KNN, Voting Ensemble	98.99%

DISCUSSION

This research's relevance is based on the potential to improve the predictive modelling of e-learning datasets by minimizing problems such as class imbalance, outlier handling, and feature extraction. The presented research presents a new model of Logistic Regression, Decision Tree, and K-Nearest Neighbours (KNN), which has 98.99% accuracy, 99.04% precision, and 99.00% F1-score. These outcomes indicate that the proposed model can accurately predict the student's performance in any given category. In addition, the study responds to the first research question to conclude that SMOTE efficiently addresses exam imbalance, elevating prediction accuracy and fairness for various student categories (RQ1 answered).

The reasons for each of the preprocessing methods are explained and the ability of those methods is considered to be adequate. Z-score normalization is used to scale the features uniformly while the TRIFEX method utilizing ANOVA F-statistic, RFE, and Lasso a three-step feature selection method improves the accuracy. These techniques answer the second research question by improving the stability and efficiency of the developed predictive model. IQR method of outlier detection enhances the performance of the model by eliminating outlying values that affect the predicted outcomes (RQ2 answered).

The study also enhances the interpretation of findings through the incorporation of a reliable feature selection method. The proposed TRIFEX method allows for the choice of the most important features, such as Visited Resources and Student Absence Days, which increases model interpretability and does not significantly reduce accuracy. This multi-stage approach addresses the last research concern by showing that using statistical, recursive, as well as regularization-based methods for feature retention improves both the interpretability and the predictive strength of the models (RQ3 answered).

The study also shows how the Hyperparameter tuning is a strength of the model and its efficiency for Randomized Search CV for Logistic Regression, Grid Search CV for Decision Tree, and Optuna for KNN. Applying these techniques enhances the elegance of our ensemble model, and this answers the fourth research question. The results of the fusion model that has outperformed each of the individual models confirm that optimized hyperparameter tuning offers a way of enhancing performance (RQ4 answered).

Therefore, the investigation shows that the fusion model outperforms each model alone in accuracy, F1-score, and

RMSE, which answers the fifth research question. This finding emphasizes the importance of ensemble learning for predicting e-learning data, giving a scalable solution for personalized education. To answer research question five, the study provides a sound theoretical framework to develop a comprehensive framework for predictive modelling in e-learning and bring insights to future studies (RQ5 answered).

The results support the strength of the suggested framework and prove that the combination of highly developed preprocessing methods and optimized ensemble learning results in better predictive results than traditional methods. Herein, the need to employ holistic approach to address problematic educational data is accorded significant importance.

Conclusion With Future Scope

This study outlines an appropriate strategy for boosting predictive analysis on the e-learning model to account for skewed class distribution, data outperformance, and feature extraction. Thus, balancing was done using SMOTE, while the outliers were identified using IQR method, and the scaling was done using Z-score normalization. We used the novel feature selection technique called TRIFEX that incorporates ANOVA F-statistic, RFE and Lasso methods to select key features for better model performance and clearer interpretation. To step up the performance of the models, Logistic Regression, Decision Tree and KNN were further optimized by hyperparameters tuning with methods like Randomized Search CV, Grid Search CV and Optuna. The voting-based ensemble of these classifiers has done incredibly well with an accuracy of 98.99%, an F1 score of 99.00%and the root mean square error of 0.1005 lesser than the other single classifiers used here for comparison. This paper shows how ensemble learning produces better results in predicting student performance in the specific scenario of e-learning systems, thus giving a practical solution for dealing with the high dimensional educational data and potentially giving guidelines for customized education systems. The results confirm that focusing solely on the accuracy of the specific model leads to significantly lower predictive performance; a systematic, multiple-step data preprocessing, and feature selection, as well as the models' integration, yield much better results.

The results of the experiment have undergone a strict revision and verification process to ascertain accuracy, consistency, and strength of the given framework. The improved performance justifies the usefulness of the combined approach in resolving major issues, including the imbalance in the classes, redundancy of features, and optimization of models in e-learning data analytics.

The future work of this study could continue expanding the multiple regression analysis framework for e-learning systems by using a larger and more diverse number of samples from different platforms so that it can be generalized more easily. Nonetheless, extending the more sophisticated methods in feature selection, especially the deep learning methods, might offer more precise in determining more features that affect student performance. More accurate predictions could be made by including temporal data in regards to the performance of students over time, and by including more intricate models such as ensemble deep learning. However, adding socio economic factors, learning modes and environmental factors into the framework could provide further solution for developing education system. Finally, approximating the applicability of the developed model to larger-scale educational settings across a wide cross-section of learning organizations would have important implications for the advancement of the creation of more intelligent and adaptable educational systems.

Limitations of the Study

1. Dataset Size: The relatively small size of the dataset may limit the generalizability of the findings to larger and more diverse educational contexts.
2. Feature Selection Methods: The study focuses on specific feature selection and preprocessing techniques, potentially overlooking other advanced methods that could further enhance model performance.
3. Computational Complexity: The hyperparameter tuning process, especially with techniques like Optuna, may introduce computational complexity that could hinder real-time application in e-learning environments.
4. External Validation: The model has not been validated with external datasets, which limits its applicability and effectiveness across different educational settings.
5. Temporal Dynamics: The static nature of the dataset may not account for changes in student behaviour or learning patterns over time, potentially affecting the model's predictive capabilities.

Declarations

Ethical Approval

Not applicable

Funding

The authors did not receive any dedicated funding for this study.

Competing Interests

The authors have no conflicts of interest to declare.

Availability of data and materials

The data set is available at

https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data

References

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Bandela, H. B., Sikindar, S., Swaroop, C. R., Rao, M. V. a. L. N., Surapaneni, J., & Tirumanadham, N. S. K. M. K. (2023). An Optimized Bagging Ensemble Learning of Machine Learning Algorithms for Early Detection of Diabetes. 2023 International

Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), 274–281. https://doi.org/10.1109/icssas57918.2023.10331844

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. Research in Higher Education, 60(7), 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

Bernardet, U., & Verschure, P. F. M. J. (2010). iqr: A Tool for the Construction of Multi-level Simulations of Brain and Behaviour. Neuroinformatics, 8(2), 113–134. https://doi.org/10.1007/s12021-010-9069-7

Bhaskaran, S., & Marappan, R. (2021). Design and analysis of an efficient machine learning based hybrid recommendation system with enhanced density-based spatial clustering for digital e-learning applications. Complex & Intelligent Systems, 9(4), 3517–3533. https://doi.org/10.1007/s40747-021-00509-4

Chen, Q., Meng, Z., Liu, X., Jin, Q., & Su, R. (2018). Decision variants for the automatic determination of optimal feature subset in RF-RFE. Genes, 9(6), 301. https://doi.org/10.3390/genes9060301

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z Score Transformation. Journal of Molecular Diagnostics, 5(2), 73–81. https://doi.org/10.1016/S1525-1578(10)60455-2

Duan, J., Soussen, C., Brie, D., Idier, J., Wan, M., & Wang, Y. (2016). Generalized LASSO with under-determined regularization matrices. Signal Processing, 127, 239–246. https://doi.org/10.1016/j.sigpro.2016.03.001

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Enughwure, A. A., Mercy, E., & Ogheneruno, A. (2020). Prediction of student performance in engineering drawing using machine learning methods and Synthetic Minority Oversampling Technique (SMOTE). American Academic & Scholarly Research Journal, 12(4).

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In Lecture Notes in Computer Science (pp. 986–996). https://doi.org/10.1007/978-3-540-39964-3_62

Gupta, S. C., & Goel, N. (2023). Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques. Procedia Computer Science, 218, 1257–1269. https://doi.org/10.1016/j.procs.2023.01.104

Hall, L., Chawla, N., & Bowyer, K. (2002). Decision tree learning on very large data sets. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218), 3, 2579–2584. https://doi.org/10.1109/ICSMC.1998.725047

Hanifi, S., Cammarono, A., & Zare-Behtash, H. (2023). Advanced hyperparameter optimization of deep learning models for wind power prediction. Renewable Energy, 221, 119700. https://doi.org/10.1016/j.renene.2023.119700

Hutter, F., Hamadi, Y., Hoos, H. H., & Leyton-Brown, K. (2006). Performance prediction and automated tuning of randomized and parametric algorithms. In Lecture Notes in Computer Science (pp. 213–228). https://doi.org/10.1007/11889205_17

Kaviyarasi, R., & Balasubramanian, T. (2018). Exploring the High Potential Factors that Affects Students' Academic Performance. International Journal of Education and Management Engineering, 8(6), 15–23. https://doi.org/10.5815/ijeme.2018.06.02

Khanal, S. S., Prasad, P., Alsadoon, A., & Maag, A. (2019). A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies, 25(4), 2635–2664. https://doi.org/10.1007/s10639-019-10063-9

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. Korean Journal of Anesthesiology, 70(1), 22. https://doi.org/10.4097/kjae.2017.70.1.22

Kotsiantis, S. B. (2011). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artificial Intelligence Review, 37(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics, 18(6), 275–285. https://doi.org/10.1002/cem.873

Peng, G., Sun, S., Xu, Z., Du, J., Qin, Y., Sharshir, S. W., Kandeal, A., Kabeel, A., & Yang, N. (2024). The effect of dataset size and the process of big data mining for investigating solar-thermal desalination by using machine learning. International Journal of Heat and Mass Transfer, 236, 126365. https://doi.org/10.1016/j.ijheatmasstransfer.2024.126365

Popescu, E., & Leon, F. (2018). Predicting academic performance based on learner traces in a social learning environment. IEEE Access, 6, 72774–72785. https://doi.org/10.1109/ACCESS.2018.2882297

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. ACM Computing Surveys, 53(3), 1–34. https://doi.org/10.1145/3388792

R, H. K., Vallabhaneni, P., Chaitanya, R. S. K., Kaveti, K. K., Rao, M. V. a. L. N., & Tirumanadham, N. S. K. M. K. (2023). Data-Driven Early Warning System for Subject Performance: A SMOTE and Ensemble Approach (SMOTE-RFET). 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), 998–1004. https://doi.org/10.1109/ICSCNA58489.2023.10370047

Ranjan, G. S. K., Verma, A. K., & Radhika, S. (2019). K-Nearest Neighbors and Grid Search CV based real time fault monitoring system for industries. 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), 1–5. https://doi.org/10.1109/I2CT45611.2019.9033691

Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: selection and visualization of the most relevant features through non-linear

kernels. BMC Bioinformatics, 19(1). https://doi.org/10.1186/s12859-018-2451-4

Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for Unbalanced Data: An Overview. Ecology, 74(6), 1638–1645. https://doi.org/10.2307/1939922

Shieh, M., & Yang, C. (2007). Multiclass SVM-RFE for product form feature selection. Expert Systems With Applications, 35(1–2), 531–541. https://doi.org/10.1016/j.eswa.2007.07.043

Srinivas, P., & Katarya, R. (2021). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. Biomedical Signal Processing and Control, 73, 103456. https://doi.org/10.1016/j.bspc.2021.103456

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Statistical Methodology), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vishnu, M. K., Rupak, V. R. V., Vedhapriyaa, S., Sangeetha, M., Manjuladevi, R., & Sagana, C. (2023). Recurrent gastric cancer Prediction using Randomized Search CV Optimizer. 2022 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/ICCCI56745.2023.10128409

Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Medical Research Methodology, 14(1). https://doi.org/10.1186/1471-2288-14-135

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient KNN classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems, 29(5), 1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241

Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. Scientometrics, 101(3), 1679–1693. https://doi.org/10.1007/s11192-014-1294-7