SSRG International Journal of Electronics and Communication Engineering Volume 12 Issue 9, 219-234, September 2025
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I9P119 © 2025 Seventh Sense Research Group®

Original Article

Ensemble Model for Educational Data Mining based on Synthetic Minority Oversampling Technique

R. Manoharan[1], M. Subi Stalin[2], Ganesh Babu Loganathan[3], K.Venkateswaran[4]

[1]Department of Ece, Apollo Engineering College, Chennai, Tamilnadu, India.
[2]Department of ECE, Arignar Anna Institute Of Science And Technology, Chennai, Tamilnadu, India.
[3]Department of Robotics & Automation, Raja Lakshmi Engineering College, Chennai, Tamilnadu, India.
[4]Department of Information Technology, St.Joseph College Of Engineering, Chennai, Tamilnadu, India.

[1]Corresponding Author : professormanoharan@gmail.com

Received: 18 July 2025 Revised: 20 August 2025 Accepted: 19 September 2025 Published: 29 September 2025

Abstract *- Educational Data Mining (EDM) is a growing field that applies data mining, statistical analysis, and machine learning techniques to analyze student-related data. Existing EDM approaches often rely on manual statistical methods, which are time-consuming and less adaptable to dynamic educational environments. This paper proposes a novel ensemble-based framework that integrates machine learning classifiers with statistical approaches for student performance classification to address these limitations. To improve predictive accuracy, the model combines multiple classifiers, including Decision Tree, Logistic Regression, Random Forest, Multilayer Perceptron, and K-Nearest Neighbor. Given the inherent class imbalance in educational data, the Synthetic Minority Oversampling Technique (SMOTE) balances the dataset and enhances classifier performance. The proposed model is evaluated using a real-world dataset comprising 6,807 student records collected from a technological college in India. Performance is assessed using eight evaluation metrics to identify the most effective configuration. Results demonstrate the model's capability to deliver accurate and fair classification, aiding data-driven educational decision-making.*

Keywords *- Educational Data Mining, Ensemble method, SMOTE, Machine Learning.*

1. Introduction

In recent years, a few fields have prompted an enormous amount of information to be gathered. Since examining the consider-capable measure of information to arrive at valuable data is a dull task for mankind, information mining procedures can be utilized to find important and critical information from the data. The educational system data mining is dramatically improved, and the educational data mining methods classify the huge data collection in the educational institutes. This mining method includes the knowledge discovery and machine learning algorithms to classify and group the results according to user needs.

This educational data mining helps to understand the learning process of the educational institutions. The proposed methodology helps to predict the student's academic performance based on the recent academic progression of the individual student. And also analyse the affecting factors of the student's academic performance. In recent research, authors may introduce new strategies for mining the educational data collection.

With the fast advancement of organization data innovation and the wide utilization of mobile phones, tablet PCs and other portable terminals, online education plays an increasingly significant part in social life. The increased demand for educational data mining has a huge impact on the construction of education data prediction graphs. K-Means is an unsupervised clustering method that can be used to analyse data to identify patterns. It provides a three-dimensional view of the data sets' observations. Here, the k-means algorithm is used to cluster similar data into one group and differentiate neighboring notes into their own groups. This method is an iterative method to process the data into clusters. The K-means algorithm follows the given steps to process the data.

To a large extent, current supervised classification algorithms rely on conventional measures, which can yield optimal results when the test size is finite. Nonetheless, just limited examples can be procured, practically speaking. From these studies, we use Support Vector Machine (SVM) as a learning strategy towards a model from the factual information and applied in various real-time applications, such as heart data monitoring, satellite data processing, etc. Times, records, and sampling methods really are not issues with the SVM approach. Hence, the above claimed technique is combined with the universally benchmarked model called the Synthetic Minority Oversampling Technique (SMOTE) for the purpose of redressing the existing social divide. Further, the performance evaluation is also statistically performed. Therefore, in order to demonstrate the efficacy of the proposed techniques.

The data collection is based on three services running in an Indian state, Uttar Pradesh, through a recognised institution. The dataset comprises both a student's educational and non-academic (demographic) information. Twenty characteristics are present in the 6807 specimens that make up the data set.

The class parameter used is "Admission Status," which indicates whether a pupil has finished the course or withdrew before finishing. The Waikato Environment for Knowledge Analysis (WEKA) is used to replicate the surroundings in which the theories are created.

The experimental results that demonstrate the importance of non-academic features towards developing a predictive framework for student achievement follow the presentation of the database and its academic and non-academic parameters in the study. The findings, employing all variables as well as just academic parameters, are displayed in tabulated form and graphically analyzed. The study's findings suggest that non-academic factors are essential in pupils' initial performance prediction.

1.1. Problem Statement

In the domain of Educational Data Mining (EDM), analyzing student performance and learning outcomes is essential for improving educational strategies and personalized learning interventions. However, existing statistical techniques often require manual preprocessing and offer limited adaptability to dynamic, real-world educational datasets. Moreover, a significant challenge lies in the presence of class imbalance within academic datasets, where high-performing or underperforming student categories are often underrepresented.

This imbalance can skew predictive modeling results, leading to biased or inaccurate classifications. Existing machine learning models also struggle to maintain performance consistency across imbalanced data. Therefore, there is a pressing need for an automated, intelligent EDM framework that not only leverages the predictive power of ensemble machine learning methods but also incorporates mechanisms to address data imbalance.

To this end, this research aims to develop a hybrid ensemble model enhanced with the Synthetic Minority Oversampling Technique (SMOTE), capable of accurately classifying student performance while ensuring robustness, fairness, and generalizability across imbalanced educational datasets.

1.2. Motivation

The increasing availability of educational data presents a unique opportunity to apply machine learning techniques for gaining actionable insights into student performance, dropout risks, and learning behavior. However, one major obstacle in effectively leveraging such data is the inherent imbalance in student datasets, where certain performance categories, such as at-risk or high-achieving students, are underrepresented. This imbalance reduces the predictive power of existing models and leads to biased outcomes, especially when using conventional statistical approaches that lack adaptability.

Furthermore, most educational institutions lack the tools to automate data analysis, relying instead on manual interpretation, which is time-consuming and error-prone. This research is motivated by the need to overcome these limitations by introducing a smart, ensemble-based classification framework that combines the strengths of multiple machine learning algorithms with SMOTE to ensure fair representation of all student groups. The goal is to support educational institutions in making informed, data-driven decisions that enhance student success and retention.

1.3. Research Gap

While Educational Data Mining (EDM) has gained momentum in recent years, most existing approaches rely heavily on conventional statistical methods or single-machine learning classifiers, which often fail to generalize across complex, real-world student datasets. These existing models typically lack the capability to handle imbalanced class distributions, leading to biased predictions—particularly for minority student groups such as low-performing or at-risk individuals.

Moreover, very few studies have explored the integration of ensemble machine learning techniques with oversampling methods like SMOTE to improve classification accuracy and fairness in EDM contexts. Although some literature acknowledges class imbalance, comprehensive frameworks that combine multiple classifiers with automated data balancing strategies are limited. Additionally, there is a lack of evaluation using real-world educational data from diverse student populations in developing regions such as India. This gap highlights the need for an intelligent, scalable, and data-balanced EDM model that ensures more equitable and accurate student performance prediction.

2. Literature Survey

Sapkota *et al.* 2019 proposed the spectral clustering approach to summarize the data set using cluster and classification algorithms. The authors use the k-means algorithm to classify the data set and form cluster to reduce the time. From this approach, the clustering error of the dataset is reduced.

Li, H, Lu, Q., *et al.* proposed the SVM classification optimization mechanism to improve the forecast accuracy of the system. The authors use a K-CV parameter optimization model to improve the SVM classification accuracy. S. Chandra and M. Kaur *et al*. (2015) proposed that the classification mechanism enhances the accuracy of the classification algorithm. The authors developed a model specifically used for medical data classification and monitored the accuracy of the system. Okfalisa *et al*. 2017 proposed the comparative analysis of the classification algorithms.

The authors comparatively take the two classification algorithms for testing. They use K-Nearest and the modified K-Nearest Neighbor Algorithm. The results of the modified k-nearest neighbor algorithm produce better results in terms of accuracy than the KNN. Pristyanto et al 2018 proposed the balance distribution model to classify the education-related data set. They use the OSS, SMOTE method to balancing the data set from the raw data. This classification improves the accuracy of the balancing of SVM classifications.

Basarslan, M. S., &Argun, I. D, *et al*. 2018 fought for a classification model for bank data classification. The authors use the UCI machine learning approach to classify the large data set with high accuracy. This model inherits the native bayes, KNN and decision algorithm behavior to classify the datasets. Baralis *et al*. (2008) proposed the lazy model to improve the accuracy of the associative classification method. The authors use the SVM and decision tree algorithm for the lazy model to improve the system's classification accuracy.

Erol, H et al. 2018 proposed the classification methodlogies in data mining techniques. The authors use the information mining clustering conclusions to classify the images and information that are remotely sensed. This model clusters the sensed image into 6 parts. Karamouzis and Vrettos *et al*. (2008) proposed an ANN system that predicts the student's achievement based on the student's profile. They trained the algorithm to cross-validate the classified data set of the student profile.

Li *et al*. 2019 proposed the medical data stream distribution pattern for the rule mining algorithm. The authors use the existing data mining methods to classify the medical data sets in an association rule manner. They also use a density estimation method to predict the accuracy of the data set.

Antonio *et al*. 2020 proposed a model to predict the decision of the student enrolment in a subject-wise manner. They use the advanced data mining method to predict the growth of students in academic success rates. This model also uses the decision tree and support vector system to reduce the dropout rate of the data.

The knowledge system to mine student information in academic institutions was proposed by Penghe Chen et al. in 2009. They automated information well-planned using the knowledge base training method. Those algorithms raise the accuracy of the information set.

Yu et al. (2019) proposed an Extreme Learning Machine (ELM) categorization model-based "Active Online-Weighted ELM (AOW-ELM)" as a successful execution. In addition to feature selection strategies, Aggarwal et al. (2019) compared research on other machine learning approaches. The author concentrated on the variation and association criteria used in selecting features.

The best potential classification algorithms to anticipate students' achievement, according to Aggarwal et al. (2019), were MLP as well as Random Forest. They were used in a test on student information with educational and non-academic characteristics. In order to create a software defect prediction model, Panda (2019) created a hybrid categorization technique that combines distributed basis balance-based instance selection with a radial basis function neural network classifier.

The forecasting model was developed using technology measurements along with readily viewable historical software reliability statistics gathered from many organizations. Based on the complex's assessment from 2012 to 2015, Abdollahi & Ebrahimi (2019) forecast how such an Iranian theatre complex would behave in 2022. Those who also offered a few insights into issues and the application of skills.

Educational Data Mining (EDM) has emerged as a vital discipline aimed at improving academic outcomes through the analysis of student data. Existing EDM techniques have relied heavily on statistical models such as linear Regression and logistic Regression to predict student performance and identify at-risk learners. While effective in controlled settings, these methods often fall short in dynamic environments due to their limited ability to handle high-dimensional, nonlinear data.

Recent studies have introduced Machine Learning (ML) models to address these limitations. For instance, it demonstrated the effectiveness of decision trees and support vector machines in predicting academic failure while exploring ensemble methods like random forests and boosting techniques, showing improved accuracy over standalone models. However, these studies often assume balanced datasets, overlooking real-world challenges like class imbalance and data sparsity.

To overcome such limitations, hybrid models and data balancing strategies have gained attention. The Synthetic Minority Oversampling Technique (SMOTE) has been widely adopted to mitigate class imbalance in educational datasets. Proposed a SMOTE-enhanced neural network that significantly improved the recall for underrepresented student categories. Implemented a KNN-SMOTE hybrid model for dropout prediction, achieving higher sensitivity in detecting minority class patterns.

Despite these advancements, limited research has been done on integrating multiple ML classifiers with SMOTE in a unified ensemble framework tailored to student performance classification.

Furthermore, studies focusing on real-world, diverse datasets from underrepresented educational contexts—particularly in developing regions like India—remain scarce. This gap highlights the need for more inclusive and adaptable EDM models that balance accuracy, fairness, and scalability.

3. Proposed Ensemble Model

3.1. Decision Tree

A decision tree is a root node-based drop-down approach. It travels the child nodes with the help of the conditions in the tree splits. We use this algorithm for learning decisions from a dataset and produce the classification tree model for further processing. This also monitors the background process of each feature in the data set.

3.2. Logistic Regression

Logistic Regression is a statistical design to analyse the deviations of a sampling dataset. This model is a regression method to process the data. Regression results in the data set in Boolean values such as 0 and 1with an overall sum of 1.

3.3. KNN

K-nearest neighbour is a type of unsupervised clustering method that can be used to identify patterns in the data. It provides a three-dimensional view of the data sets' observations. These groups tend to cluster algorithms start by differentiating neighbor cluster points from similar points and then iterate until convergence.

The table format is taken as an iterative approach to the KNN clustering algorithm. In the proposed model, KNN is used for data constraint cross-validation.

3.4. RF

Random forest is an ensembling learning model to classify the Regression and other constructed operations based on the decision tree approval flows. This model predicts the data set's mean and average rate of the classification samples.

3.5. MLP

Multi-layer perception is a model of an artificial neural network. This model refers to the Multiple-Layer Perceptrons in the processing terminology. We use this model for the training propagation for the system model.

3.6. MOTE-ENN

SMOTE-ENN is a connection model that improves the accuracy of the results. This model connects the SMOTE and ENN models, two different models of over- and undersampling methodology. This model is based on the hybrid sampling approach.

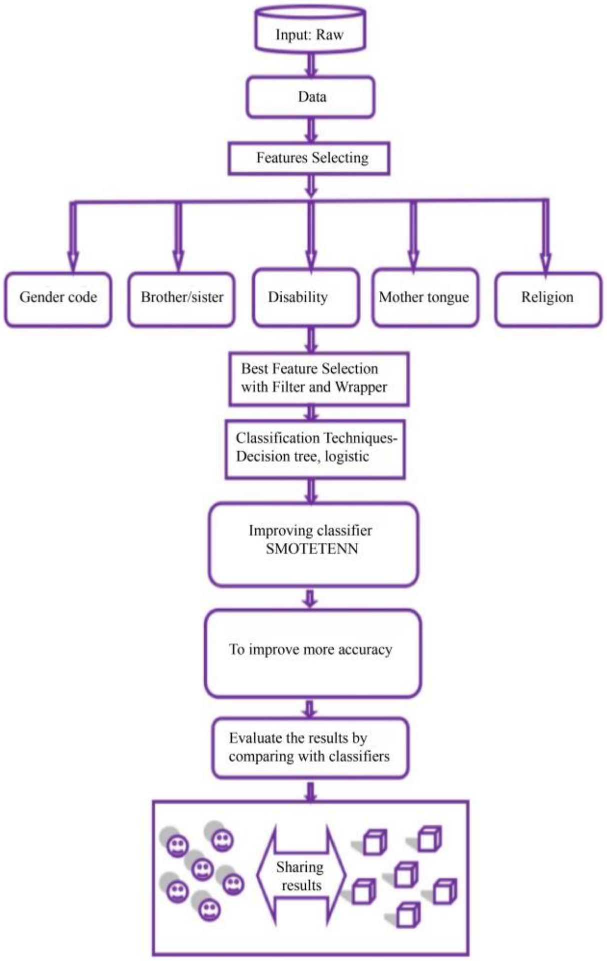

Input: Raw
Data
Features Selecting
Gender code
Brother/sister
Disability
Mother tongue
Religion
Best Feature Selection with Filter and Wrapper
Classification Techniques-Decision tree, logistic
Improving classifier SMOTETENN
To improve more accuracy
Evaluate the results by comparing with classifiers
Sharing results

Fig. 1 Steps involved in the Proposed System

The proposed methodology contains seven stages, which are shown step by step in Figure 1.

3.7. SMOTE

Smote is a statistical technique that increases the number of minority samples in the data set by generating new objects. These methods also reduce the duplication of the minority sample in the existing data set. Here, we use this method in combination with the nearest neighbour algorithm to find the target class for every neighbouring node.

3.8. SMOTE-Tomek

This is a commonly used hybrid method in data classification. We use this method to oversample the data sets. SMOTE –Tomek is connected with SMOTE to produce the enhanced sampling result.

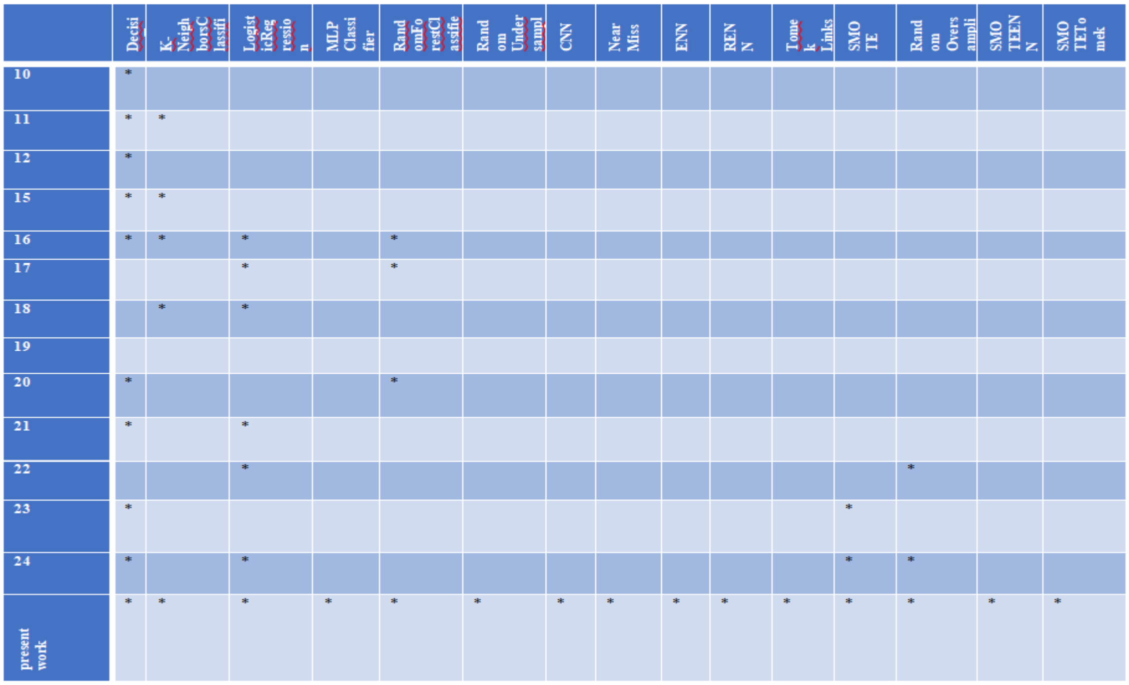
Feature scaling is a method that is utilized to standardize the scope of the free factors. Usually, machine learning models utilize the Euclidean distance calculation to ascertain the distance between two points. Without the scaling feature, Euclidean calculation does not work. To empower scaling features, regularly utilized techniques are normalization, mean normalization, unit vector and min-max scaling. To ascertain the understudy execution, go and rescale the component of the dataset using normalization strategy. Accordingly, all the highlights have the standard typical dissemination attributes with μ = 0 as well as σ = 1, where μ is the mean, and σ is the root-mean-square deviation from the normal. The equation used to determine the qualities is as follows.

In this paper, various machine learning classification algorithms were used to measure the deviation in each algorithm. It incorporates Random Forest, Artificial Neural Network, XGBoost, K-Nearest Neighbor, SVM, naiveBayes, logistic Regression, and the Decision Tree algorithm. Previously mentioned algorithms are grounded classifiers in a multiclass order. It incorporates the K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Logistic Regression (LR), and Random Forest (RF). In certain situations, vector and XG-help calculations do not uphold the multiclass order. Apply one-versus-one technique to help vector machine and one-to-many strategies for X the G-support algorithm.

3.9. Cross Validation

Cross-validation is an approval method applied to assess the factual examination results summed up into an autonomous dataset. This paper uses two well-known distinctive cross-validation methods: arbitrary weight and mixed 5-fold cross-validation. It will separate the haphazard data from the training data up to 80% of the original data. The preparation set is utilized with a resampling technique, as noticed evidently, and classes used in testing ought never to be adjusted by any stretch of the imagination. Accordingly, every iteration of resampling strategies is applied to the preparation set while utilizing diverse model approval shown in Table 1.

Table 1. State of the art – Algorithms' performance measure



	Decisi	K-NeighborsClassifi	LogisticRegression	MLP Classifier	RandomForestClassifie	Random Undersampl	CNN	Near Miss	ENN	RENN	Tomek Links	SMOTE	Random Oversampli	SMOTEENN	SMOTETomek
10	*														
11	*	*													
12	*														
15	*	*													
16	*	*	*		*										
17			*		*										
18		*	*												
19															
20	*				*										
21	*		*												
22			*										*		
23	*											*			
24	*		*									*	*		
present work	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

The above table shows the regression result of the data set in Boolean values, such as 0 to 1with the overall sum of 1. The colour denotes the combination of imbalance and the machine learning algorithm's positive rate accuracy of the system.

3.10. Dataset Description

The dataset used in this study was collected from a technological college in India and contains records of 6,807 students, including attributes such as student ID, attendance, internal and external marks, department, and gender. This primary dataset serves as the foundation for performance classification tasks, as shown in Table 2. Due to the inherent class imbalance in the data—where certain student performance categories were underrepresented—a synthetic dataset was generated using the Synthetic Minority Oversampling Technique (SMOTE), increasing the dataset size to approximately 12,000 records while maintaining the original attribute structure. A feature-engineered version of the dataset was created to improve model accuracy and consistency by applying normalization, encoding, and ranking of variables based on importance. These preprocessed datasets were used to train and calibrate various classifiers. Additionally, an evaluation metrics log was maintained to capture eight key performance indicators, including accuracy, precision, recall, and F1-score, allowing a comprehensive assessment of the ensemble model's effectiveness in educational classification.

Table 2. Dataset description

Dataset Name	Source	Size	Attributes	Purpose
Student Performance Dataset	Technological College, India	6,807 records	Student ID, Attendance, Internal Marks, External Marks, Department, Gender	Primary dataset used for performance classification
Synthetic Balanced Dataset	Generated using SMOTE	~12,000 records	Same as above + SMOTE-generated minority class samples	To resolve class imbalance and improve classification accuracy
Feature-Engineered Dataset	Derived from the original dataset with preprocessing	6,807 records	Normalized and encoded values, feature importance-ranked variables	Used for model calibration and comparison of ML algorithms
Evaluation Metrics Log	Output from model training and testing	8 evaluation sets	Accuracy, Precision, Recall, F1-Score, ROC-AUC, Log Loss, etc.	To assess the performance of classifiers and an ensemble model

3.11. Preprocessing Steps

Data Cleaning: Removed duplicate records and entries with missing critical values (e.g., marks, attendance). Handled incomplete data using imputation techniques (mean for numeric values, mode for categorical).

Data Transformation: Normalized numerical attributes (e.g., internal and external marks, attendance) using Min-Max scaling to bring values into the [0, 1] range. Applied label encoding for binary categorical features (e.g., gender) and one-hot encoding for multiclass features (e.g., department).

Feature Selection: Correlation analysis and information gain were used to identify and retain the most relevant features for classification. Eliminated redundant or weakly correlated features to improve model efficiency.

Handling Class Imbalance: Applied Synthetic Minority Oversampling Technique (SMOTE) to balance the distribution of class labels and ensure fair model training.

Data Splitting: Split the dataset into training (70%) and testing (30%) sets while maintaining class distribution using stratified sampling.

3.12. Performance Analysis

Performance evaluation is the essential method for the classifier to differentiate and identify the finest machine learning model. Machine learning algorithms are evaluated in various ways. Here, various evaluation methods are used to predict the precision, accuracy, and sensitivity, as well as the F1-score; also, the factual assessment system is utilized for a more trustworthy and ground-breaking examination and comparison. Investigating and looking at the classifiers' presentation is a critical method. Despite the fact that it is easy to utilize assessment quantifiers, the results might be deluding. In this manner, identifying the finest model or strategy dependent on their capacities is a basic test. Measurable essentialness tests are intended to tackle this issue.

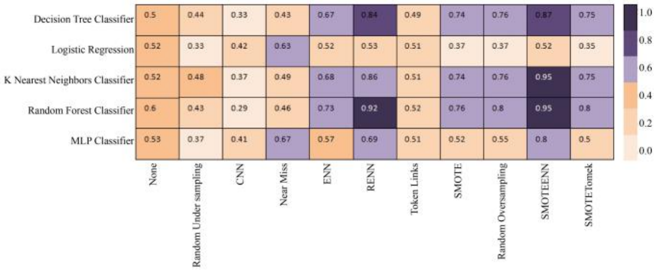
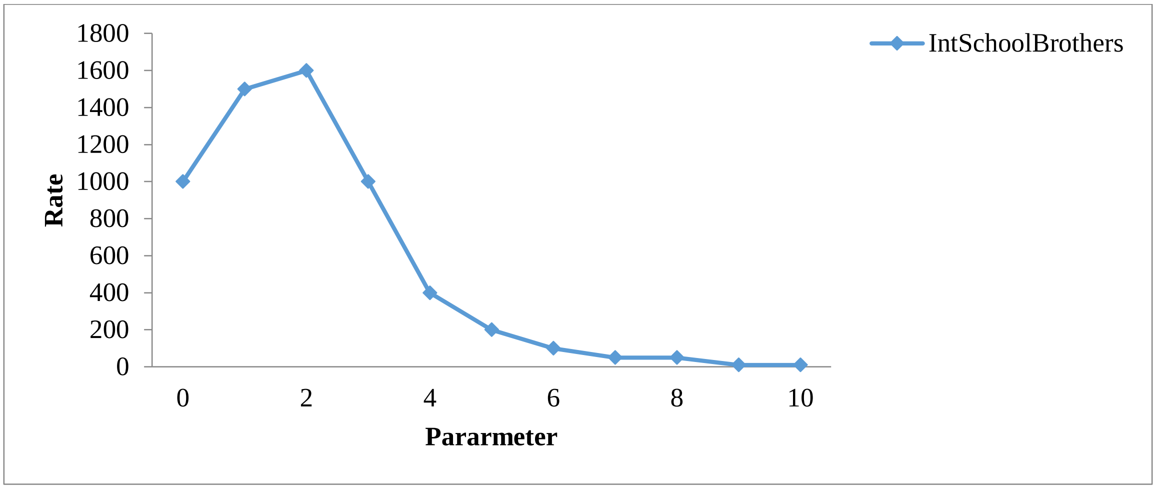
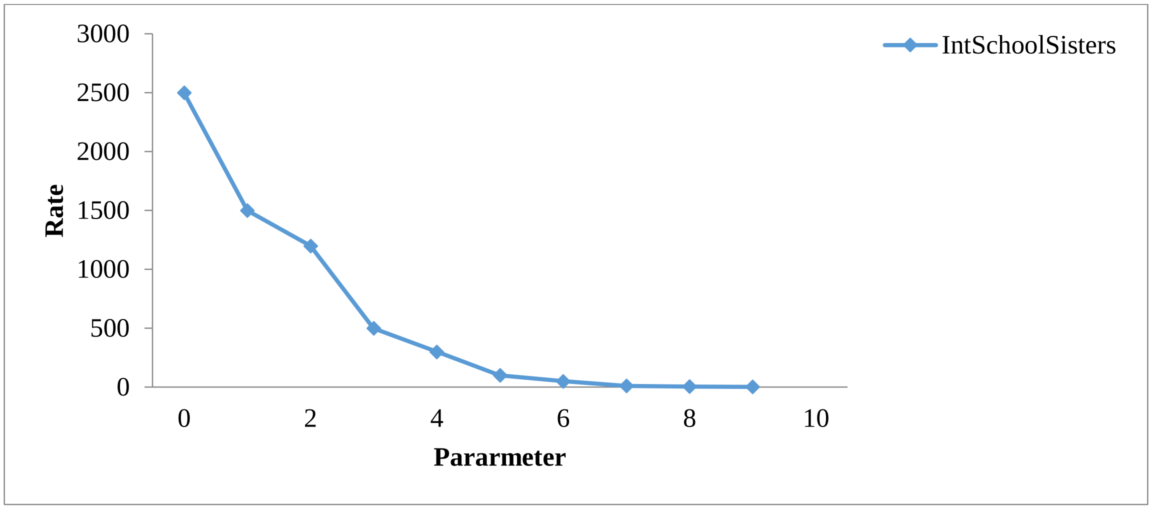


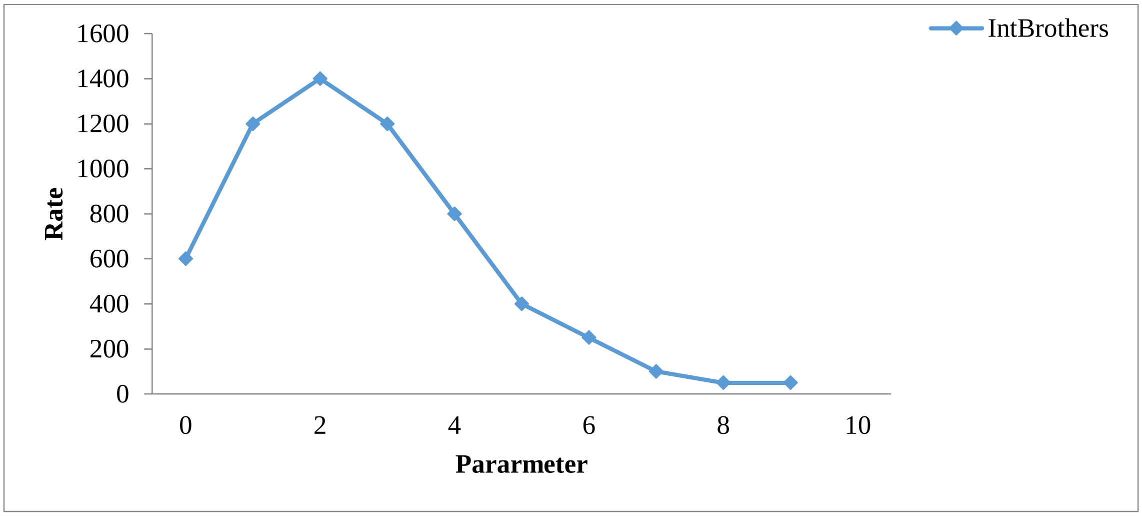
Fig. 2 Confusion matrix

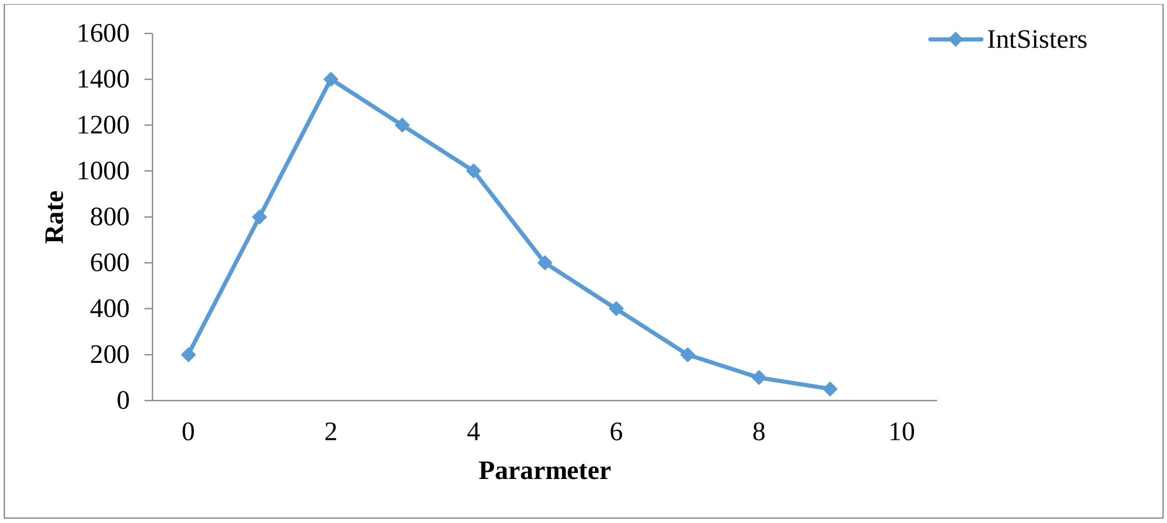
The repeated measures ANOVA is the standard real test procedure used to choose the differentiations between more than two related models. ANOVA strategy resampling the invalid hypothesis from the examiner's data models. The ANOVA test considers three presumptions, as can be seen. The accompanying text explains where these questions come from: First, the models need to be widely adopted, as shown in Figure 3. There should be no interdependence between the model scenarios. Third, it is important that the parties' differences (the techniques being tested) are comparable. In this study, we examine data normality using the Anderson-Darling test. Compared to the Kolmogorov-Smirnov test, this one is different. In light of this, if the p-estimation of this ordinariness is not exactly (α= 0.05), the faulty supposition will be discarded, and the data will not have a typical conveyance. The ANOVA concerns were unfounded. As a multivariate alternative to the ANOVA test, the Friedman test can be used to compare various models and approaches. The

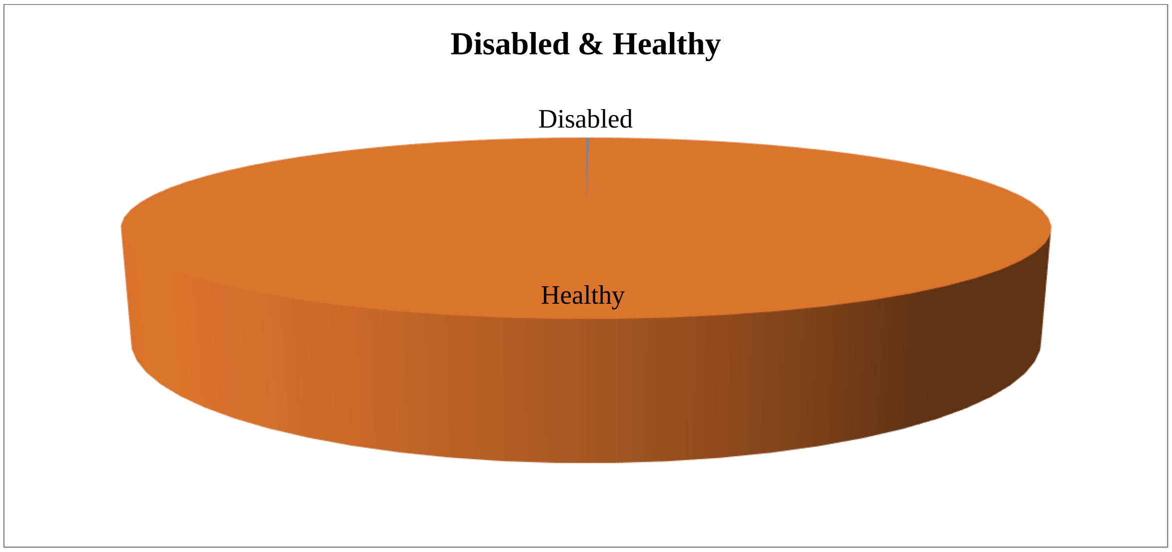
Friedman test is flawed because it assumes that all resampling methods produce the same results. If this assumption is rejected, then it implies resampling approaches produce different results. For each resampling method, this research makes use of the precision data collected using Mix 5-overlay cross-approval. For each resampling method, the Friedman test first estimates where the information came from for each and every classifier. The Friedman test appropriately provides several positions for each resampling strategy that aid in describing the optimal resampling strategy.

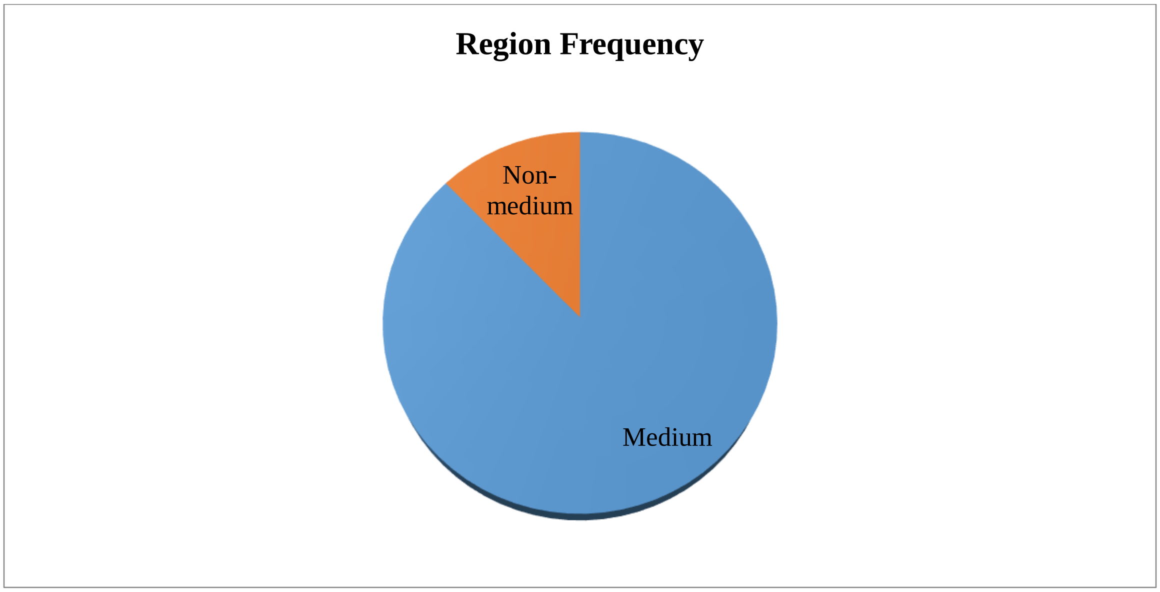







1600
1400
1200
1000
800
600
400
200
0
Rate
0
2
4
6
8
10
Pararmeter
IntSisters


Disabled & Healthy
Disabled
Healthy


Region Frequency
Non-medium
Medium

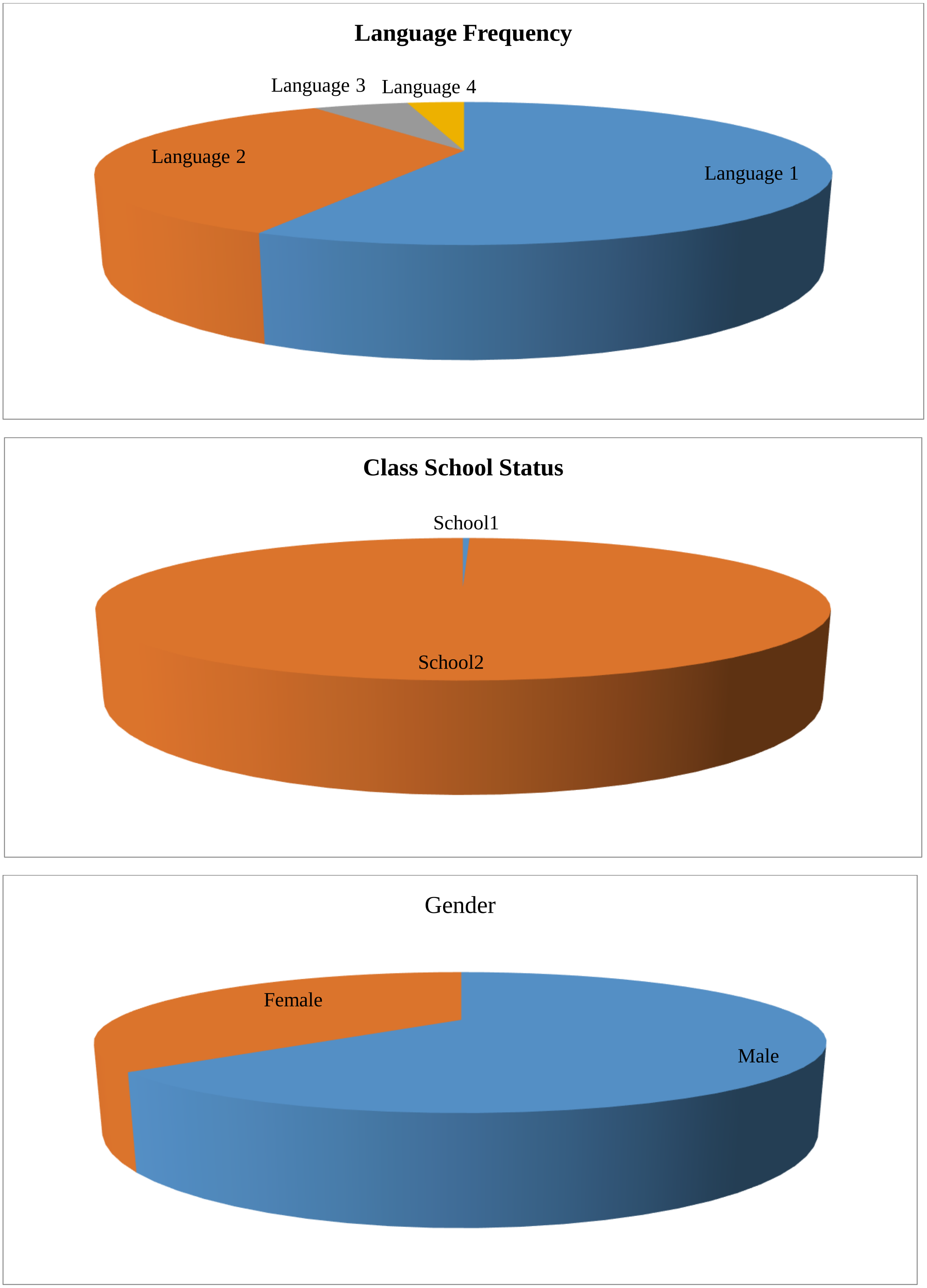


Fig. 3 Performance measures of various parameters and their visualizations

Table 3. Feature set vs. Data type representation

Student Data Set –A		**Student Data Set –B**	
Feature Name	Type	Feature Name	Type
GenderCode	Boolean	GenderCode	Boolean
IntBrothers	Number	IntBrothers	Number
IntSisters	Number	IntSisters	Number
IntSchoolBrothers	Number	IntSchoolBrothers	Number
IntSchoolSisters	Number	IntSchoolSisters	Number
ClassSchoolStatus	Boolean	ClassSchoolStatus	Boolean
Disability01	Boolean	Disability01	Boolean
VchMotherTounge	Number	VchMotherTounge	Number
MotherTongueBin	Binary	MotherTongueBin	Binary
Lang1	Boolean	Lang1	Boolean
Lang2	Boolean	Lang2	Boolean
Lang3	Boolean	Lang3	Boolean
Lang4	Boolean	Lang4	Boolean
Religion	Number	Religion	Number
Result	String	Result	String

Table 3: Eight classification methods were employed to construct the forecasts in the initial experiment: SVM, J48 Decision Tree, logistic Regression, multilayer perceptron, voting, random forest, bagging, and AdaBoost. X%age, XII%age, X, Gap Year, Pass Year, Branch, Program, Admission Through, Entrance Test Year, and Program Completed in Stipulated Time are the sole scholastic characteristics considered when building the models. Tables 3 and 4 display the precision and recall statistics and the F1 measure statistics for every design.

The table reveals that 79.6% is the greatest F1-measure that may be obtained using classifiers (using voting meta-classifier, logistic Regression, and multilayer perceptrons), where the voting meta-classifier is an ensemble learning method using the J48 Decision Tree as well as the multilayer perceptron. Exercise 2 In the second experiment, eight classification algorithms with SMOTE—Logistic Regression, J48 Decision Tree, SVM, Multilayer Perceptron, Random Forest, Voting, AdaBoost, or Bagging—are used to construct the predictive model. When building the system, both academic and non-academic variables were taken into account. Table 4 displays the F1-score, precision, and values for every predictor. According to the tables, classifiers can get an F1-score of up to 93.8%. (using the Random Forest meta-classifier).

4. Results and Discussion

This study aims to demonstrate the effects of the unbalanced information issues and address them by employing a variety of resampling strategies; other goals include selecting the most appropriate resampling method, evaluating the relative merits of the various remaining models, distinguishing between multiclass and paired order, and assessing the significance of the highlights' structure. Python, a universally applicable high-level programming language, was used to code all of the introduced models and tactics. A 2GHz Intel Core i7 processor with 4GB of RAM handles all of the mundane tasks. It is important to note that all classifiers are run on the unbalanced information to demonstrate how the problem in asymmetric data affects the presentation of the models. Then, all classifiers are run on information that has been readjusted using resampling techniques in order to provide a more accurate assessment of the efficacy of these methods in resolving the lop-sidedness problem.

This paper attempts to show the impact of imbalanced information issues, handle this issue utilizing the different resampling strategies; also, deciding the best resampling technique and the best classifier compared to every other model and analyzing the contrast between two-class classification and multiclass, and the significance of the highlights' highlights'structures are among the points from this studies this paper. The most mainstream assessment method to quantify a classifier's performance is exactness. This measurement is the extent of the quantity of right expectations, and the overall quantities of tests were analyzed. Even though exactness is simple to comprehend, it overlooks numerous fundamental factors that ought to be considered in evaluating the performance of a classifier.

Table 4. Performance measures of students

Classifier	Precision	Recall	F1-Score
J48 Decision Tree	93.6	93.1	93.3
Logistic	91.2	89.7	90.4
Regression	92.6	90.6	91.6
Multilayer	96.2	89	92.5
Perceptron	100	85.8	92.5
SVM	97.1	87.2	91.9
AdaBoost	97	90.9	93.9
Bagging	93.3	91.5	92.5
Random Forest			
Voting			

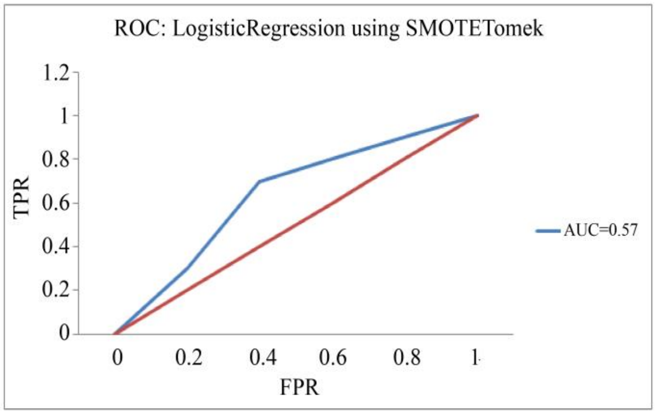
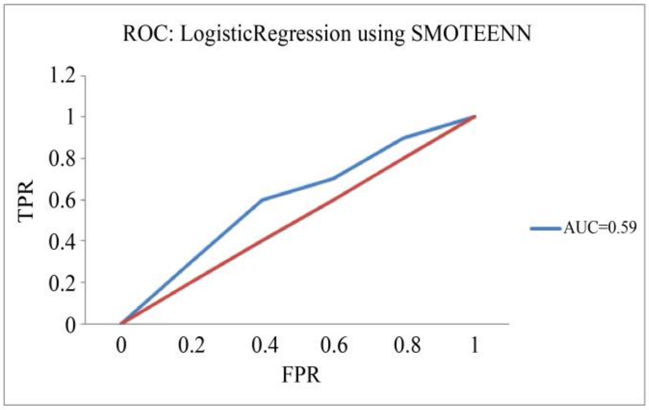
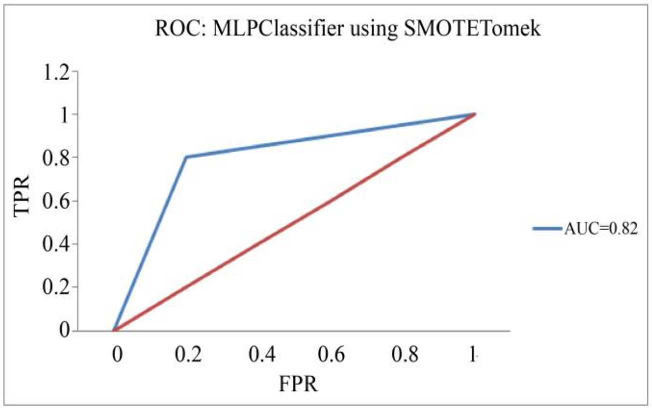
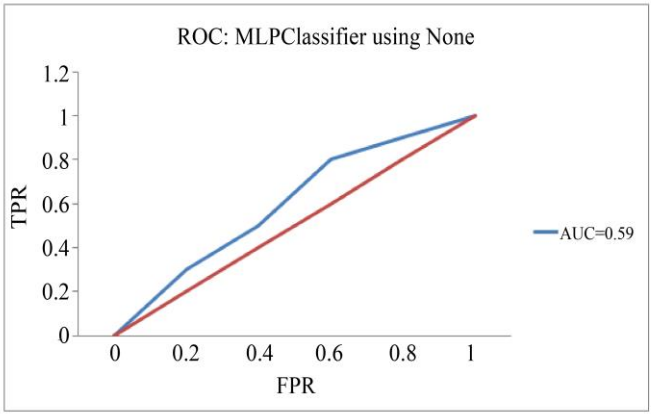
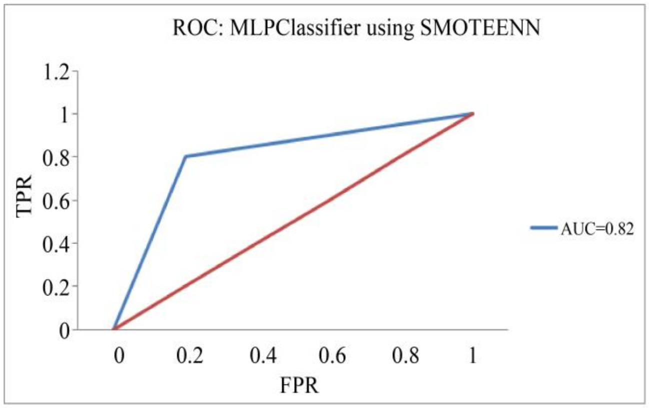


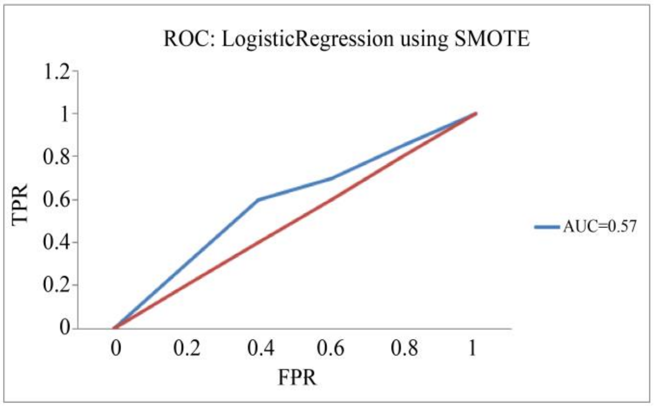
Fig. 4 AUC of various classifiers on Experiment 1

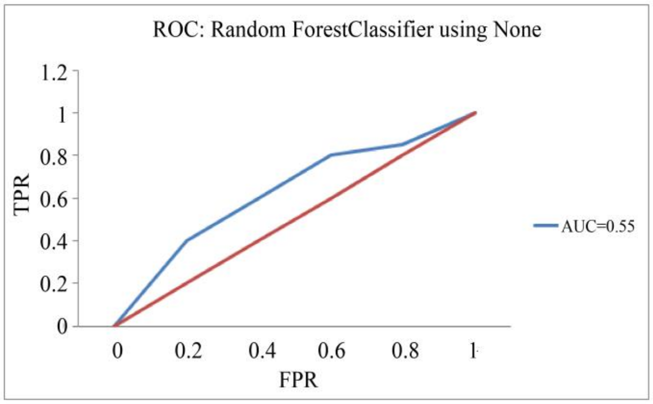












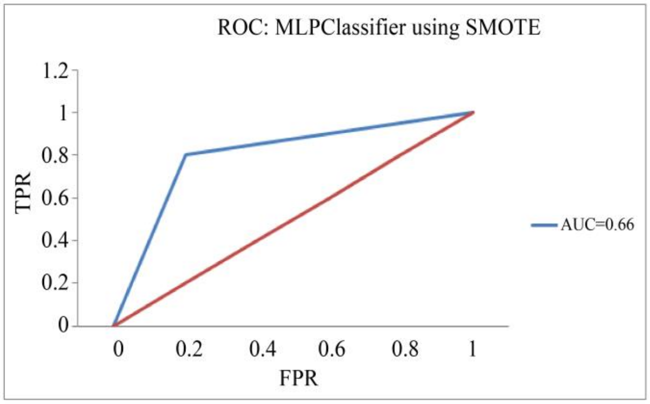


Fig. 5 AUC of various classifiers on Experiment 2

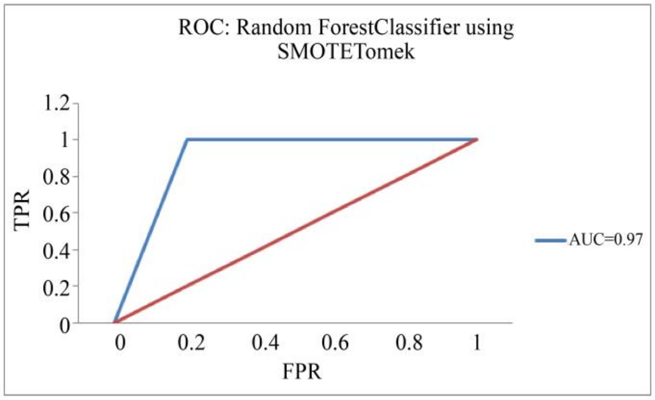


Fig. 6 AUC of various classifiers on Experiment 3

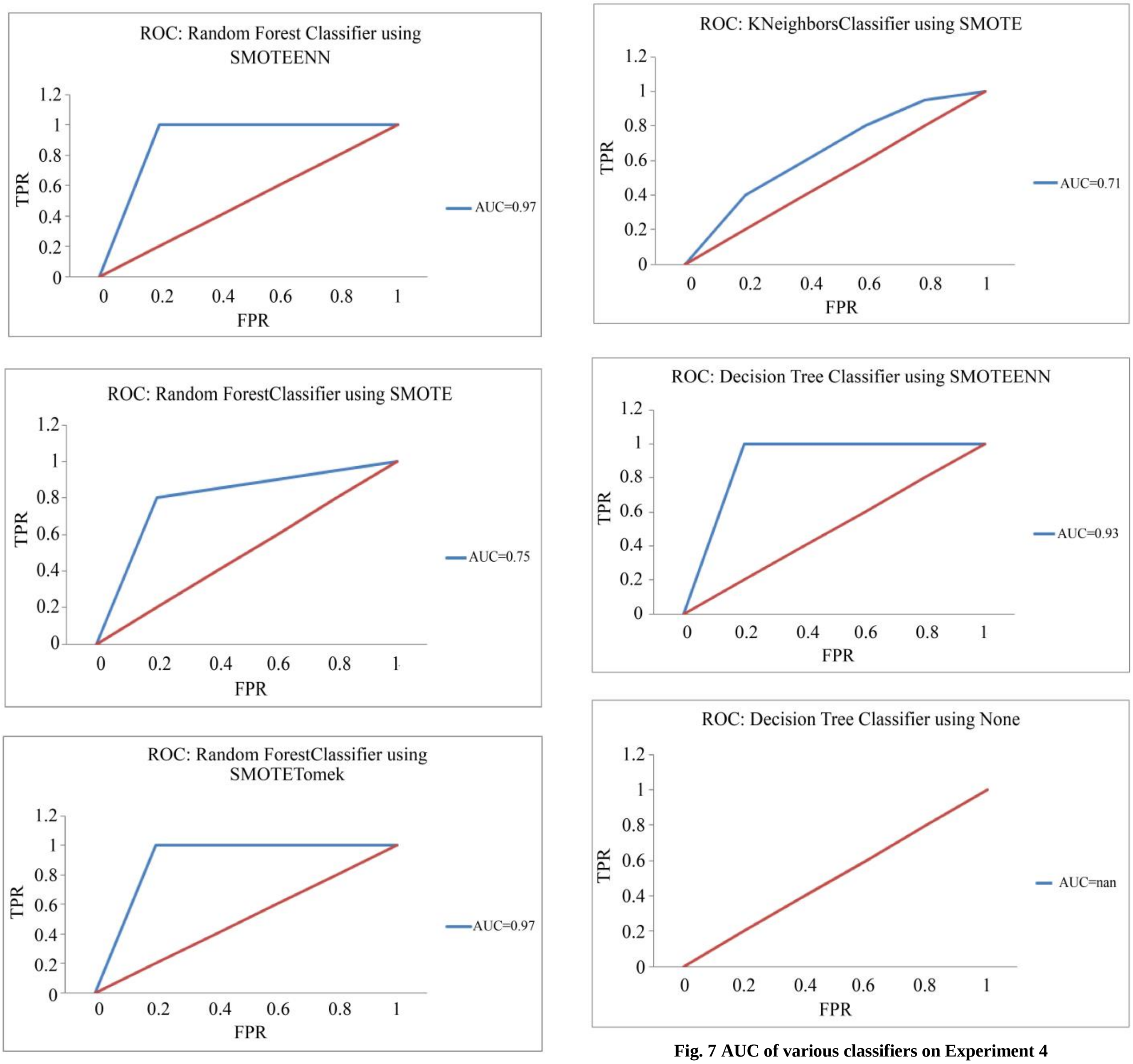


Fig. 7 AUC of various classifiers on Experiment 4

Table 5. Performance analysis of state-of-the-art classifiers vs. feature models

	None	SMOTE	SMOTETomek	SMOTEENN
DecisionTreeClassifier	0.542018	0.725589	0.965854	0.963506
KNeighborsClassifier	0.616430	0.712414	0.930275	0.933919
LogisticRegression	NaN	0.565586	0.606834	0.589493
MLPClassifier	0.652504	0.655894	0.824258	0.815152
RandomForestClassifier	0.550091	0.747669	0.970436	0.974411

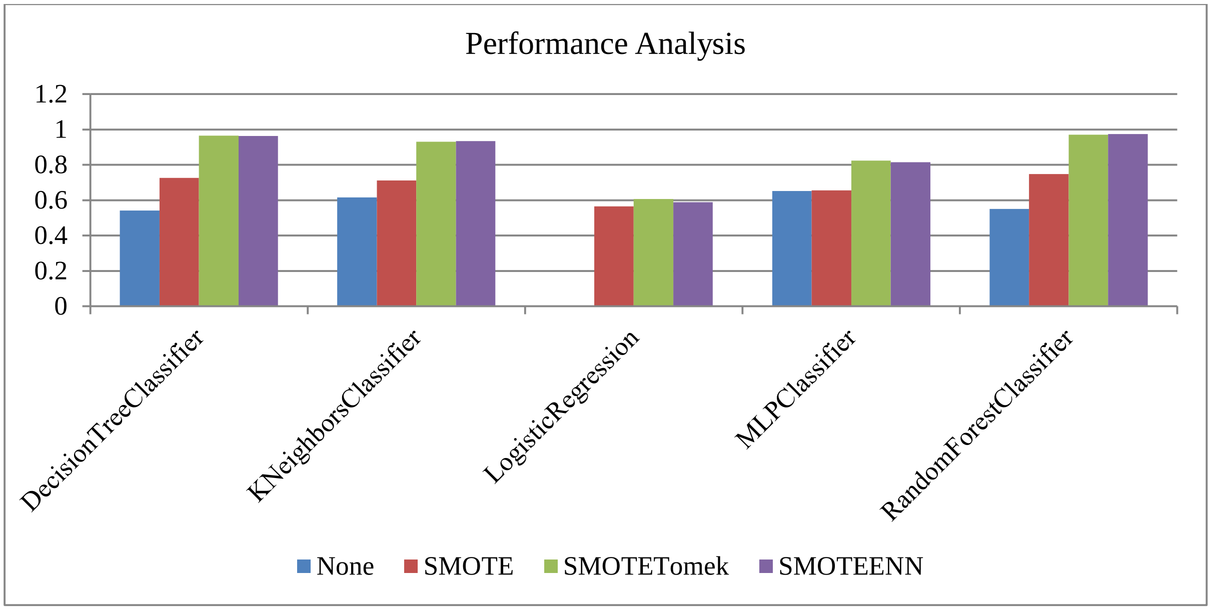


Fig. 8 Performance Analysis of classifiers vs. feature models

Table. 6 Data classification measure using cross-validation

Classifiers	**None**	**Random Under sampling**	**CNN**	**Near Miss**	**ENN**	**RENN**	**Tomek Links**	**SMOTE**	**Random Oversampling**	**SMOTEENN**	**SMOTT omek**
Decision Tree Classifier	0.496118	0.440419	0.329967	0.424171	0.666050	0.834829	0.489381	0.728201	0.758342	0.862706	0.754452
Logistic Retrogression	0.509190	0.331147	0.416988	0.617306	0.524899	0.527254	0.511774	0.364268	0.359198	0.514448	0.341385
K Neighbors Classifier	0.507561	0.469819	0.358006	0.489123	0.668403	0.860551	0.505590	0.718269	0.735444	0.943194	0.734714
Random Forest Classifier	0.502880	0.422818	0.281351	0.449868	0.730360	0.924528	0.513475	0.761615	0.803596	0.938769	0.798183
MLP Classifier	0.520391	0.372000	0.408957	0.673992	0.572032	0.690398	0.516574	0.505113	0.543419	0.698833	0.504265

When faced with a multi-classification problem, it can be difficult to determine how much weight to give to each possible classification method and which results will be best if resampling and a classifier are used. Using imbalanced data in the multi-grouping problem shows that AI systems can not provide exact answers with imbalanced datasets and that most classifiers cannot predict all objective classes.

As a result, rectifying the information disparity is crucial. Table 5 and Figure 9 summarise the results of applying six different resampling models to both altered datasets, revealing each AI technique's resulting accuracy. The precision outcome obtained by using an unbalanced dataset is subpar. Precision can be helpful if there are roughly the same number of tests for each category in the data. However, precision is completely useless with a lopsided set of experiments, since the algorithm anticipates classifier estimates. The precision check of the balancing datasets is not ideal. Since most classifiers consider all possible classes, it is reasonable to assume that they will predict less accurate outcomes when using fair data shown in Table 6.

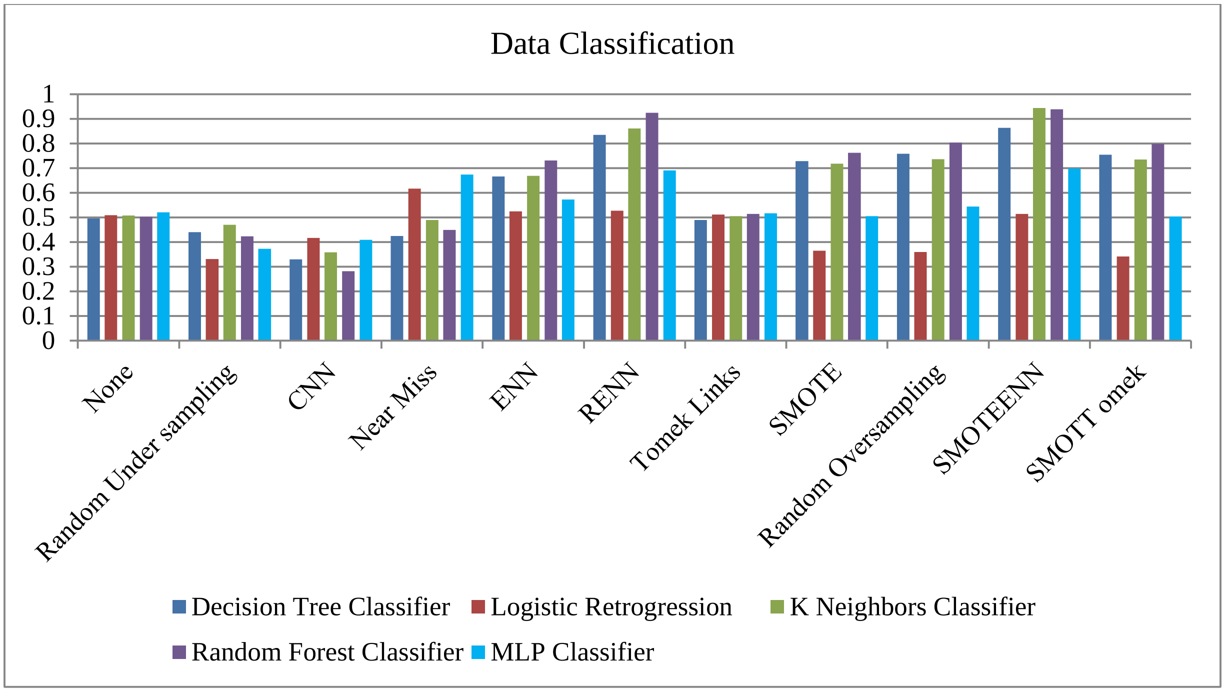


Fig. 9 Data classification using Cross Validation

Since the skewed data problem has been fixed by resampling methods, the precision may now be relied upon. The conclusion of both the precision tests and recalls is shown together in Figures 5 to 8. The review test's findings are comparable to the accuracy; however, specific machine learning models have made great strides in this area. For example, on the student model-2 dataset, Support Vector Machine attained a result of 44.80% with the exactness test when utilizing unbalanced information, but this climbed to 57.31% after using the SVM-SMOTE approach to fix the data. XG-accuracy Boost's test result in the Portugal dataset using unbalanced information was 64.85%, but this was increased to 76.32% after training, with adjusted information applying the Borderline SMOTE technique. Using the F1-score, as was mentioned, makes it easier to analyze the results of the review and accuracy tests. To be fair, classifiers don't do great with all the classes and don't achieve a great F1-score result when using imbalanced data.

Correcting the knowledge gap is crucial to resolving this underlying problem. In all datasets, all four classes are expected and analyzed, demonstrating that classifiers do not miss any classes after employing skewed data through various resampling techniques. This is a crucial factor in our reliance on skewed statistics. For instance, the Artificial Neural Network model may choose to ignore one of the classes during training if the data is not evenly distributed. However, after the inequality issue is resolved, this model takes into account all demographics. Table 5 and Figure 9 display the F1-score outcomes for all applied AI models.. Classifiers' performances might vary widely when using the corrected data produced by alternative resampling methods, and vice versa. As a result, determining the optimal resampling strategy for maximizing the performance of AI models remains a formidable challenge. The challenge of determining the optimal resampling technique can be mitigated through the use of factual essentialness tests. In accordance with what was announced, this study makes use of the NH, gathered by a mixed five-overlay cross-validation for every resampling approach that depends on various AI designs.

5. Conclusion

In conclusion, this study introduces a comprehensive ensemble-based framework for Educational Data Mining that addresses critical issues such as class imbalance and limited model generalizability. By integrating multiple machine learning classifiers-Decision Tree, Logistic Regression, Random Forest, Multilayer Perceptron, and K-Nearest Neighbor-with the Synthetic Minority Oversampling Technique (SMOTE), the proposed model achieves improved classification accuracy and fairness, especially for underrepresented student groups. Experimental validation using a real-world dataset of 6,807 students confirms the model's effectiveness across diverse performance metrics. This work extends existing EDM research by offering a scalable, automated solution that bridges traditional statistical analysis with modern machine learning practices. Future directions include the integration of more advanced balancing techniques such as ADASYN or Borderline-SMOTE, real-time prediction models using streaming data, and deployment in institutional dashboards for early intervention systems. Additionally, the framework can be adapted for cross-institutional datasets to support national-level education policy and personalized learning pathways.

References

[1] Niroj Sapkota et al., “Data Summarization Using Clustering and Classification: Spectral Clustering Combined with k-Means Using NFPH,” *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 146-151, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Haijun Li, and Qingping Lu, “K-CV Parameter Optimization Method in the Application of SVM Classification Data,” *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, Beijing, China, pp. 25-29, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sneha Chandra, and Maneet Kaur, “Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining,” *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, pp. 376-381, 2015. [Google Scholar] [Publisher Link]

[4] Okfalisa et al., “Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification,”*2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, pp. 294-298, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yoga Pristyanto, Irfan Pratama, and Anggit Ferdita Nugraha, “Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification,” *2018 International Conference on Information and Communications Technology*, Yogyakarta, Indonesia, pp. 310-314, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[6] Muhammet Sinan Başarslan, and İrem Düzdar Argun, “Classification of a Bank Data Set on Various Data Mining Platforms,” *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting*, Istanbul, Turkey, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] Elena Baralis, Silvia Chiusano, and Paolo Garza, “A Lazy Approach to Associative Classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 156-171, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hamza Erol, Bala Mikat Tyoden, and Recep Erol, “Classification Performances Of Data Mining Clustering Algorithms For Remotely Sensed Multispectral Image Data,” *2018 Innovations in Intelligent Systems and Applications*, Thessaloniki, Greece, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Lei Zuo, and Junfeng Guo, “Customer Classification of Discrete Data Concerning Customer Assets Based on Data Mining,” *2019 International Conference on Intelligent Transportation, Big Data & Smart City*, Changsha, China, pp. 352-355, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[10] Stamos T. Karamouzis, and Andreas Vrettos, “An Artificial Neural Network for Predicting Student Graduation Outcomes,” *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, USA, pp. 991-994, 2008. [Google Scholar]

[11] Rosângela Marques de Albuquerque et al., “Using Neural Networks to Predict the Future Performance of Students,” *2015 International Symposium on Computers in Education (SIIE)*, Setubal, Portugal, pp. 109-113, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[12] Samy Abu Naser et al., “Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology,” *International Journal of Hybrid Information Technology*, vol. 8, no. 2, pp. 221-228, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[13] Tismy Devasia, T.P. Vinushree, and Vinayak Hegde, “Prediction of students performance using Educational Data Mining,” *2016 International Conference on Data Mining and Advanced Computing*, Ernakulam, India, pp. 91-95, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[14] Zlatko J. Kovačić, “Early Prediction of Student Success: Mining Students Enrolment Data,” *Proceedings of Informing Science & IT Education Conference*, pp. 1-19, 2010. [Google Scholar] [Publisher Link]

[15] Anal Acharya, and Devadatta Sinha, “Early Prediction of Students Performance using Machine Learning Techniques,” *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37-43, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[16] Farshid Marbouti, Heidi A. Diefes-Dux, and Krishna Madhavan, “Models for Early Prediction of at-Risk Students in a Course Using Standards-Based Grading,” *Computers & Education*, vol. 103, pp. 1-15, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[17] Martin Hlosta, Zdenek Zdráhal, and Jaroslav Zendulka, “Ouroboros: Early Identification of at-Risk Students without Models Based on Legacy Data,” *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, Vancouver British Columbia Canada, pp. 6-15, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[18] Vaibhav Kumar, and M.L. Garg, “Comparison of Machine Learning Models in Student Result Prediction,” *International Conference on Advanced Computing Networking and Informatics*, pp. 439-452, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[19] Dorina Kabakchieva, “Predicting Student Performance by Using Data Mining Methods for Classification,” *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 61-72, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[20] Paulo Cortez, and Alice Silva, “Using Data Mining to Predict Secondary School Student Performance,” *Proceedings of 5th Annual Future Business Technology Conference*, pp. 5-12, 2008. [Google Scholar] [Publisher Link]

[21] Mushtaq Hussain et al., “Using Machine Learning to Predict Student Difficulties from Learning Session Data,” *Artificial Intelligence Review*, vol. 52, pp. 381-407, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Ya-Han Hu, Chia-Lun Lo, and Sheng-Pao Shih, “Developing Early Warning Systems to Predict Students’ Online Learning Performance,” *Computers in Human Behavior*, vol. 36, pp. 469-478, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ahmed Mueen, Bassam Zafar, and Umar Manzoor, “Modeling and Predicting Students’ Academic Performance Using Data Mining Techniques,” *I.J. Modern Education and Computer Science*, vol. 8, no. 11, pp. 36-42, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[24] Dech Thammasiri et al., “A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition,”*Expert Systems with Applications*, vol. 41, no. 2, pp. 321-330, 2014. [CrossRef] [Google Scholar] [Publisher Link]