MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

Predicting Students' Performance and Its Influential Factors Using Ensemble Approach - LRN

Pooja [1] and Dr Rajni Bhalla [2]
[1] Research Scholar (Computer Application) Lovely Professional University, Phagwara, Punjab (INDIA)
[2]Associate Professor, School of Computer Application, LPU Phagwara, Punjab (INDIA)
Email: ·Poojagori51@gmail.com[1], rajni.b27@gmail.com[2]

ABSTRACT:
The effectiveness of any educational institution can be measured by the academic growth of their students. In light of emerging technologies such as artificial intelligence, machine learning and data mining, institutions are integrating technology into conventional teaching approaches. In today's education system, prediction of students' performance is extremely important. Predicting student performance in advance can help the students, teachers as well as the institutions. Prediction of students' performance has been crucial issue to forecast whether a student will complete his/her study within stipulated time period or not.Students should be advised well in advance to concentrate their efforts in specific area in order to improve their academic achievement. The objective of this study is to identify uncover/hidden factors that could assist in predicting the students' performance. We used educational data mining (EDM), which is the collection of techniques used to uncover hidden patterns in massive amount of existing data. In this study, primary data set has been collected from the students of different schools of Jalandhar city . The size of dataset is 399 records with 17 attributes. Features that are redundant or unnecessary are removed from the dataset via feature selection approaches, which have been used to improve the prediction accuracy of standalone classifiers while lowering overall costs. We used and analyzed the different machine learning algorithms such as SVM, Logistic Regression, Naïve Bayes (NB), Random Forest, Boosting, Bagging and Stacking. To lessen sample related bias in our investigation, we used 5-fold cross validation. One of the causes of accurate prediction outcomes is this. The whole dataset is divided into two sets of training set and testing set 75:25 proportions respectively. In this research, an ensemble model LRN proposed to identify prediction of students' performance. For better performance, with hyper-parameter tuning three base classifiers are gathered and added to the proposed ensemble model LRN (Logistic Regression+Random Forest+Naive Bayes). By using ensemble techniques; we will have a good result that demonstrates the dependability of the proposed model. Our proposed model gave accuracy 96% which is the highest among SVM, Logistic Regression, Naive Bayes, Random Forest and other boosting algorithms.

Keywords: Educational data mining (EDM), Machine Learning(ML), Feature selection and feature extraction, students' performance prediction, Ensemble model .

HIGHLIGHTS OF THE PAPER:

(i) An ensemble model- LRN proposed to identify prediction of students' performance.
(ii) By combining various machine learning techniques with bagging, stacking and voting the model enhances precision, accuracy, and recall.
(iii) It offers a more detailed analysis of student performance, leveraging factors like attendance, pending assignments and reappear exams are affecting the performance of students.

1. INTRODUCTION:

In the modern world, a vast quantity of data related to students can be store into any educational institution. Education is a key factor of achieving long term institutional growth. Students are the main asset of any organization. Prediction of students' performance is the major goal of any institution. During the last decade, education level has improved. Educational data mining (EDM) refers to data mining techniques used to analyze educational data[1]. This data contains academic and non-academic attributes of the students. Machine learning generally looks into algorithms that make inferences from the input dataset in order to create broad hypotheses that forecast future occurrences[2]. On the other hand, data mining is essential for the purpose of finding pertinent information among a vast quantity of data. Data mining is very helpful for researchers in the field of education. Enhances analytic focuses on the collecting and evaluation of data from learners in order to optimise learning materials and to enhance learners' learning experiences [3].Data mining and machine learning techniques are very useful in prediction of student' performance and correct decision making. For example, DM approaches could be used to address a number of interesting questions in this field [4, 5]. Which students are enrolled in the most credits? What is the choice of students about the type of course? What are the main reasons for student's dropout their course. Is it possible to predict the student performance? With hyper parameter tuning, the accuracy of propose model is to be increased. Also, feature selection and extraction is applied on the dataset.There are many factors influencing students' academic performance to ensure they are completed their course on time. According to earlier studies, excellent student performance may be attributed to the teaching and learning process, university infrastructure, peer and family pressure, and student finances. Other hidden elements that may contribute to students' extension include their health, albeit it appears that only 25% of students are suffering from significant illnesses at this time while studying that they need to take a special study leave [7].

In next section, feature selection methods namely chi-square, gain ratio, correlation are described and next section describes baseline classifiers used in this study to evaluate predictive performance of feature selection methods. For selecting appropriate attributes from pool of attributes in dataset, feature selection technique is used. It is a pre-processing technique help in reducing the dimensionality of the data by eliminating irrelevant and redundant features. The irrelevant features, if included in classification process, not only reduce its accuracy but also consume more space and execution time. So, feature selection is very important as it facilitates deep understanding of data by studying only relevant features and enhancing accuracy, speed and the predictive capability of classifier [6].

Next section, reviews the existing literature on factors affecting on prediction of students' performance, feature selection based study in prediction of student' performance. Section 3 describes methodology which includes description of feature selection methods and classifiers used in this paper. Section 4 description of dataset utilized for experiment purposes is given. Section 5 presents proposed ensemble model LRN, experiments are carried out and results are summarized, and finally section 6 concludes the research.

2. LITERATURE REVIEW

In education's field, data mining consists hierarchy of data which is different from other fields as stock marketing, fraud detection, heart disease detection etc. Data mining in educations are categorized into 5 dimensions such as classification, prediction; clustering etc. one dimension is predictions[11]. Several researchers focused on predicting the students' performance based on diver's factors using ML algorithms. Mainly regression, classification and clustering techniques are used to predict the students' performance. However, we observed that classification is one of the most popular techniques used in predicting the academic performance. Various models and classification algorithms have been used to predict the students' performance. Findings have shown that student academic achievements are typically influenced by a myriad of factors, ranging from academic and non-academic attributes [8]. Prediction of students 'performance can be considered a difficult task as it's depends on many attributes related to the students. These attributes can be categorized as demographics, psychological, academic progress, CGPA, grade and educational background [9]. According to [10], the predictions of student achievements are calculated mainly using previous semesters' grades and current coursework assessments, such as assignments, midterms, and projects, and final exams. The most important attribute used to predict the performance is student's CGPA. The demographics information that consists of the family background, the gender,

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

and age is also considered an important attributes.Bravo Agapito et al [12] used exploratory factor analysis, linear regression (multiple) and cluster analysis in their study that "age" is a factor that affects students' academic performance of 802 undergraduate in completely online learning. Hamsa et al [13] used dataset of students having Graduates and Master degree in Computer Science and Electronic communication for the prediction of students' performance. They applied classification methods such as fuzzy genetic and decision tree algorithms. Used association rules and clustering for the same factors such as age, gender, annual income, parents' occupation, parents' education, are included [14]. According to [15] they used dataset consisted of academic achievement grades of 1854 students who took the Turkish Language-1 course in state University in Turkey during the fall semester of 2019-20. They considered Random Forests, nearest neighbour, KNN, SVM, Logistic Regression were calculated and compared to predict the final grades of the students.77% accuracy of KNN produced. According to [16] sentiment analysis also plays important role in the prediction of students' performance.[17] student' grade and educational background considered as important factors. According to [18], different machine learning algorithms such as decision tree, naïve bayes, SVM, Random Forest are compared to predict the students' performance, on student academic dataset, Portugueses. Accuracy rate and F measure of data set is measured with two forms that is with attribute selection and without attribute selection. With attribute selection DT performs better and without attribute selection RF performs better. NB mining technique for data extraction for useful information provides more accuracy [19]. It takes students' academic history as input. As per [22], there is a relationship between big data and the education environment. They mainly focused on tools/techniques, and different big data algorithms used in education context to facilitate and give benefits in the learning process. Students had high acceptance of blending learning other than online learning [23]. One effective strategy to improve educational outcomes is to use the data mining area for educational purposes [24]. According to [25-26], to lessen the effect of data imbalance on the model and increase the precision of student performance prediction, they integrated the SMOTE technique with machine learning techniques. [26] used NB, SVM, NN, Random Forest and Logistic Regression algorithms to predict the performance of students. Midterm exam result used as input. The performance of these algorithms was between 69-75% but random forest outperformed among these algorithms.Their researches concluded that midterm exam result is the best attribute to predict performance. According [28] presents a model based on SVM and logistic regression for predicting students' academic performance. The sequential minimal optimization algorithm outperforms logistic regression in accuracy. The research aims to help educational institutes predict future student behavior and identify impactful features like teacher performance and student motivation, ultimately reducing dropout rates. On the other hand, many machine learning and data mining techniques have been used to predict the students' performance such as: Artificial Neural Network (ANN), Random Forest (RF), Lasso, K-Nearest Neighbor (KNN), Support Vector Machine (SVM); Linear Regression, Logistic Regression, Decision Tree (DT), Naïve Bayes (NB), Principal Component Analysis (PCA).

Table 1 shows a summary of the research papers that relate to this study.

Ref	Features	Dataset Size	Machine Learning Algorithms	Best Algorithm
13	Internal grades, sessional grades and admission score	168	Fuzzy Genetic Algorithm and DT	FGA model is less strict than DT
14	Age, gender, annual income, parents' occupation, parents' education		Association rules and clustering	Association rule
15	Midterm, faculty, department	1854	Random Forests, KNN, SVM and Logistic Regression	KNN
16	Sentiment Analysis	396	-	-
17	Grades, Backgrounds	1169	Linear Regression, Logistic Regression, RF, kNN, Proposed Progressive Prediction algorithm	Proposed progressive prediction algorithm
20	Medu,Fedu,Fam_size,Pstatus, Traveltime,G1,G2	649	Naïve Bayes, SVM, ID3, C4.5	SVM
21	Program,ethnicity, gender, age group, scholarship, transfer status, course load, admitted on prohibition, result, math level, English level, School system,	1491	ANN, NB, SVM, DT, K-means Cluster, K-Nearest Neighborand Linear Regression	NB
27	Personal and demographics information, student satisfaction and integration	149	ANN, Logistic Regression	Logistic Regression
34	Attitudes of students, Personal development	78	Chi-square	Chi-square
35	sociodemographic characteristics, individual ,lifestyle, behavioral factors ,familyand psychosocial variables	659	Multi stage sampling technique, Simple Random Sampling (SRS),Chi square test, Bivariable and multivariable logistic regression, Cross Sectional Technique	Cross Sectional Technique
36	Age, academic performance, location, and online learning behavior.	15K	Naive Bayes, Support Vector Machines, Multi-Layer Perceptron,and Logistic Regression,Ensembled custom 1D Convolutional Neural Network (CNN)	Proposed ensemble model

Dataset Used:

In this paper, we have used 2 primary datasets and 1 secondary dataset.

Description of Primary Dataset1: We have collected the data from various Govt./Private schools of Jalandhar city through online/offline mode and grouped the attributes into 7 categories.

Table 2 shows a summary of primary Dataset1

Sr No.	Attribute	Description
1	Personal Information	Name,Class and Age
2	School Information	Name of the School,Affiliation of the school,Type of the school
3	Family information	Parents Annual Income, Family Type(Nuclear/Joint),Total no. of family members, Living place(Urban/Rural),Father's Education level, Mother's Education Level
4	Perception metrics	Education's role in future job prospects, Parental encouragement for studies, Parents harmonious
5	Admission Preference	Factors considered when choosing a school such as Distance from home,Transportation Facility,Quality of Education, Previous school results, Popularity of school, Facilities for disable children, ICT facilities, Co-curricular activities ,Extra Curricular Activities
6	Instructor Evaluation	Effectiveness of lecturers, Presentation Clarity and organization, Stimulation of student interest, Effective use of class time, Instructor Availability and helpfulness, Prompt grading and feedback
7	Miscellaneous	Medium of teaching,Types of Students(Day scholar/hostler),Gap year(if any),Study time at home,Distance to school,Scholarship/incentives availed, Previous class result

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

Description of Primary Dataset2: We have collected the data from an organization. There are 17 features (11 numerical features such as RegdNo, Attendance15161, Attendance15162, PendingE, PendingF, ReappearExamGiven, ReappearSummerTermRegistered,, IsAbove75T1, IsAbove75T2, SatisfactoryTerms, Performance) and 6 categorical features such as Section, School, StudentStatus, Country, Region,Economy**.**
Table 3 shows a summary of primary Dataset2

Sr No.	Attribute	Description	Type
1	RegdNo	Registration Number of student	Int64
2	Section	Section of the student	object
3	School	Type of school	object
4	StudentStatus	Status of the student(Active, Not Active)	object
5	Country	Student belongs to which country	object
6	Attendance15161	Attendance in 15161	Int64
7	Attendance15162	Attendance in 15162	Int64
8	PendingE	Pending Exam details	Int64
9	PendingF	Pending Files	Int64
10	ReappearSummerTermRegistered	Is he/she register for reappear summer term	Int64
11	ReappearExamGiven	Reappear exam given(yes/No)	Int64
12	IsAbove75T1	Is Marks >75 in term1	Int64
13	IsAbove75T2	Is Marks >75 in term2	Int64
14	SatisfactoryTerms	Yes/No	Int64
15	Region	Region	Object
16	Economy	Economy	object
17	Performance	Performance of Student	Int64

Description of Secondary Dataset: We have online dataset from kaggle.33 attributes and 396 records in each course (Math Course and Portuguese Course)
Table 4 shows a summary of Secondary dataset

SrNo.	Attribute	Description	Possible Values
1	School	student's school	binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira
2	Sex	student's sex	binary: "F" - female or "M" - male
3	Age	student's age	numeric: from 15 to 22
4	Address	student's home address type	(binary: "U" - urban or "R" - rural)
5	Famsize	family size	(binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6	Pstatus	parent's cohabitation status	(binary: "T" - living together or "A" - apart)
7	Medu	mother's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8	Fedu	father's education	(numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9	Mjob	mother's job	(nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10	Fjob	father's job	nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11	Reason	reason to choose this school	(nominal: close to "home", school "reputation", "course" preference or "other")
12	Guardian	student's guardian	(nominal: "mother", "father" or "other")
13	TravelTime	home to school travel time	(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14	StudyTime	weekly study time	(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	Failures	number of past class failures	(numeric: n if 1<=n<3, else 4)
16	Schoolsup	extra educational support	(binary: yes or no)
17	Famsup	family educational support	(binary: yes or no)
18	Paid	extra paid classes within the course subject	(binary: yes or no)
19	Activity	extra-curricular activities	(binary: yes or no)
20	Nursery	attended nursery school	(binary: yes or no)
21	Higher	wants to take higher education	(binary: yes or no)
22	Internet	Internet access at home	(binary: yes or no)
23	Romantic	with a romantic relationship	(binary: yes or no)
24	Famrel	quality of family relationships	(numeric: from 1 - very bad to 5 - excellent)
25	Freetime	free time after school	(numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends	(numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption	(numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption	(numeric: from 1 - very low to 5 - very high) attended nursery school
29	Health	current health status	(numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences	(numeric: from 0 to 93)
31	G1	first period grade	numeric: from 0 to 20
32	G2	second period grade	numeric: from 0 to 20
33	G3 (output target)	final grade	(numeric: from 0 to 20,

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

3. PROPOSED METHODOLOGY

This section describes the structure of model.

(1) Data Collection: There are different datasets are available online which help us to predict the performance of students or we can collect real time dataset from educational institutions. In this proposed methodology, we used real time dataset with 399 records and with 17 different attributes related to the students.

(2) Data Preprocessing.:Since data was collected from students, there were missing values in it. Basic prepocessing techniques was applied to ensure consistency and readiness for analysis.

(3) Feature Selection/Extraction: Select the best correlated features using correlation matrix or feature extraction is applied on the dataset to find the best 3 features that will effect on performance of student. Train the model with training set.

(4) Data Splitting: Dataset was splitted into training set and testing set to build and validate the model.

(5) Model Training and testing: After the model has been trained with the labels and tested the model with testing dataset, performance of model will be evaluated.

(6) Performance Evaluation :Accuracy, precision, recall and F1_Score are standard metrics used for evaluation. After that we compared the accuracy of proposed model with existing models to determine improvement or superiority.

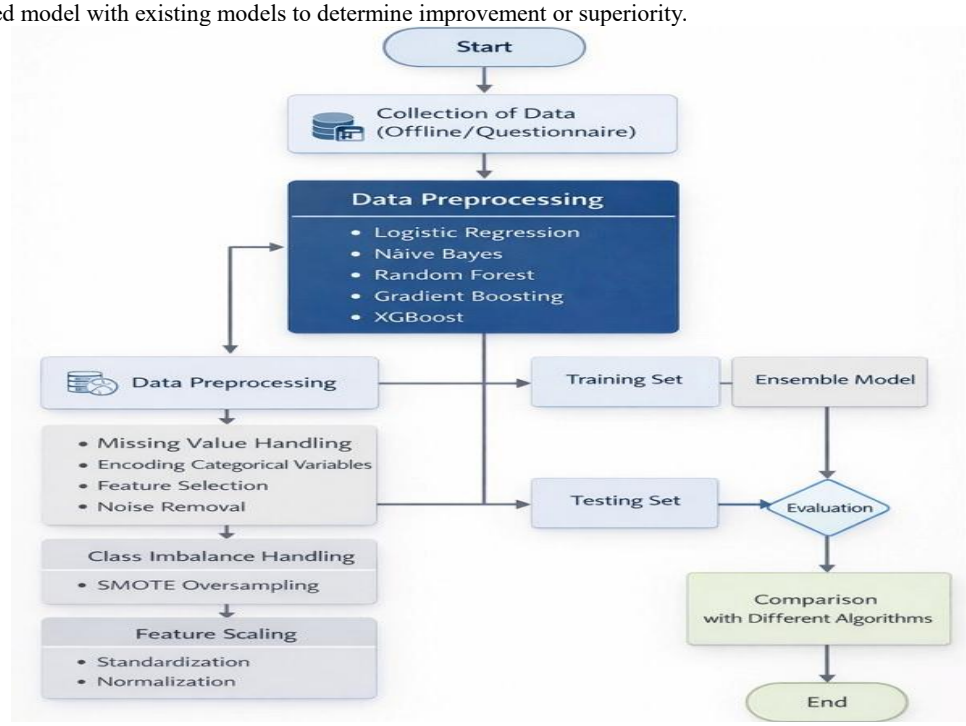


Fig 1: Proposed Methodology

3.1 Data Pre-processing:

After the collection of dataset, preprocessing methods are applied to develop the data set quality. The data pre-processing is regarded as an essential step in the process of knowledge discovery which involves cleaning of data, feature selection, data transformation and data reduction. Before applying the data mining algorithm the data pre-processing is the step which transforms the actual information into an applicable shape to be used by a specific algorithm of mining as shown in fig.2 (a)

For accurate analysis, data preprocessing is a most important step. As our collected real-time dataset has inconsistent, incomplete and inaccurate data. It is also containing blank and missing values. Data processing is used to convert the raw data useful information and reduces the errors which help in decision making.

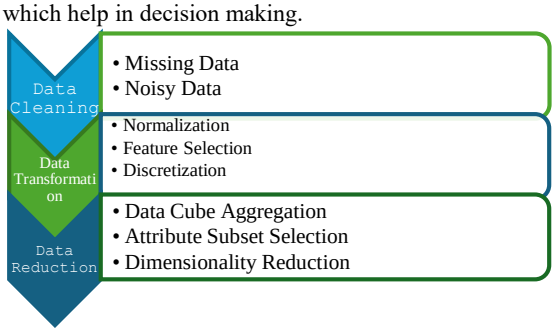


Fig 2(a): Data preprocessing techniques

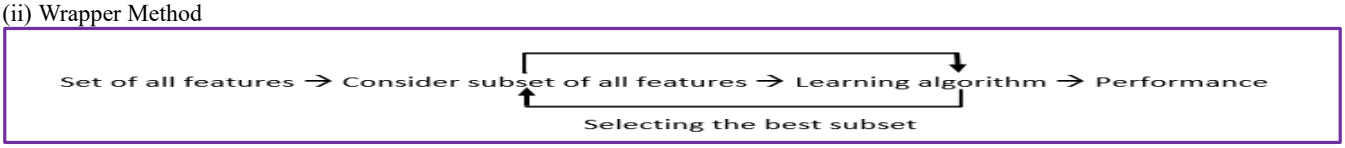
3.2 Features Selection and Features Extraction

Preprocessing technique such as feature selection has been applied on dataset which helps in selecting appropriate attributes. Since, there is massive amount of data, the feature selection techniques help in reducing dimensionality of data by removing redundant and irrelevant attributes. By removing these attributes from the dataset, it facilitates deep understanding of the data by studying only relevant features and algorithms consume less space and fast execution time .Feature selection methods are mainly classified in three categories as shown in Fig2(b).There are three types of feature selection techniques.

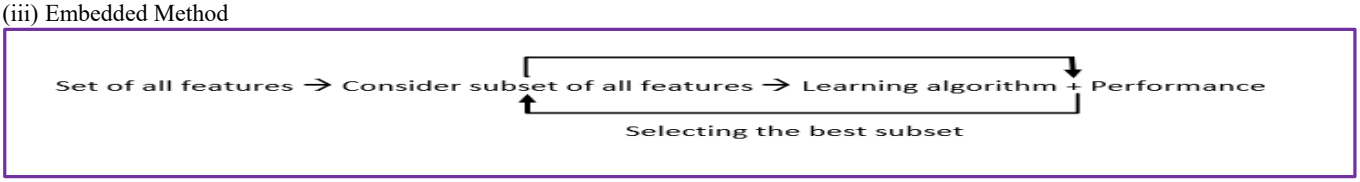
(i) Filter Method

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

(ii) Wrapper Method



(iii) Embedded Method



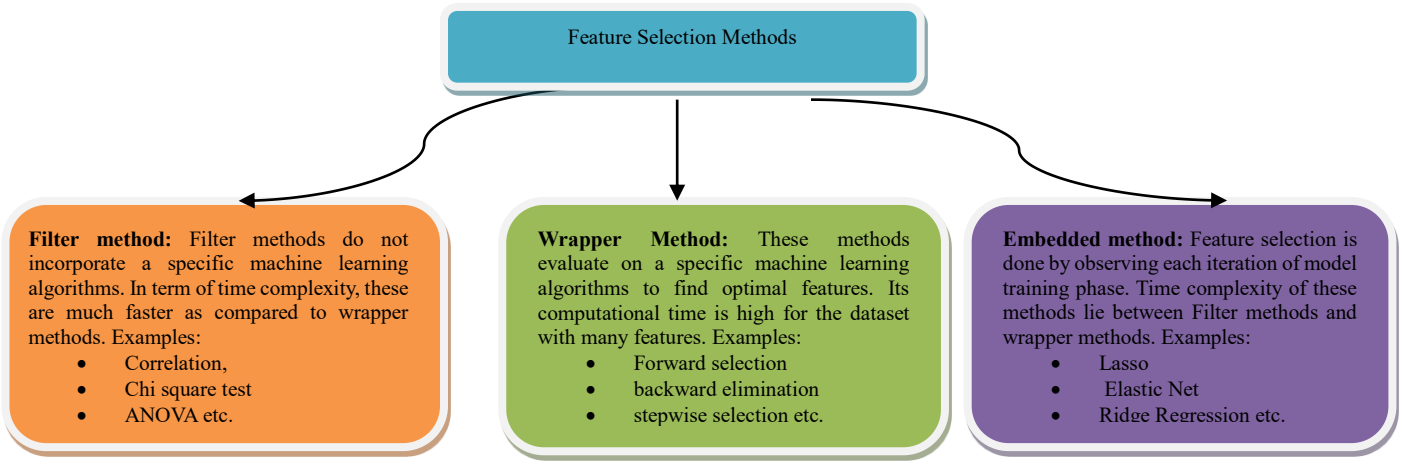


Fig 2(b): Feature selection techniques

Feature extraction techniques aim to create new features that synthesize the majority of the information in the original feature set, hence reducing the amount of features in a dataset [21].

Feature selection methods used:

- **Chi square test:** It is a statistical test of independence to determine dependency of two variables [30]. In order to determine the value of an attribute, a filter-based feature selection method is used to calculate the value of chi-squared statistics with respect to class. This method assesses the ability to predict the value of an attribute from the value of the class by determining whether or not there are statistical links and testing their independence.
- **Correlation:** It is a filter based feature selection method which is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It is often represented by coefficient correlation value lies between -1 and +1, where 0 indicates no correlation, and values closer to -1 or +1 indicate stronger relationships.
- **Information Gain (IG):** In Decision Tree (DT) and Random Forest (RF), Information gain is used to decide the best split. IG helps us understand how much a particular feature contributes to making accurate predictions in a decision tree.
- **Forward Selection:** It is an iterative procedure start with an empty set of features and keep adding a feature which improves our model best after each iteration.
- **Recursive Feature Elimination:** In this method features are selected by recursively considering the smaller and smaller set of features. The least important features are then removed from the current set of features till we are left with the required no. of features.
- **Tree Based Methods:** DT and RF perform feature selection by selecting the most important features for splitting nodes based on criteria like information gain (IG) or Gini impurity.

Baseline Classifiers Used:

- **Support vector machine (SVM):** This method was introduced by Cortes,C.,Vapnik,V.,1995 is very popular machine learning technique which is used for both classification and regression analysis[31]. SVM constructs a hyperplane which separates the data in two classes- positive and negative to maximize the margin between classes which is applied in various fields such as pattern recognition, text categorization etc. [32] used mathematical optimization to address the issues in SVM, such as the detection of relevant features or the accommodation of measurement costs associated with the variables.Kernel functions used in SVM are: Linear Kernel function, Polynomial Function, Gaussian Function and Radial kernel function. To solve the binary classification problems, and due to its kernel functionality, this model is mostly used in machine leaning.
- **Naive Bayes (NB):** It is a probabilistic based supervised machine learning algorithm introduced by John G.H..It is based on Bayesian theorem to predict the class. To solve the classification problem, this probabilistic method is used. It handles discrete as well as continuous data.
- **K-NN:** It is non parametric classifier that uses a distance measure to make predictions without building a model. In K-NN, training stage comprises of storing feature vector and class labels of training dataset. For testing stage, distance from new vector to all stored vectors is computed and closest K samples are selected. Then class label is assigned according to the class of majority of K-NN. It is a non-parametric method that makes predictions based on the similarity of data points in a given dataset.
- **Logistic Regression**: To solve the binary classification problems, this method is used. This algorithm predicts one of the possible outcomes based on distinctive features pertinent to the problem.
- **Random Forest (RF):** It is an ensemble algorithm developed by Breiman,L. [33] used for classification as well as regression. It is based on multiple decision trees and uses bootstrap aggregating. By combining the result of various decision trees, it is constructed. Bagging and voting techniques are

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

applied on the dataset. Then x features are selected from a pool of actual features. Fresh selection of attribute is done at each split. Splitting is continued till the depth d where a decision tree is completed. The result is a random forest, which is made up of many of these decision trees.

- **Stacking Classifier (stacking):** It is an ensemble classification or regression model which consists of two layers estimators. All baseline models exist in layer1 that are used to predict the output on testing dataset. Regressor or Meta classifier exist in layer2 which takes all the predictions of baseline models as input and produce new predictions.
- **XGBoosting Classifier (Boosting):** It is used for classification, regression and ranking. Due to its accuracy and scalability used mostly.

Table 5: Feature Selection on Primary Dataset1

Features Selection Methods Used (Primary Dataset1)					
Name of the attribute	**Chi Square**	**Correlation Coefficient**	**Information Gain**	**Forward Selection**	**Tree Based Methods**
Personal Information	√	√		√	
School Information		√	√		
Family information	√	√	√		
Perception metrics	√		√	√	√
Admission Preference	√			√	
Instructor Evaluation	√	√	√		√
Miscellaneous	√	√	√		√

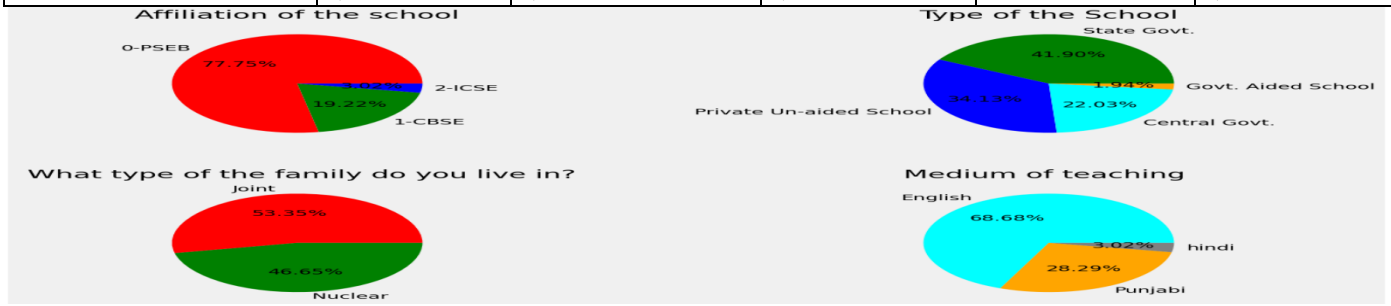


Fig 8: Visualization of Dataset1

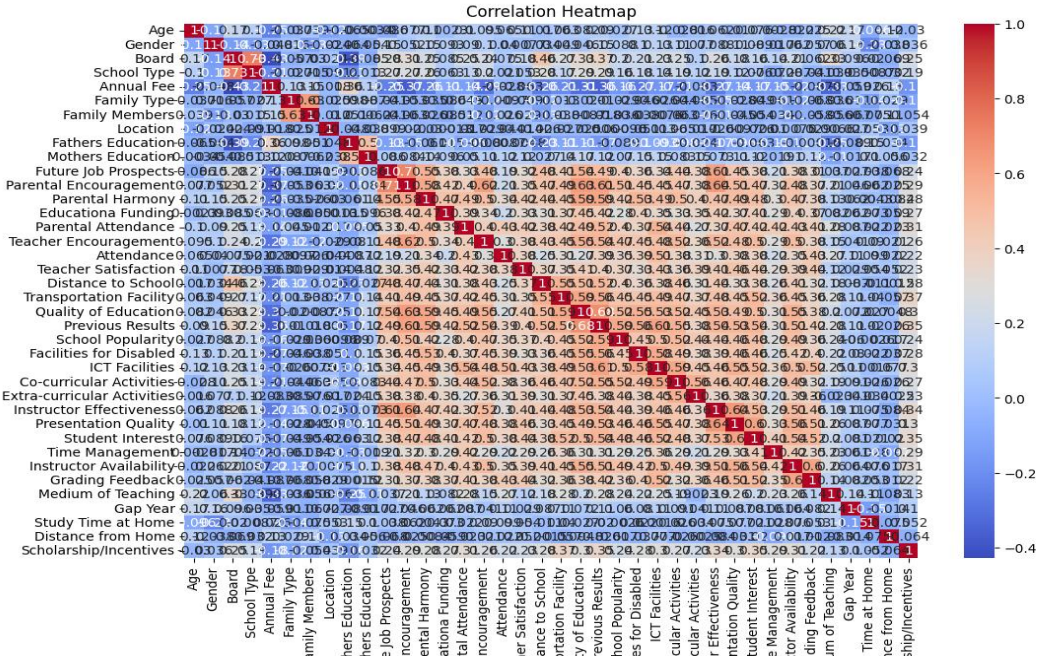


Fig 9: Correlation Heatmap (Dataset1)

Table 6: Feature Selection on Primary Datast2

Features Selection Methods Used (Primary Dataset2)					
Name of the attribute	**Chi Square**	**Correlation Coefficient**	**Information Gain**	**Forward Selection**	**Tree Based Methods**
RegdNo					
Section			√		
School					
StudentStatus			√	√	√
Country					
Attendance15161	√	√	√		√
Attendance15162	√	√	√		√
PendingE	√	√	√	√	
PendingF	√	√	√	√	
ReappearSummerTermRegistered	√	√	√	√	√
ReappearExamGiven	√	√	√	√	√
IsAbove75T1	√		√		
IsAbove75T2	√	√	√		
SatisfactoryTerms	√	√	√	√	√
Region	√				

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

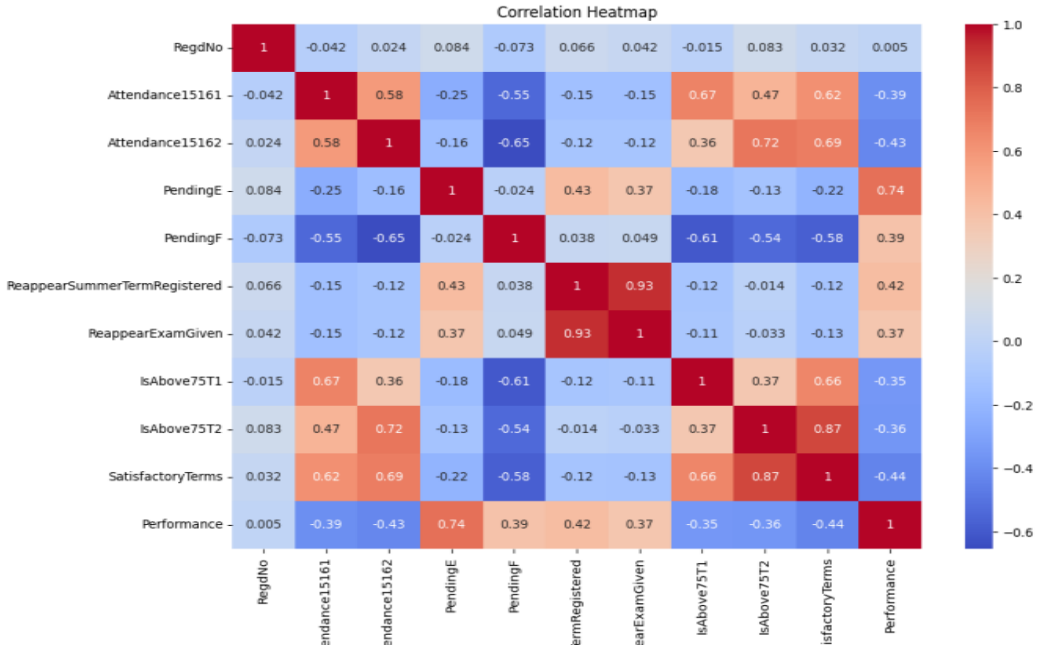


Fig.3: Result of Correlation (Dataset2)

	Features	Score
9	Region	0.735913
6	IsAbove75T1	3.877542
7	IsAbove75T2	6.688248
8	SatisfactoryTerms	12.604874
5	ReappearExamGiven	48.974507
4	ReappearSummerTermRegistered	61.401957
0	Attendance15161	62.356513
1	Attendance15162	180.712138
10	Performance	276.000000
3	PendingF	518.908417
2	PendingE	976.712880

Fig 4: Result of Information Gain

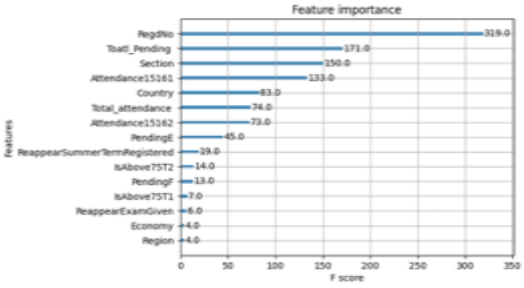


Fig.5: Result of ChiSquare

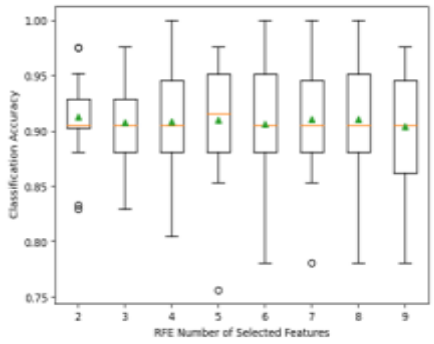


Fig 6: Output of Recursive feature Elimination based on Lasso

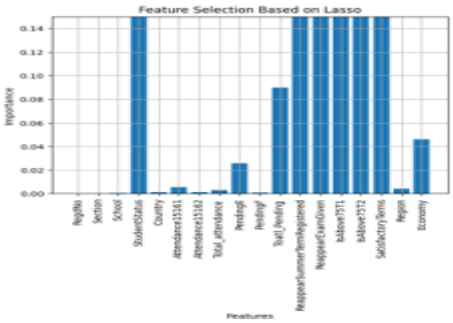


Fig 7: Result of Feature Selection

Dataset Used	Total Attributes	Attributes selected
Real time datset1 (Schools of Jalandhar City)	Personal Information, School Information, Family information, Perception metrics, Admission Preference, Instructor Evaluation, Miscellaneous	Family Information, Admission Preference, Miscellaneous
Real Time Dataset2 (From Institiution)	RegdNo, Section, School, Student Status, Country,Attendance15161,Attendance15162, Pending E,PendingF, Reappear Summer Term Registered, Reappear ExamGiven,IsAbove75T1, IsAbove75T2,Satisfactoryterm,Region	StudentStatus,Attendance15161,Attendance15162, PendingE, PendingF, Reappear Summer Term Registered, Reappear Exam Given, IsAbove75T1, IsAbove75T2,Satisfactoryterm,
Online Dataset kaggle	School, Sex,Age,Address,Famsize,Pstatus,Medu Fedu,Mjob,Fjob,Reason,Guardian,TravelTime,StudyTime,Failures, Schoolsup,Famsup,Paid,Activity Nursery,Higher,Internet,Romantic,Famrel,Freetime,goout,Dalc, Walc,Health,absences,G1,G2	Famzize,Pstatus,MEdu,Fedu,TravelTime,StudyTime,Schoolsup,Health,Absences, G1 and G2

3.3 Proposed Ensemble Classifier (Voting): In ensemble classifier (voting), we create a single model which trains by ensemble of numerous models and predicts output based on their combined majority of voting for each output class.To improve the performance of model on unseen data, cross validation technique is used. Available data is divided into multiple folds or subsets, using K-fold as a validation set, and training the model on remaining fold. Finally, the result from each validation step is averaged to obtain a more robust estimate of model's performance.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

Number of CV Scores used in Average	Cross Validation Scores	Average CV Score
3	[0.92481203 0.90977444 0.68421053]	0.8395989974937343
5	[0.9 0.9375 0.95 0.9375 0.94936709]	0.9348734177215189
7	[0.89473684 0.9122807 0.96491228 0.92982456 0.9122807 0.98245614 0.92982456]	0.932330827067669
8	[0.9 0.9 0.92 0.96 0.88 0.92 0.98 0.93877551]	0.9248469387755102
10	[0.875 0.925 0.9 0.95 0.95 0.925 0.875 0.975 0.975 0.92307692]	0.9273076923076923

Table 7: Result of various K fold cross validation

3.4 Splitting of the Dataset

We must split the dataset into Training and Testing. Here, we have split the dataset in 75:25proportion. The training set is used for training the model and for testing the model test set is used.

3.5 Applying machine learning algorithms

Different machine learning algorithms were applied to the dataset to find out the finest prediction model. Here 8 types of algorithms are compared such as SVM, Naïve Bayes, Logistic regression, KNeighbour classifier, Stacking Classifier, XGBoosting classifier, Random Forest Classifier and proposed ensemble algorithm.

3.6 Evaluation Metrics

There are various evaluation matrices used to evaluate classification techniques in machine learning. Accuracy, Precision, Recall and F-Measure.

- Accuracy is the proportion of all classifications that were correct, whether positive or negative.
  Accuracy = (TP+TN)/(TP+TN+FP+FN)
  Where TP is true positive, TN is true negative, FP is false positive and FN is false negative.
- Precision is the proportion of all the model's positive classifications that are actually positive.
  Precision = TP/(TP+FP)
- Recall is the proportion of all actual positives that were classified correctly as positives.
  Recall = TP/(TP+FN)
- F-Measure is used to evaluate the performance of machine learning model.
  F-measure = (2*precision*recall)/ (Precision +Recall)

- Confusion Matrix

Table 8 shows statistical summary of primary Dataset 2

3.7 Visualization of Dataset 2:

In this fig.10 we show the relationship between region and economy.

	Attendance15161	Attendance15162	PendingE	PendingF	ReappearSummerTermRegistered	ReappearExamGiven	IsAbove75T1	IsAbove75T2	SatisfactoryTerms
Mean	88.19549	84.64912	1.899749	0.593985	0.125313	0.110276	0.922306	0.87218	0.837093
Median	91	88	0	0	0	0	1	1	1
Mode	94	93	0	0	0	0	1	1	1
Min	48.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Max.	100.000000	100.000000	15.000000	19.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	83.000000	79.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.00000
50%	91.000000	88.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
75%	95.000000	94.000000	3.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
StdDev	9.445792	14.53715	2.932229	2.26411	0.331489	0.313626	0.268026	0.334308	0.369745
Variiance	89.22299	211.3288	8.597965	5.126195	.109885	0.098361	0.071838	0.111762	0.136711
ANOVA (F-Value)	72.750419	88.254664	470.249991	70.649378	84.759188	63.533528	56.756803	59.921677	95.507641

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

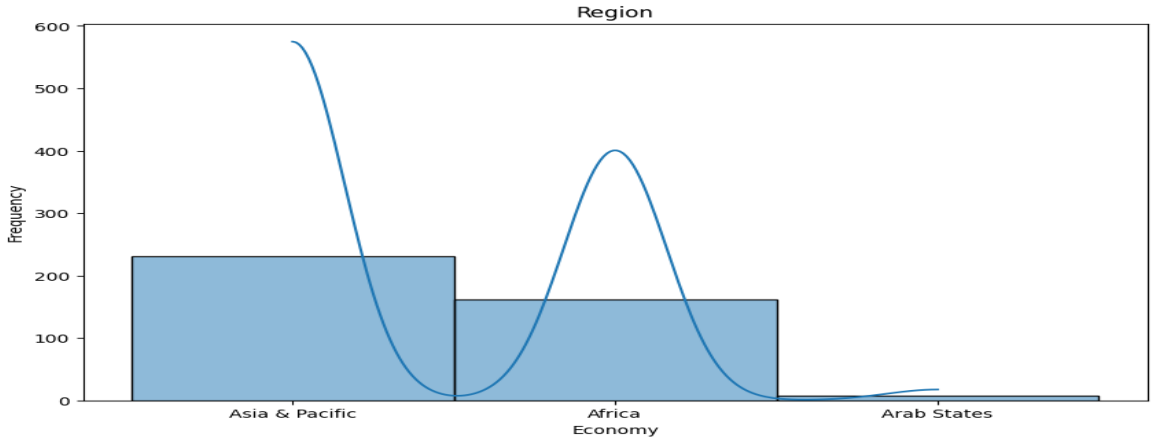


Fig 10: Region Vs Economy

In fig 11 we have represented the relationship between attendance and performance of the students. The students who secured total_attendance between 150-200, their performance0 is 267 and performance1 is 78 whereas students who secured attendance between 100-150, their performance0 is 10 and performance1 is 39.

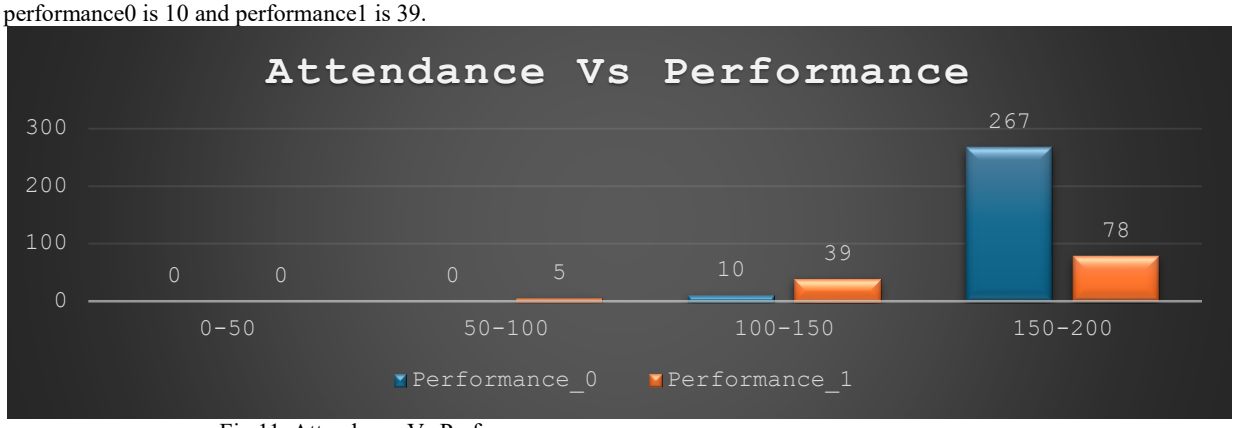


Fig 11: Attendance Vs Performance

Fig.12 represented 349 students did not registered themselves into ReaaperSummerTerm whereas 50 students registered for the same.

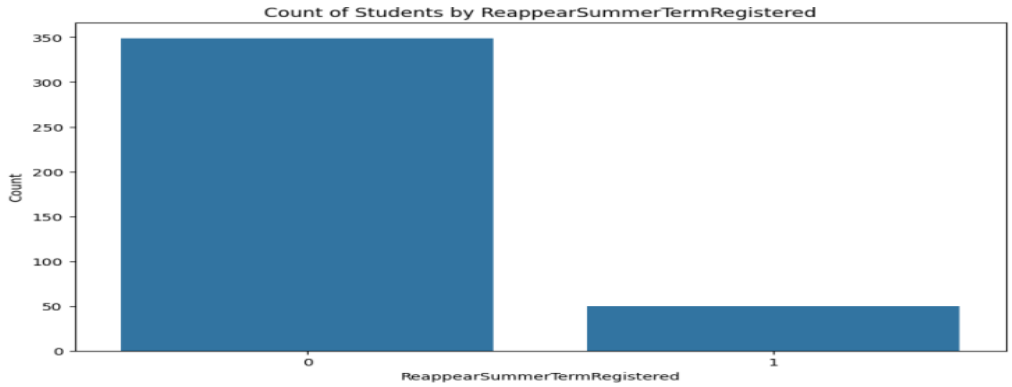


Fig12: No. of Students Registered in ReappearSummerTerm

In Fig.13 hierarchical clustering dendrogram with complete linkage is shown. It is used to represent to store each step as a memory that the Hierarchical cluster algorithm performs. In the dendrogram, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset. Datapoint 4 and 16 are combined together to form a cluster (say c1). Datapoint 7 and 4 are combined together to make new cluster (say c2). Now, cluster c2 and datapoint 5 combined together to form new cluster (say c3) as height of this cluster is higher than previous cluster
(c1 and c2) as the Euclidean distance between datapoint 5 and c2 is a little bit greater than c1 and c2. This process continues until the final dendrogram is created that combines all the points together.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

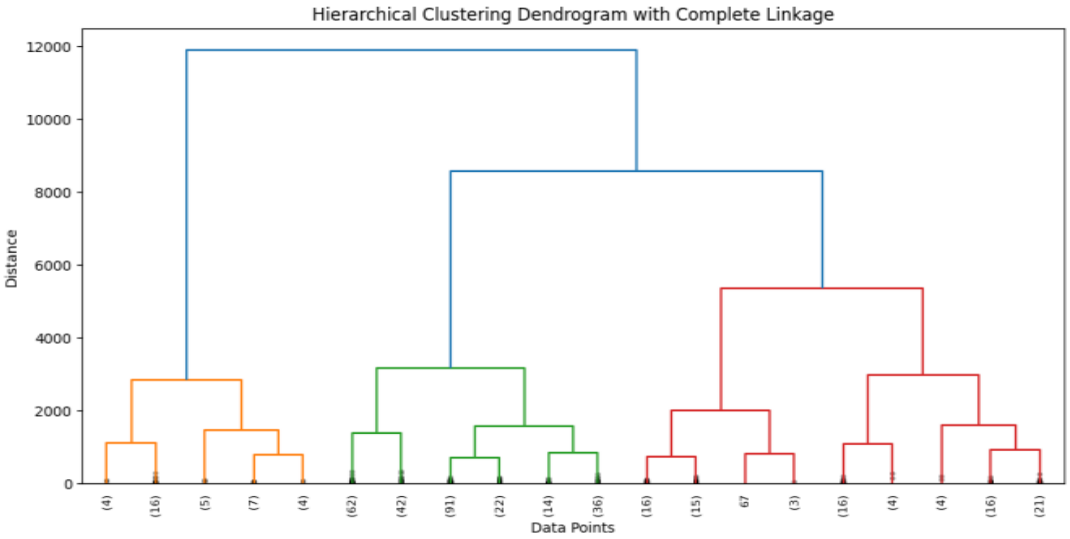


Fig 13:Hierarchical Clustering Dendogram with complete Linkage

Fig 14 shows the different types of clusters are formed according to the relationship between attendance of students and their performance.

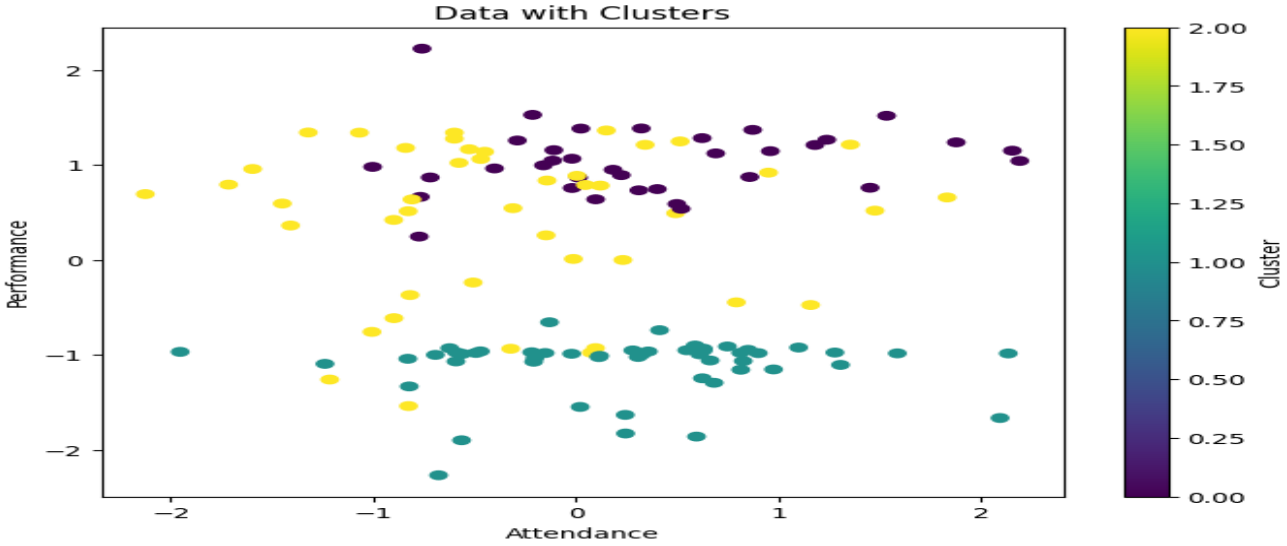


Fig.14: Clusters of Attendance Vs Performance

RESULTS: Confusion matrices for different algorithms:

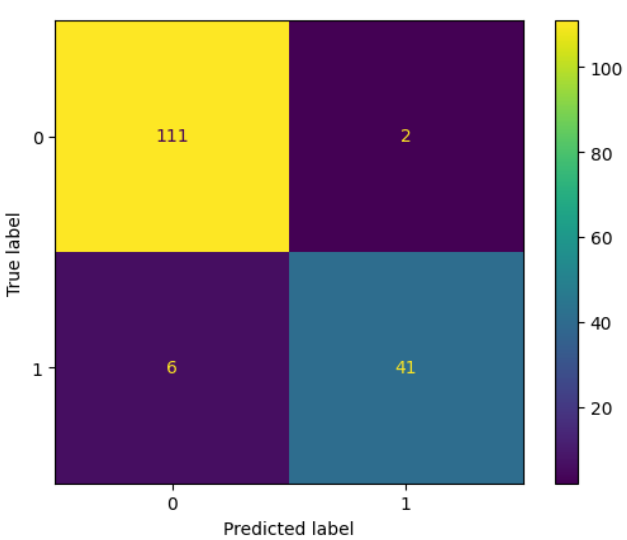


Fig15 (a): Confusion Matrix (SVM)

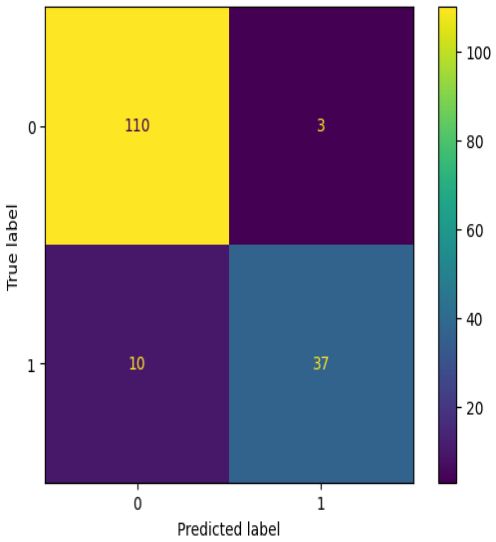


Fig15(b) : Confusion Matrix (Naïve Bayes)

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

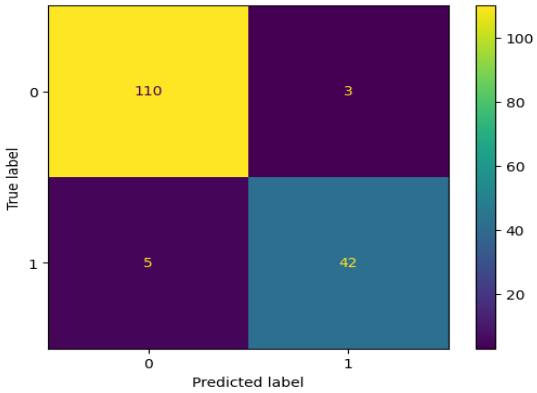


Fig15(c): Confusion Matrix (KNClassfier)

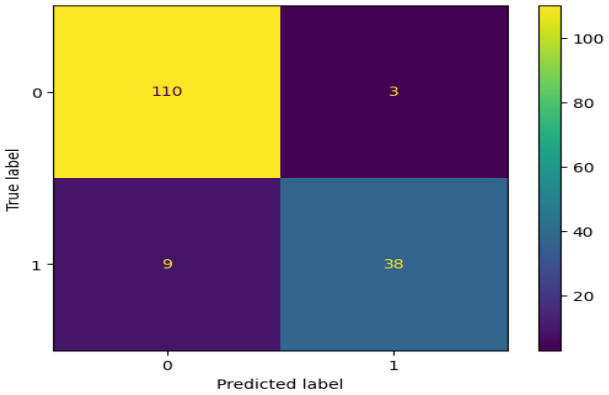


Fig.15(d): Confusion Matrix (Random Forest)

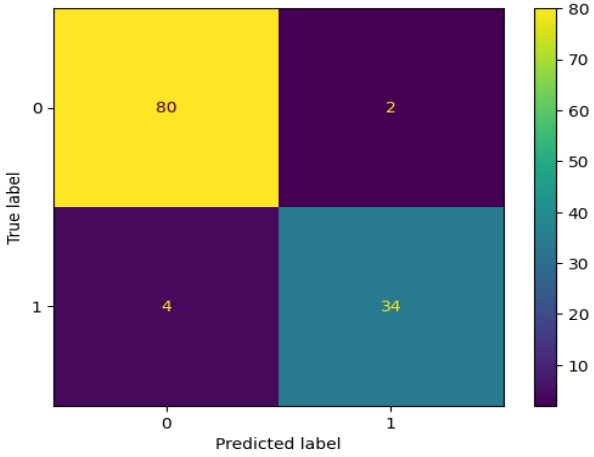


Fig15(e) : Confusion Matrix (Blending)

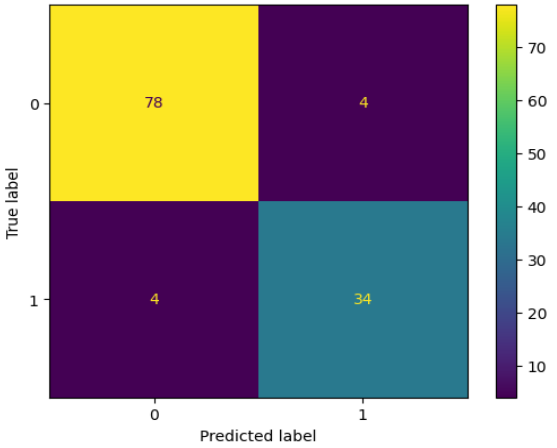


Fig15(f) : Confusion Matrix (Boosting)

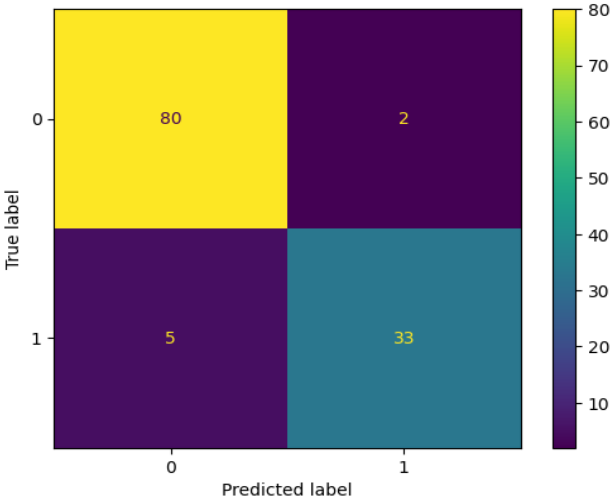


Fig. 15(g): Confusion Matrix (Stacking)

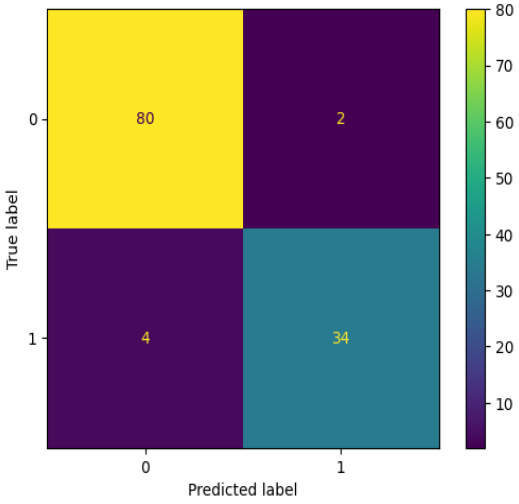


Fig.15(h) : Confusion Matrix (Bagging)

Table 9: Result of different machine learning algorithms on dataset2

Algorithms Used	**Accuracy**	**Precision**	**Recall**	**F1_Score**
SVM	0.95	0.9534	0.8723	0.9506
Naïve Bayes	0.9187	0.925	0.7872	0.9207
Logistic Regression	0.9583	0.9459	0.921	0.9584
KNeighbour Classifier	0.925	0.9268	0.8085	0.9265
Stacking Classifier(stacking)	0.9416	0.9428	0.8684	0.9423
XGBoosting Classifier (Boosting)	0.9333	0.8947	0.8947	0.9333
Random Forest Classifier (Bagging)	0.95	0.9444	0.8947	0.9503
Proposed Ensemble Classifier - LRN	**0.96**	**0.9548**	**0.922**	**0.9604**

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

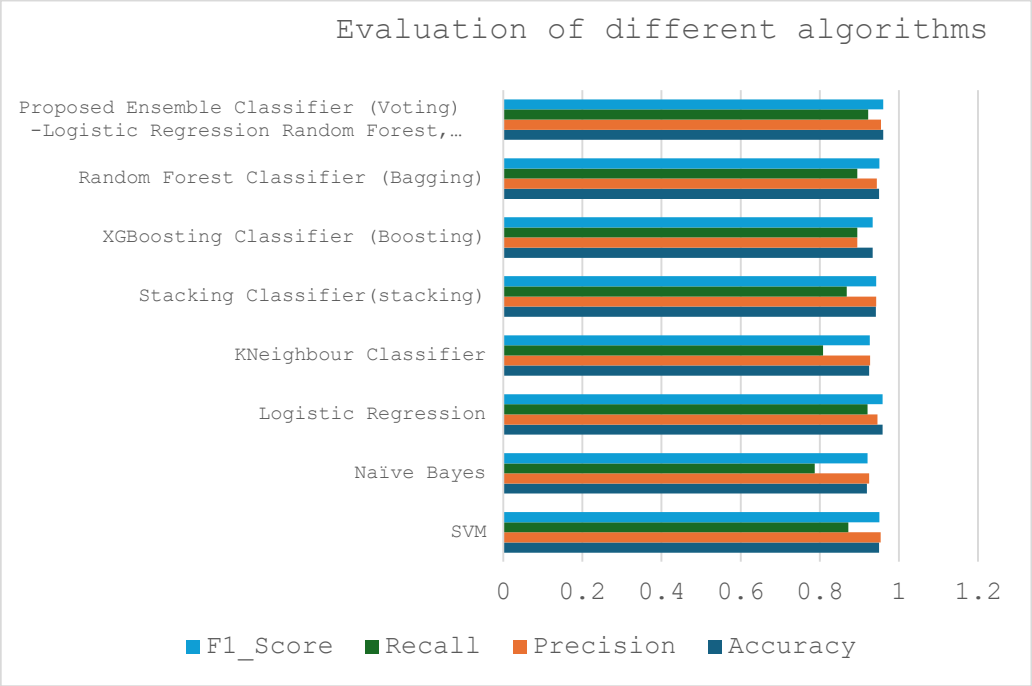
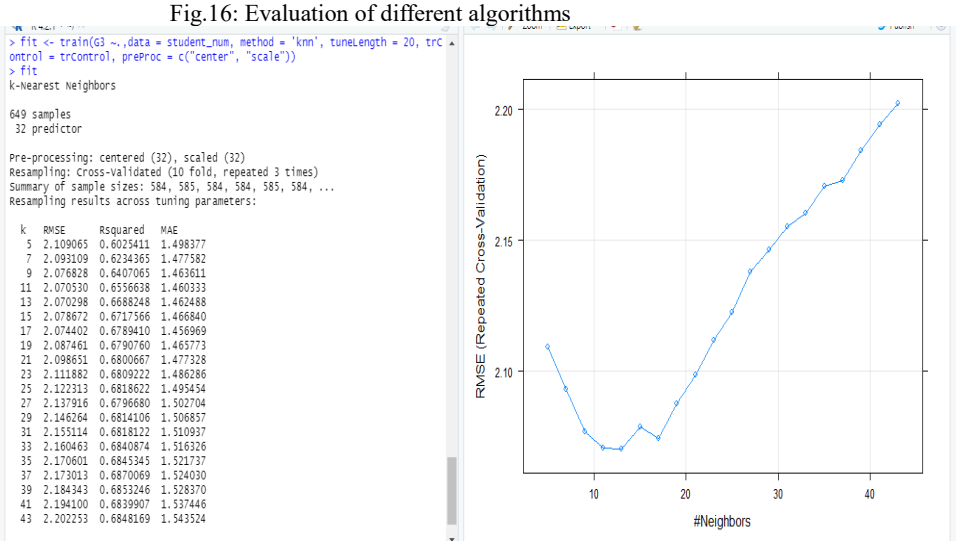


Fig.16: Evaluation of different algorithms



```
> fit <- train(G3 ~.,data = student_num, method = 'knn', tuneLength = 20, trC
ontrol = trControl, preProc = c("center", "scale"))
> fit
k-Nearest Neighbors

649 samples
 32 predictor

Pre-processing: centered (32), scaled (32)
Resampling: Cross-validated (10 fold, repeated 3 times)
Summary of sample sizes: 584, 585, 584, 584, 585, 584, ...
Resampling results across tuning parameters:

  k   RMSE      Rsquared   MAE
   5  2.109065  0.6025411  1.498377
   7  2.093109  0.6234365  1.477582
   9  2.076828  0.6407065  1.463611
  11  2.070530  0.6556638  1.460333
  13  2.070298  0.6688248  1.462488
  15  2.078672  0.6717566  1.466840
  17  2.074402  0.6789410  1.456969
  19  2.087461  0.6790760  1.465773
  21  2.098651  0.6800667  1.477328
  23  2.111882  0.6809222  1.486286
  25  2.122313  0.6818622  1.495454
  27  2.137916  0.6796680  1.502704
  29  2.146264  0.6814106  1.506857
  31  2.155114  0.6818122  1.510937
  33  2.160463  0.6840874  1.516326
  35  2.170601  0.6845345  1.521737
  37  2.173013  0.6870069  1.524030
  39  2.184343  0.6853246  1.528370
  41  2.194100  0.6839907  1.537446
  43  2.202253  0.6848169  1.543524
```


Fig. 17: Result of KNN Model

Figure 17 shows the result of KNN model where the value of k differs (i.e. k=5,7,9,11….41,43).The value of RMSE , Rsquared and MAE also varies. The value of RMSE=2.070298 is lowest when k=13,and RMSE =2.202253 at k=43

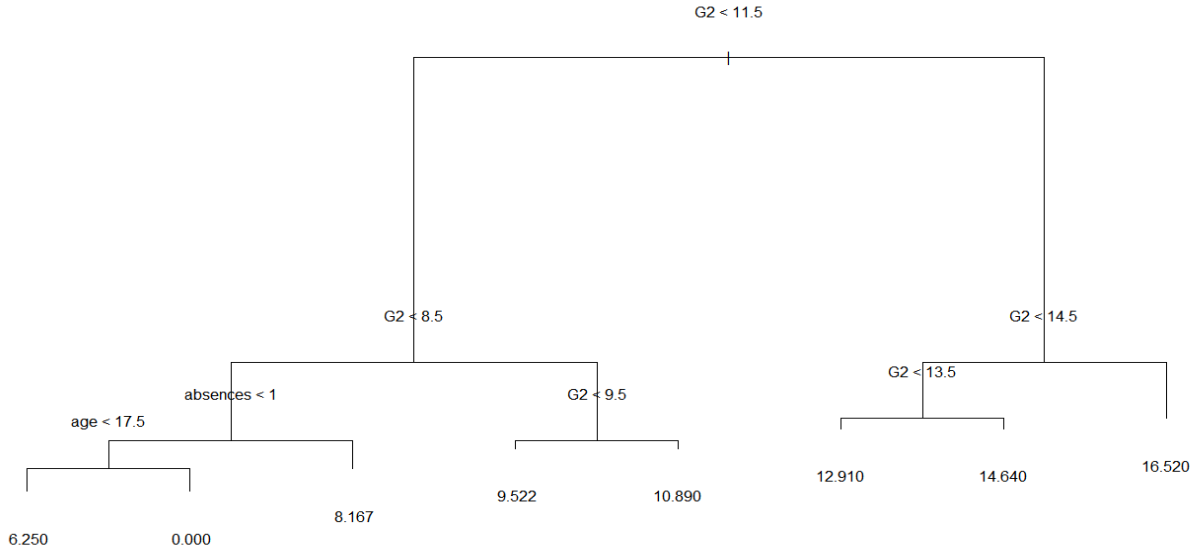


Fig. 18: Visualization of Decision Tree Regression (online dataset, nodes=8)

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

Fig.18 represents Decision Tree Regression with nodes=8. Depending upon the values of G2, the decision tree was created.

Table 10: Result of different machine learning algorithms on online dataset

Algorithm Used	Accuracy	Precision	Recall	F-Measure
Naive Bayes	88.7%	62.7%	74.2%	67.1%
Logistic Regression	92.5%	72.8%	85.5%	76.7%
Deep Learning	90.3%	81.04%	53.1%	62.5%
Decision Tree	77.9%	41.5%	100%	58.2%
Random Forest	91.4%	69.1%	76.8%	72.0%
Support Vector Machine	91.4%.	68.9%	74.5%	70.6%
Gradient Booted Tree	94.5%	**84.3%**	78.8%	80.0%
Proposed Ensemble Classifier - LRN	**95.5%**	83.3%	**96.5%**	**90.2%**

In the current study, online dataset from UCI repository was used in research and despite the novelty of this topic, real time dataset is to be collected. We show in our experiment that our proposed ensemble classifier-LRN gave highest accuracy 95.5% whereas Gradient Booted Tree shows the better outcomes as shown in Table 10.

4. CONCLUSION

Prediction of students' performance is considering the wide competition, an accurate decision is needed to be taken by the institution to increase the students' performance. We proposes LRN approach for prediction of students' performance in the continuation [37]. One of the preprocessing requirement in developing advanced methods is feature selection, as they help in reducing the size of dataset as well as at the same time improves accuracy and execution time of the algorithms. In this paper,firstly we selected the attributes by applying different types of feature selection techniques. We found that student'status (active participation or not) , class_attendance, PendingExams are major factors that affected the students'performance. As discussed in this study, various methods, including AI, machine learning, data mining and IoT, are being explored by researchers to predict student performance. However, there is a persistent need for improved results in this domain.Our proposed model (LRN) displayed a higher accuracy rate 96% in predicting performance of students in this study as compared with other algorithms such as SVM, NB, Knn, XgBoosting etc.

References:

[1] L.Ji.X.Zhang, L. Zhang,"Research on the Algorithm of Education Data Mining Based on Big Data",in*2020 IEEE 2nd International Conference on Computer Science and Educational Informatization(CSEI),* 2020,pp. 344-350.

[2] F.Y. Osisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms:classification and comparison, "*International Journal of Computer Trends and Technology(IJCTT)*, 2017,vol.48, no. 3, pp.-128-138.

[3] J.T. Avella, M. Kebritchi,S.G.Nunn, and T. Kanai, "Learning analytics methods, benefits, and challenges in higher education: A systematic literature review.,"*Online Learning*, 2016,vol. 20, no. 2, pp. 13-29.

[4] Luan J.,"Data Mining and Its Applications in Higher Education", *New Directions for Institutional Research*, 2002, 113, pp.17–36.

[5] Minaei-Bidgoli B.; Kashy D.; Kortemeyer G.; and Punch W., "Predicting student performance: an application of data mining methods with an educational web-based system". 2003 ,*In Proc. of IEEE Frontiers in Education*. Colorado, USA, pp.13–18.

[6] WangD, Zhang Z A hybrid System with filter approach and multiple population Genetic Algorithm for feature selection in Credit Scoring., science direct,*Journal of Computation and applied mathematics*
http://dx.doi.org/10.1016/j.cam.2017.04.036

[7] Wan Maziah Wan Ab Razak , Zulhamri Abdullah et al. 2nd ICIEBP: The 2nd International Conference on Islamic Economics, Business, and Philanthropy (ICIEBP) Theme: "Sustainability and Socio Economic Growth" Volume 2019 .
DOI: 10.18502/kss.v3i13.4285

[8] A. M. Shahiri, W. Husain, and N. A. Rashid, ``A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414_422, Jan. 2015.

[9] K.P.Shaleena and S. Paul, "Data mining techniques for predicting student performance," in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.

[10] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica,T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, ``Predicting academic performance: A systematic literature review," in Proc.Companion 23rd Annu. ACM Conf. Innov. Technol. Comput. Sci. Edu.,2018, pp. 175_199.

[11] S.N.U. Manzoor, " An agent based system for activity monitoring on network-ABSAMN", *Expert System Application*, 2009,vol 8, no.36, pp-10987-10944.

[12] Bravo Agapito. J. Romero.S.J.,& Pamplona. S.:Early prediction of Undergraduate Student's Academic performance in completely online learning: A Five year study",*Computers in Human Behavior*,2020 106595.https://doi.org/10.1016/j.chb.2020.106595

[13] Hamsa. H. Indiradevi.S. & Kizhakkethottam.J.J. "Student academic performance prediction model using decision tree and fuzzy genetic algorithm",*Procedia Technology*,2016,25.326-332.

[14] Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J., "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", *Computers in Human Behavior*, 2017,vol 73, pp 247-256.

[15] Mustafa Yağcı Educational data mining: prediction of students' academic performance using machine learning algorithms, *Smart Learning Environments*,2022, 9:11 https://doi.org/10.1186/s40561-022-00192-z

[16] Pooja and Bhalla R: "A Review Paper on the Role of Sentiment Analysis in Quality Education",*SN Computer Science*,2022, 3:469 https://doi.org/10.1007/s42979-022-01366-9

[17] J. Xu, K. H. Moon, and M. Van Der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.

[18] Leila Ismail, Huned Materwala, Alain Hennebelle, "Comparative Analysis Of MachineLearning Models For Students' Performance Prediction", https://www.researchgate.net/publication/350057919, 2021.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 190-203

[19] Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde, "Prediction Of Students Performance Using Educational Data Mining", ResearchGate, 2020
[20]Ahmed, A., Nipa, F. A., Bhuyian, W. U., Mushfique, K. M., Shahin, K. I., Nguyen, H.-H., & Farid, D. M,"Students' performance prediction employing Decision Tree." CTU Journal of Innovation and Sustainable Development , 2024, 16(Special issue: ISDS), 42-51. https://doi.org/10.22144/ctujoisd.2024.321
[21] Hassan Zeineddine and Udo C. Braendle, "Enhancing Prediction of Student Success:Automated Machine Learning Approch", Article in Computers & Electrical Engineering,2021.
[22], S.; Khan, M.Q. "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning",*Ann. Data Sci*. 2021.
[23] F. Su, D. Zou, L. Wang, and L. Kohnke, "Student engagement and teaching presence in blended learning and emergency remote teaching," *Journal of Computers in Education*, Jun 2024,vol. 11, no. 2, pp. 445–470 https://doi.org/10.1007/s40692-023-00263-1.
[24] Ahmed, A., Nipa, F. A., Bhuyian, W. U., Mushfique, K. M., Shahin, K. I., Nguyen, H.-H., & Farid, D. M,"Students' performance prediction employing Decision Tree." *CTU Journal of Innovation and Sustainable Development* , 2024, 16(Special issue: ISDS), 42-51. https://doi.org/10.22144/ctujoisd.2024.321
[25] Jawad, K.; Shah, M.A.; Tahir, M. "Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing", *Sustainability* (2022), 14, 14795.
[26] Bujang, S.D.A.; Selamat, A.; Ibrahim, R.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H.; Ghani, N.A.M. "Multiclass prediction model for student grade prediction using machine learning", *IEEE Access* ,2021, 9, 95608–95621.
[27] F. Sarker, T. Tiropanis, and H. C. Davis, "Linked data, data mining and external open data for better prediction of at-risk students," in Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014, 2014.
[28] E. S. Bhutto, I. F. Siddiqui, Q. A. Arain, and M. Anwar,"Predicting students' academic performance through supervised machine learning," in *Proceedings of the 2020 International Conference on Information Science and Communication Technology (ICISCT)*, pp. 1–6, Karachi, Pakistan, April 2020.
[29]H.H. Htun, M. Biehl, N. Petkov,"Survey of feature selection and extraction techniques for stock market prediction",*Financial Innovation,* 2023,9 (1) ,pp. 26
[30] Liu, H. &Setiono R.(1995) Chi2 : feature selection and discretization of the numeric attributes. In Anon(Ed.), Proceedings of the International Conference on Tools with Artificial Intelligence (pp 388-391),IEEE
[31] Cortes, C. &Vapnik, V. Machine Learning (1995) 20: 273. https://doi.org/10.1023/A:1022627411411
[32] Carrizosa, E & Morales, DR 2013, 'Supervised Classification and Mathematical Optimization Computers & Operations Research, vol 40, no. 1, pp. 150–165.
[33] Breiman, L. (2001). Random forests. Machine Learning, 45(1),5–32, https://doi.org/10.1023/A:1010933404324
[34] Sheena M.G., Missy S.M., Frinvie A., Jerald C. M.International Journal of Novel Research in Education and Learning Vol. 6, Issue 1, pp: (21-34), Month: January - February 2019, **www.noveltyjournals.com**
[35] Tadese et al. *BMC Medical Education (2022),* https://doi.org/10.1186/s12909-022-03461-0
[36] Kaur H.,Kaur T.,Bhardwaj V.,Kumar M.;An ensemble deep learning model for classification of students as weak and strong learners via multiparametric analysis, *Discover Applied Sciences,*Nov. 2024.
[37] Pooja and R. Bhalla, "Predicting the Students' Performance Using Machine Learning Tools- A Survey on Jalandhar City (Punjab)," *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, Greater Noida, India, 2025, pp. 871-876, doi: 10.1109/IC3ECSBHI63591.2025.10991068.

Authors

	Dr Rajni Bhalla is Professor in Department of Computer Application at Lovely Professional University. Her research interest includes data mining, data analysis, feature extraction, prediction, machine learning and clustering techniques. She can be contacted at email: rajni.b27@mail.com
	Pooja,Research Scholar from Lovely Professional University, Phagwara, Punjab(India). Her research area is Machine learning, Data mining, prediction techniques. She can be contacted at email:Poojagori51@gmail.com