Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 3 (2024)

Enhancing Predictive Accuracy in Education: A Detailed Analysis of Student Performance Using Machine Learning Models

Vratika Gupta[1], Dr. Priyank Singhal[2], Dr. Vipin Khattri[3]

[1]Research Scholar, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, India,

[2]Associate Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, India,

[3]Associate Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, India

Abstract:- Predicting student performance is crucial for enhancing educational outcomes and providing targeted interventions. This study employs various machine learning models to comprehensively analyze student performance and enhance predictive accuracy in educational settings. It addresses the challenge of predicting student outcomes by considering diverse predictors such as demographic details, academic history, and behavioral factors. The aim of this research is to develop robust models capable of generalizing across different educational environments and providing actionable insights into factors influencing academic achievement. The research utilized various approaches, including Random Forest, Linear Regression, Decision Tree, Support Vector Machine, Logistic Regression, XGboost, K-Nearest Neighbor, and Multilayer Perceptron Classifier. Results revealed significant outperformance of these traditional statistical methods. Attendance, previous academic performance, and engagement level emerge as critical predictors of success. Using the Student Attendance Dataset from Kaggle, Logistic Regression achieved an accuracy of 98%, Random Forest 100%, Multilayer Perceptron 100%, and XGBoost 96%. Findings underscore the potential of machine learning models as proactive tools for educators to effectively address student needs. The research highlights the transformative impact of machine learning in educational assessment and intervention strategies, advocating for a more data-driven approach in education. Future research may refine models by incorporating additional features, ultimately offering educators effective tools for supporting students and enhancing academic outcomes.

Keywords*: Attendance, Classification, Logistic Regression, Machine learning, Random Forest, Student performance.*

1. Introduction

Learning management systems, test data, student activities, library systems, etc., produce massive amounts of student information in today's secondary and tertiary education institutions [1]. Therefore, the quantity and diversity of educational data gathered by all institutions are increasing. For a long time, educational data has been subject to predictions of student performance using Data Mining (DM), Learning Analytics (LA), and Machine Learning (ML). These methods have demonstrated the utility of various algorithms and methodologies for delving into this area, which is difficult for humans to access. Colleges and universities, as well as secondary schools, now use student achievement as a key performance indicator. Students' grades are the most important factor in deciding their future careers at universities [2]. Their academic performance can assure the trustworthiness and quality of a candidate.

Collecting and processing information that connects with the value to be predicted is essential for building and applying a predictive model. Several factors influence students' performance; these include their prior education, their use of online learning, their demographics, the information about their social networks, variables pertaining

to their behaviour, evaluations from outside sources, the relationship between students' extracurricular activities and their academic achievement, the layout of their school, the level of parental involvement, and many more [3-8]. This study evaluates and categorizes student performance projections provided by researchers for high schools, colleges, and universities based on their significant traits and methodologies. With an abundance of data stored in educational systems, accurately predicting students' performance has become an enormous challenge [9]. Students have different learning interests, and online courses can support those who prefer to learn independently.

On the other hand, some students may benefit from having teachers tailor their lessons to their specific requirements. However, with access to all the resources offered by educational institutions, some students may do well on their exams [10]. Families with more disposable income might be able to look for other ways to pay for their children's supplementary education. Because of the wide range of topics covered, the authors set out to determine which characteristics have the greatest impact on students' academic performance and which machine-learning methods most useful in assisting school leaders in creating optimal learning environments. On the flip side, teachers have the power to shape lesson plans according to what their students are interested in and guide their students' learning journeys.

Managers have challenges in educational institution administration due to the complex data structure, multiple sources, and vast data quantity. While overseeing instructional processes, educational institutions encounter an abundance of additional administrative, monetary, and educational challenges. In order to help decision-makers with educational process management and coordination, it is necessary to examine all of these issues and come up with recommendations and findings [11, 12].

Educational data mining, or EDM, is a type of data mining that is used in schools and universities. It is theoretically oriented and seeks to help improve the academic performance of these institutions' students, graduates, and professors by developing computational methodologies that integrate theory and data [13,14]. EDM employs a variety of methods, including Logistic Regression, K-Nearest Neighbour (KNN), Decision Trees (DTs), Random Forest, Naïve Bayes (NB), Association Rule Mining, and Neural Networks. Using these strategies, one can uncover various forms of knowledge, including association rules, grouping, classifications, and prediction. Educational institutions can benefit from EDM. By employing Educational Data Mining (EDM), colleges can accurately predict the academic performance of students on tests, identify those who excel or struggle in specific areas, and determine the overall graduation rate. Afterwards, these schools can work on their educational practices to better assist students who have not been successful in the past to either continue their studies or find a concentration that fits their interests, skills, and background. Institutions can improve their academic performance with the help of EDM's improved rules and measures [15,16].

EDM is a popular place to look for studies that use DM and ML to investigate educational data [17]. The goal of machine learning research is to develop smart data that can automatically detect complicated hidden patterns and inform decisions [18, 19]. A unique DM process, the predictive route, uses current data to make predictions [20]. In the machine learning technique, getting a computer to adapt the action is the major focus, and this focus can improve the accuracy of certain actions or experiences.

Logistic Regression is one of the commonly used EDM method to predict the student performance that analyze and explain a binary variable, such as "pass" or "failed", along with a group of expected variables [21]. The goal is to identify the most appropriate model for explaining the connection between the sets of dependent and independent variables. Logistic Regression and Linear Regression are both built on top of each other, but they differ in how they handle responses with binary and continuous variables, respectively [22].

The most popular method for predicting student performance is Random Forest (RF), a supervised ensemble machine-learning approach for tasks like classification and regression. It works by building a number of decision trees during training and then producing the class output [23]. This is because supervised learning, which includes the classification approach, trains algorithms with less computing complexity using labelled

data. Moreover, the utilization of this labeled data in training procedures enables the models to acquire knowledge gradually and ultimately generate accurate predictions or data classifications [24].

The need for this study arises from the critical importance of accurately predicting student performance in educational contexts, which directly impacts academic success and career growth. With institutions increasingly leveraging data-driven approaches, there is a pressing demand for robust models that can utilize student attendance data to forecast academic outcomes effectively. This study addresses this need by evaluating various machine learning algorithms, such as Logistic Regression and Random Forest, to determine which techniques offer the highest accuracy in predicting performance. Such predictive models are essential for educators and administrators to identify at-risk students early, tailor interventions, and ultimately enhance the educational experience and success rates.

The scope of this study encompasses the analysis and prediction of student performance using machine learning algorithms applied to attendance data. By focusing on a dataset collected from graduate students and integrating various classification techniques such as Random Forest, Logistic Regression, and others, the research aims to identify the most effective models for predicting academic success. The study includes data preprocessing, implementation of multiple machine learning models, and comparison of their accuracies. The findings are intended to provide actionable insights for educational institutions to enhance student engagement, optimize resource allocation, and improve overall academic outcomes through targeted interventions.

Section 1 provides an overview of the the effect of student attendance on academic performance. Section 2 provides an overview of the past studies that have been published in the field of academic performance. Section 3 introduces the proposed methedology for analysis of student performance. Section 4 presents the primary results of the research and offers a comprehensive analysis and explanation of those outcomes within the wider context of the area. Section 5 contains the conclusion of the research.

2. Related Works

A number of researchers have conducted multiple research investigations to determine the effect of student attendance on academic performance.

Nakhipova et al., (2024) [25] generated and evaluated a strategy for assessing students' academic achievement using the Naive Bayes classifier. Researchers have developed a powerful instrument that can optimize and automate the evaluation of academic achievement by making use of state-of-the-art machine-learning techniques. The proposed strategy used a Python algorithm that used the Naive Bayes analysis method to forecast students' performance. Improvement and evaluation of performance were the key areas of emphasis in the suggested paradigm. In addition to providing a useful resource for schools and teachers, the results highlight the uniqueness of this method.

Yağcı et al., (2022) [26] suggested a novel model that utilizes machine learning techniques to forecast undergraduates' final exam scores using their midterm exam scores as input. Researchers computed and compared the performances of different machine learning algorithms to forecast the students' final exam scores. A total of 1,854 students' academic performance ratings from the 2019–2020 fall semester of a Turkish public university's Turkish Language-I course made up the dataset. The findings demonstrated a classification accuracy of 70–75% for the suggested model. Only three kinds of data were used to make the predictions: grades from midterms, data from the departments, and data from the faculty.

Error! Reference source not found.et al. (2022) [27] suggested a model for predicting and analyzing engineering students in college and university that utilizes the Least Square Method along with a New arctan − exp Logistic Regression Function. To meet the goals of alertness and prediction, this analysis approach allowed student managers as well as students to anticipate their learning conditions in advance. Analysis of the results offered strong evidence for faculty to properly modify educational management and investigate alternative methods for instruction.

Geetha et al., (2021) [28] evaluated the learner's development using individual and course-related information collected from the college's progress report. The dataset was utilized to make predictions utilizing the classification methods KNN, Support Vector Machine (SVM), and XGBoost, and the resulting accuracy was compared. According to the accuracy comparison table, the XGBoost classifier outperformed the other algorithms and was determined to be the best fit for the dataset.

GE et al., (2020) [29] evaluated student record data for use in training decision tree algorithm-generated classifiers to make predictions. Decision tree analysis for COS101 had the best overall accuracy of 96.7%, accurately predicting 95 out of 103 students' answers while mistakenly predicting 8 out of 103. But in case of STA172 it got 31 students wrong, the same system still managed a 71.91% accuracy rate in its predictions.

Hashim et al., (2020) [30] developed a model to forecast students' performance in final exams by analyzing data from courses offered by the Bachelor of Science in Computer Science and Information Technology program at the University of Basra during the 2017–2018 and 2018–2019 school years. The logistic regression classifier outperformed the others in terms of precisely predicting students' final grades (68.7% for passed as well as 88.8% for failed), according to the results.

Canagareddy et al., (2019) [31] established a machine-learning model for annual prediction of college students' performance. By using the model to predict students' performance, researchers can take action before it's too late. In order to train the suggested model, an existing student dataset was utilized, along with the classification technique. Students' attendance, grades, study time, health, and average performance were some of the criteria that can be utilized to forecast their success once the training phase has produced a training model.

Error! Reference source not found.) [32] defined student performance in Kenyan elementary education. The failure rate was so high and occurred so frequently that researchers created a model to help students improve their academic performance. Two data sets were defined that is: 2426 and 1105. The accuracy of the model was 80%. In the future, researchers would apply this concept to more nations and enhance it to work with mobile devices.

Error! Reference source not found.[33] defined the logistic regression model for predicting mathematics students' performance. Researchers addressed the issue of failure rates and the impact of poor marks on CGPA. After reviewing the document, the researchers concluded that students were disinterested in MTH101 courses and attended lectures outside of the class. The logistic regression model was also defined to predict the performance of students in the medical field to finish dropout and failure scores. The model was applied to 1205 data sets in medical students.

3. Research Methodology

This section details the methodology for developing a student performance model using the Student Attendance Dataset from Kaggle, including the techniques used and the proposed architecture.

3.1 Dataset

The dataset used in this study is obtained through secondary sources, which are available on the website of Kaggle and known as the Student Attendance Data set. The data set has been collected from schools and colleges with 2034 records. First, apply the data cleaning, remove all the missing values and change the data in string to integer values. The gender "male/female" in the original test data set is converted into a value of 0 or 1. Convert subject string values into integer values. Convert month values into string values. In this paper, the author selected 1603 boys and 431 girls who participated in the test as data sets, respectively. Data collection includes a total of ten parameters taken: Enrollment No, Month, Course, SEM, subject, Gender, Total lecture, lecture, percentage, and performance, as illustrated in Table 1.

Table 1: Attributes, Domain and Description

S.no	Attributes	Domain	Description
1	Enrollment No	TCA210101 –TCA21013266	Enrollment no helps for students to identify in college.
2	Month	Jan {0}	Month wise attendance in data set.
3	Course	BCA {0}	Students are enrolled in different courses.
4	Sem	{2nd sem}	Semester wise performance of students
5	Subject	{Digital logic subject}	Subject wise attendance in data set
6	Gender	{Male, Female} {0,1}	Students of, male and Female
7	Total Lecture	{30}	Total lectures in the semester
8	Take lecture	{0,1,2,3……30}	Take the lecture by students in the class
9	Percentage	{10%,20%,……100%}	The percentage of those students who take the lectures
10	Performance	{good, low, poor, below average, average, best}	Performance of students is good, poor, best and average

Head and Tail used in the data set: Table 2 shows the value of starting 5 rows of data using the Head () function of Python. Table 3 shows the Tail () is applied for the last five values. The size of the data set is mentioned in 2035 rows and 8 columns.

Table 2: Head applying in the data set

S.No	Enrollment No	Month	Course	Sem	Subject	Gender	Total Lecture	Present Lecture
1	TCA2101001	Jan	BCA	2	Digital Logic	Male	30	10
2	TCA2101002	Jan	BCA	2	Digital Logic	Male	30	10
3	TCA2101003	Jan	BCA	2	Digital Logic	Male	30	10
4	TCA2101004	Jan	BCA	2	Digital Logic	Male	30	10
5	TCA2101005	Jan	BCA	2	Digital Logic	Male	30	25

Table 3: Tail applying in the data set

S.No	Enrollment No	Month	Course	Sem	Subject	Gender	Total Lecture	Present Lecture

2029	TCA2103262	Jan	BCA	2	Digital Logic	Male	30	14
2030	TCA2103263	Jan	BCA	2	Digital Logic	Male	30	21
2031	TCA2103264	Jan	BCA	2	Digital Logic	Male	30	24
2032	TCA2103265	Jan	BCA	2	Digital Logic	Male	30	23
2033	TCA2103266	Jan	BCA	2	Digital Logic	Male	30	22

3.2 Techniques Used

There are various techniques used for prediction of student performances as discussed in this section.

A. Adaboost Classifier

Adaboost classifier is used for binary and multiclass classification problems. It is also called the boosting technique [34]. Decision trees with one level, or Decision trees with only one split, are the most popular estimators used with AdaBoost. Another name for these trees is Decision Stumps. The main purpose of AdaBoost is to convert weak learners to high learners [35].

Step 1: Calculate the sample weights for student data set.

$$w(xi, yi) = \frac{1}{n}, i = 1,2,3 \dots \dots . n \quad (1)$$

Step 2: To calculate the error of the model

$$Performance\ of\ stumps = \frac{1}{2} log \frac{1-Total\ Error}{Total\ Error} \quad (2)$$

The total error of the model is between 0 and 1. The performance of stump is representing by α.

Step 3: updated the weight for correct and incorrect classified points.

$$\text{For corrected classify weights} = weight * e^{-performance\ of\ stumps} \quad (3)$$

$$\text{For in corrected classify weights} = weight * e^{performance\ of\ stumps} \quad (4)$$

Step 4

$$\text{Final prediction} = \alpha_{1(o1)} + \alpha_{2(o1)} + \dots \dots \dots \dots \alpha_{2(on)} \quad (5)$$

Where,

- α (Alpha): Alpha represents the contribution of each decision stump to the final prediction.
- xi represents the input features or attributes of the dataset
- yi represents the corresponding class labels or targets associated with each input feature xi.
- n represents the total number of data points or samples in the dataset

B. Random Forest

Random forest is used in both regression and classification techniques. It is an algorithm of the bagging technique [36]. To apply the random forest algorithm, the student data is selected randomly. It combines student features for preprocessing [37]. The classifier uses a Decision Tree (DT) to choose random attributes. In the

model, a random forest algorithm is applied to predict the accuracy of the student dataset. The working procedure of RF algorithm in student academic performance:

1. Random Student features are selected from the student dataset
2. DT framing for each tuple is done, and results for each Tree are obtained
3. The level for each result is determined.
4. The result that is in lead place is considered as final.

C. Decision Tree

The decision tree is used for both classification and regression techniques. It is a tree-like structure [38]. In this paper, researchers use to decision tree to predict the performance of students. A decision tree classifier for student academic performance is a predictive model that uses features such as attendance, study habits, and test scores to classify students into different performance categories. By recursively splitting the data based on these features, the model creates a tree-like structure to make predictions about students' likelihood of achieving certain academic outcomes.

D. Logistic Regression

A logistic regression model is used as a classification model. It is a well-known supervised machine-learning algorithm [39]. The result is based on discrete or categorical values. The output is in the form of a 0 or 1. LR issues are very similar to regression problems, with the exception that regression problems are solved using continuous data, while logistic regression problems are solved using classification [40].

E. Multilayer Perceptron Classifier

MLP is a concept of supervised algorithms. It is based on the complex architecture of Aritifical Neural Network (ANN). The multilayer perceptron layer defines the one input layer, many hidden layers and the output layer. The hidden layer defines the activation function [41]. The input layer defines the gender, age, course, enrollment, and subject. The hidden layer is working to process the model. The output of the model is the student's performance.

F. XgBoost Classifier

XgBoost stands for Xtreme gradient boosting algorithm. It is used for regression and classification problems. It is a boosting technique and an ensemble learning method. In the XGBoost ensemble approach, fresh models are built to eliminate residuals or other mistakes from earlier models before being integrated to provide the final forecast. In comparison, XGBoost is faster than a lot of other ensemble classifiers (like AdaBoost). In this paper, researchers are using the xgboost classifier to predict the performance of students. The mathematical working of xgboost is also defined below.

The intuition of xgboost in the mathematical representation

- Construct the tree with root
- Calculate similarity weight $= \frac{\sum_{(Residual)}{}^{2}}{\sum prob(1-prob)+\lambda}$ (6)
- Calculate gain.

G. Support vector classifier

Support vector machine is used for classification and regression techniques. The support vector classifier works as a logistic regression. It worked for binary and multiclass classification [42]. Researchers are using a random basis kernel to predict the accuracy of the model. The support vector intuition is based on hyperplane, support vectors, and margin. The main purpose of using the SVM is to give the data in the best decision boundary & differentiate n space into classes so researchers can put new data points in the correct place. The hyperplane is also called the best decision boundary. It basically creates two margin lines and contains distance. It is helpful to

easily separate positive and negative values Radial basis function kernel is a general-purpose kernel. It is used when researchers have no prior knowledge about the data. The following equation defines the RBF kernel on two samples, x and y.

$$k(x, y) = exp\left[\frac{[x-y]^2}{2\sigma^2}\right] \quad (7)$$

H. K-Nearest Neighbor

KNN is used to solve both classification and regression problems. The simplest classification method is KNN. It is used to calculate the distance between points on a graph. To calculate the distance between neighbors, researchers are using the Euclidean distance formula [43]. In the implementation of the model, the value of k is 5. In this paper the author is defining the comparative study of KNN, SVM and Decision Tree to predict the performance of students [44]. To calculate the distance between neighbor's points, the equation in KNN is:

$$D= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (8)$$

3.3 Proposed Architecture

The first step of the proposed architecture represents the collection of student attendance datasets in schools and colleges. To collect the data set, apply the data cleaning and data mining techniques. There are no null values and missing values in the data set. The next step is to apply the analysis techniques using matplotlib, NumPy and seaborn libraries. In the next step, machine learning algorithms such as Random Forest, Linear Regression, Decision Tree, Support Vector Machine, Logistic Regression, XGboost, K-Nearest Neighbour, Multilayer Perceptron Classifier are used to best fit the dataset. There are two types of techniques: regression and classification. In the next step, prediction analysis is shown using the confusion matrix. The confusion matrix is defined in the result section. The below Figure 1 defines the structure of the model.

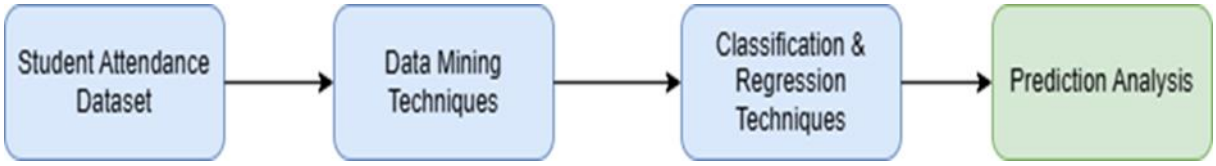

Figure 1: Framework of the model

4. Results and Discussion

This section presents the key findings of the research and provides a thorough analysis and interpretation of those results within the broader context of the field.

4.1 Data Description

Table 4 shows the data description defining the count, std, size, min and max of the data.

Table 1: Description of the parameters

Parameters	Sem	Total Lecture	Present Lecture
Count	2034.0	2034.0	2034.00000000
Mean	2.0	30.0	17.453786
Std.	0.0	0.0	5.995906
Min	2.0	30.0	1.000000
25%	2.0	30.0	15.000000
50%	2.0	30.0	20.000000

75%	2.0	30.0	21.000000
Max	2.0	30.0	27.000000

4.2 Count of Male & Female

The graph, as shown in Figure 2, is used to define the comparison between male and female students in the course. Male strength is more than female students. The count of males is 1603, and the count of females is 431. The data type of parameters is integer type.

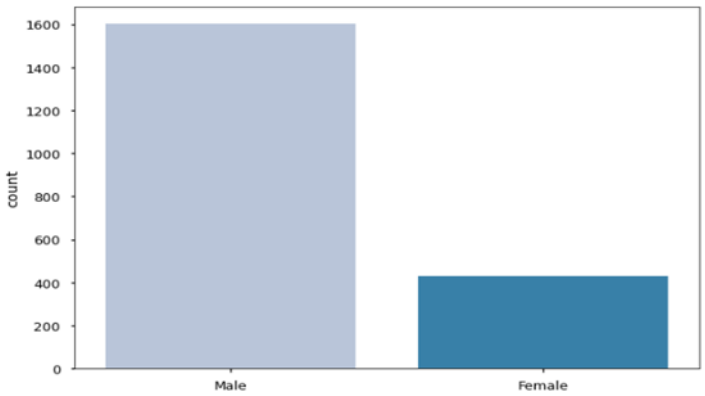

Figure 2: Count of Male & Female

After attending all of the lectures in the particular subjects, the author classified student performance into several levels. The essential purpose of any lecture is for students to understand the subject and attendance. If a student attends classes regularly and likes the lecturing style of the professors, their performance is good. If a student doesn't attend classes frequently and is unable to attend lectures, the result is skipping lectures. They don't attend lectures or miss classes. Therefore, it performs in the low and poor category. Table 5 defines the performance of students at different levels, the percentage of students and the count of students in different categories. Figure 3 defines the performance of students are poor, average, good, low, best, and below average.

Table 5: Performance, Percentage and Count of Students

Performance of students (Levels)	Percentage of students	Count of students
Best	>=90%	68
Average	>= 80%	146
Below average	>=70%	302
Good	>= 60%	729
Low	>= 40%	439
Poor	<=40 %	350

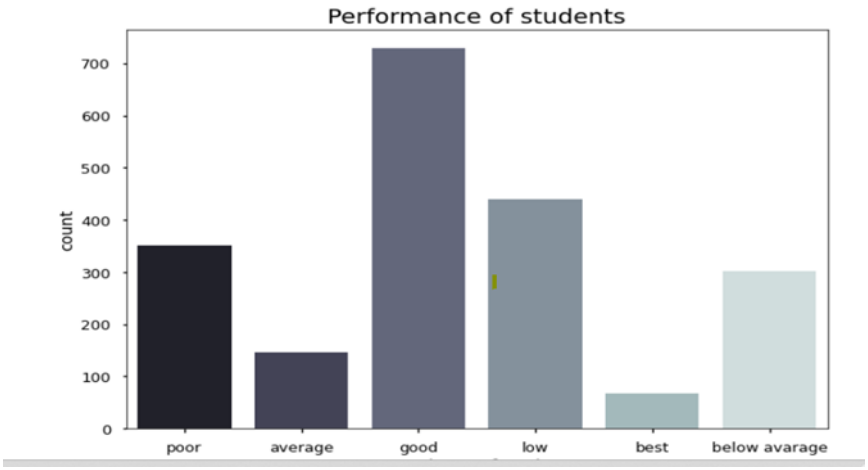

Figure 3: Performance of students using different parameters

4.3 Performance of male and female students

In today's time girls and boys both are more focused on study. Figure 4 represents the gender-wise performance of students. 1 represents the male, and gender 0 represents the female. In Figure 4 the poor performance comparison is male is high and female is low.

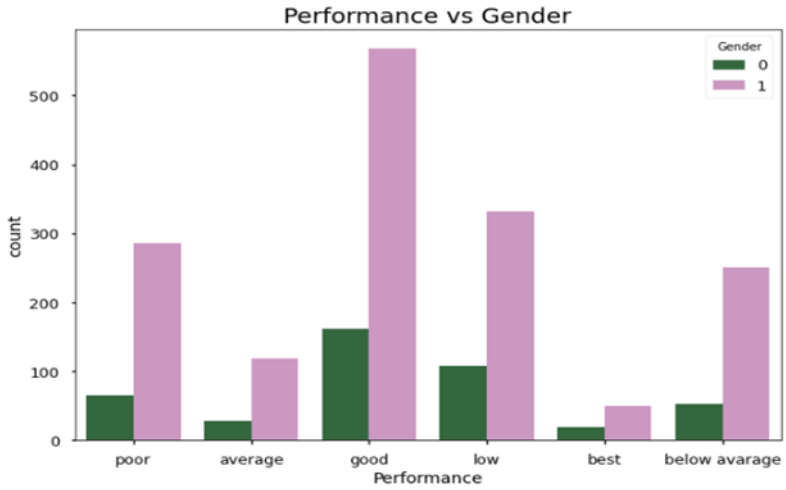

Figure 4: Gender vise Performance of students using different parameters

4.4 Confusion Matrix for Accuracy

A confusion matrix is a specialized table used to evaluate the effectiveness of an algorithm. Multiple parameters are utilized for both actual and predictive purposes. The outcome of the study is determined by utilizing a confusion matrix, namely Figure 5 to Figure 14. Various algorithms are employed for processing numerous data elements [0,1,2,3,4,5]. Based on the data received from the conducted experiments, it was determined that the KNN algorithm, namely the Random Forest variant, is the most accurate algorithm for classification.

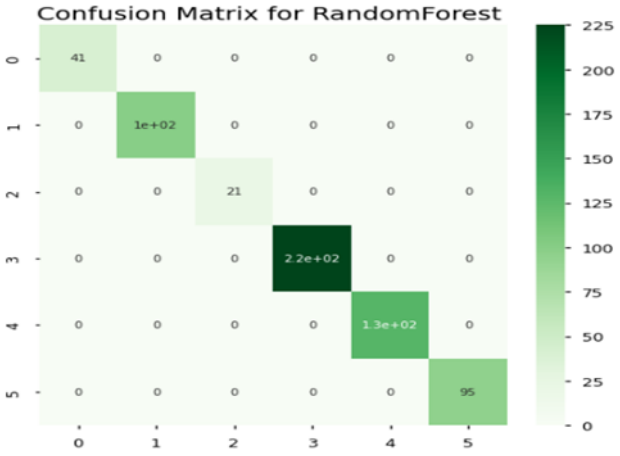

Figure 5: Confusion matrix (Random Forest)

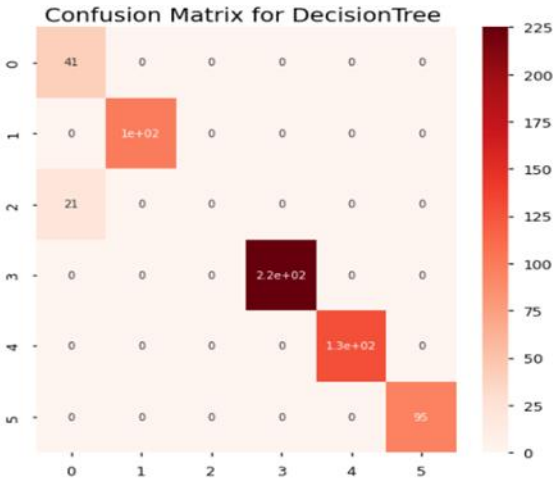

Figure 6: Confusion matrix (Decision tree)

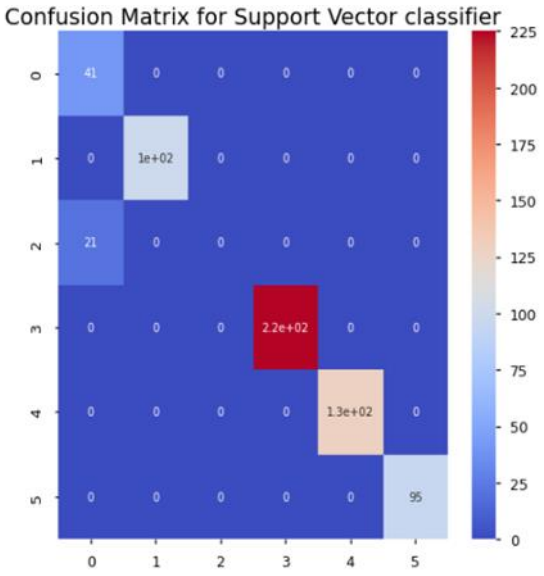

Figure 7: CM (Support Vector classifier)

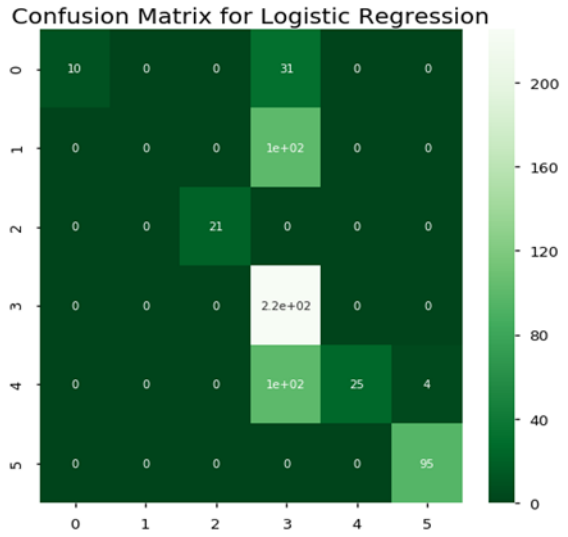

Figure 8: Confusion matrix (Logistic Regression)

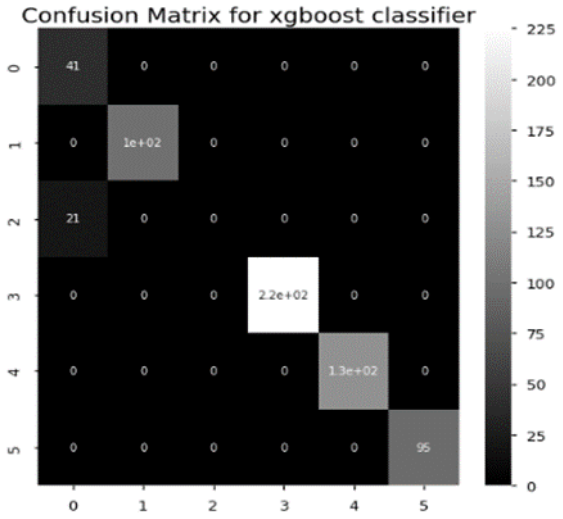

Figure 9: Confusion matrix (XGBoost Classifier)

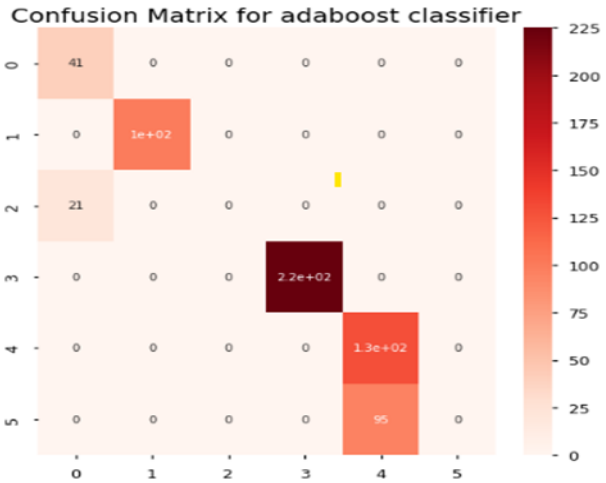

Figure 10: Confusion matrix (AdaBoost classifier)

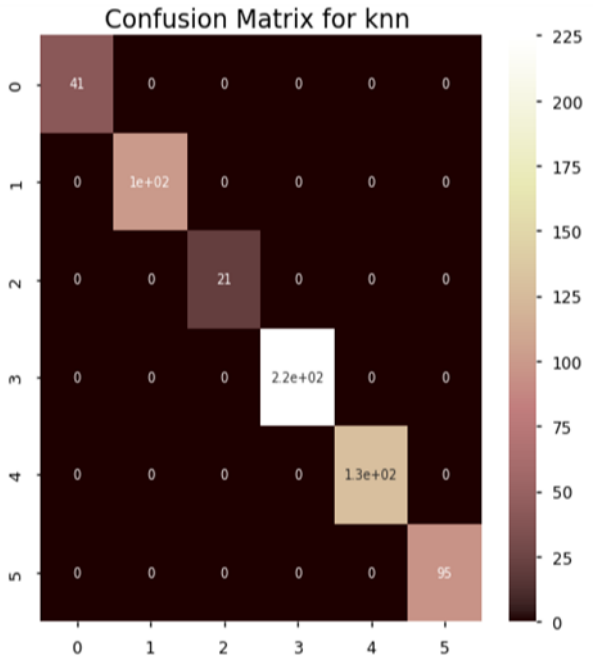

Figure 11: Confusion matrix (KNN)

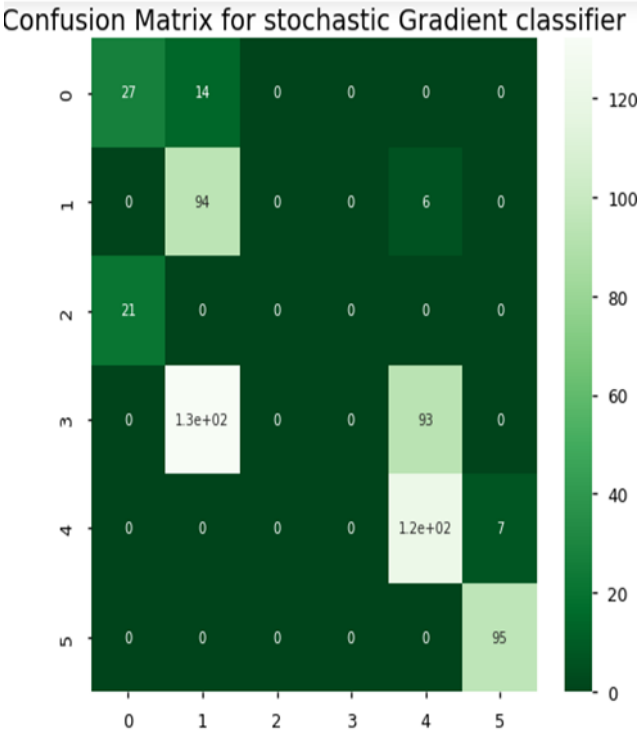

Figure 12: Confusion matrix (SGC)

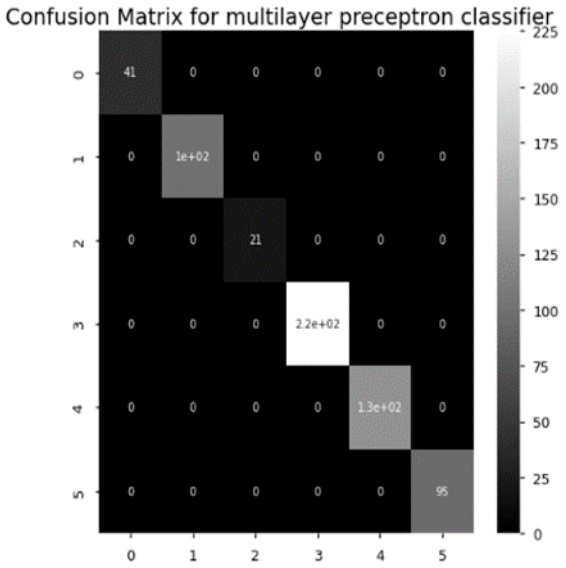

Figure 13: Confusion matrix (MPC)

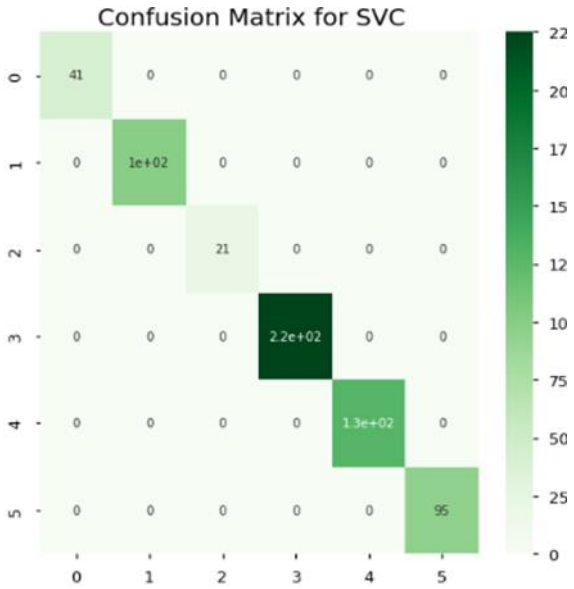

Figure 14: Confusion matrix (SVC)

4.5 Prediction using machine learning algorithms

In this model, authors have predicted the accuracy on the basis of the machine learning algorithm. The model accuracy and Obtained result accuracy is good. Linear regression is a part of supervised machine learning algorithms. It provides the output continuously. The accuracy of the linear regression is 80 %. It is shown in table 6. Logistic regression is used to classify data sets. It works on binary classification data sets. The accuracy of logistic regression is 100%. Random forest is used for both classifier and regression methods. It is a supervised learning algorithm. Random Forest Algorithm works on both continuous and categorical variables, it performs better in classification algorithms in categorization. The accuracy of random forest is 100%. Decision tree is a classification method. It's shaped like a tree structure. The accuracy of the decision tree is 96%.

Table 6: Accuracy of models, algorithms, Testing and Training Accuracy

Algorithm	Accuracy	Training Accuracy	Testing Accuracy
Random Forest	100%	1.0	1.0
Linear Regression	80%	1.0	1.0
Decision Tree Classifier	96%	1.0	1.0
Support Vector Classifier	96%	0.91	0.90
Logistic Regression	100%	1.0	1.0
XGboost classifier	96%	1.0	1.0
K-Nearest Neighbour	99%	1.0	1.0
Multilayer Perceptron Classifier	100%	1.0	1.0
Extra tree Classifier	100%	1.0	1.0
Adaboost classifier	81%	0.91	0.90
Stochastic Gradient Descent	55%	0.91	0.90

The graph, as shown in Figure 15, provides a visual comparison of the accuracy of various algorithms. The bars corresponding to each algorithm's accuracy clearly depict which algorithms perform better in terms of accuracy. Algorithms like Random Forest, Logistic Regression, Multilayer Perceptron Classifier, and Extra Tree Classifier show the highest accuracy, achieving 100%. In contrast, the Stochastic Gradient Descent algorithm shows the lowest accuracy at 55%. This visual representation helps in quickly identifying the performance effectiveness of each algorithm.

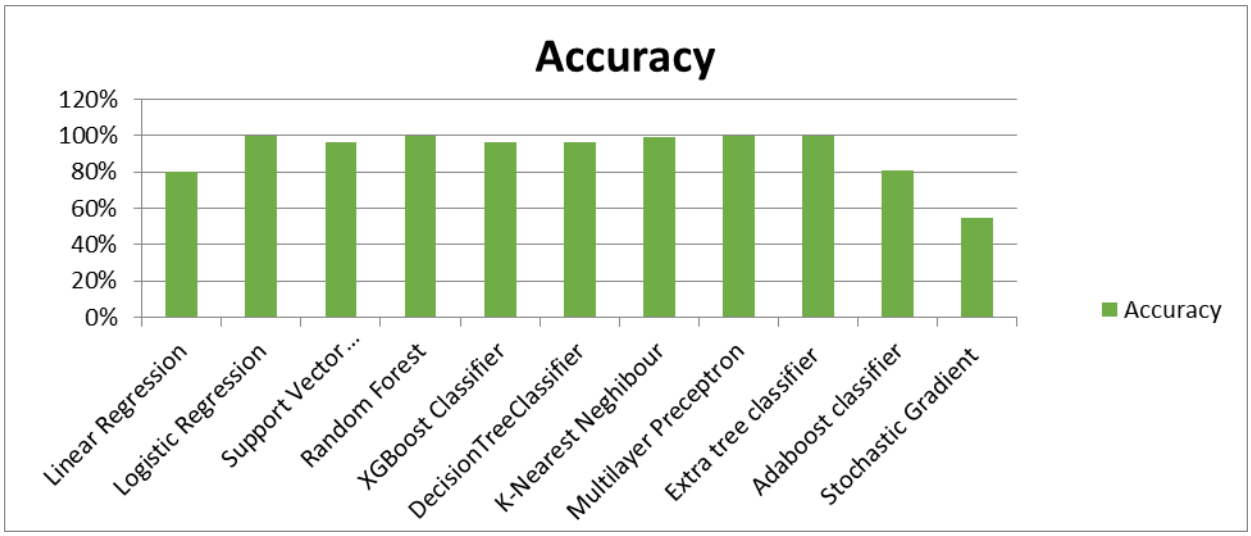

Figure 15: Accuracy comparison of different algorithms

5. Conclusion and Future Scope

This study addressed the critical issue of predicting student performance using machine learning algorithms based on attendance records. Various machine learning algorithms, including Logistic Regression, Random Forest, Multilayer Perceptron, and Extra Tree Classifier, were evaluated using a dataset of 2043 student attendance records. The methodology involved data collection, preprocessing, and applying multiple classification techniques to predict student performance. Logistic Regression achieved an accuracy of 98%, Random Forest 100%, Multilayer Perceptron 100%, and XGBoost 96%. The findings indicate that Multilayer

Perceptron and Random Forest outperformed other models in predicting student performance based on attendance data. This research demonstrates the potential of machine learning models to effectively forecast academic outcomes, providing valuable insights for educational institutions to identify at-risk students and tailor interventions. Future work can explore the integration of additional features, such as socio-economic factors and extracurricular activities, to further enhance the predictive accuracy and applicability of these models in diverse educational settings.

Refrences

[1] Gaftandzhieva S, Docheva M, Doneva R. A comprehensive approach to learning analytics in Bulgarian school education. Education and Information Technologies. 2021; 26(1):145-63.
[2] Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education. 2020; 17(1):1-21
[3] Ferreira SA, Andrade A. Academic analytics: Anatomy of an exploratory essay. Education and Information Technologies. 2016; 21(1):229-43.
[4] Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2015; 2(1):1-21.
[5] Khatib KC, Kamble TD, Chendake BR, Sonavane GN. Social media data mining for sentiment analysis. International Research Journal of Engineering and Technology. 2016; 3(04):373-6.
[6] Ozdemir D, Opseth HM, Taylor H. Leveraging learning analytics for student reflection and course evaluation. Journal of Applied Research in Higher Education. 2019; 12(1):27-37.
[7] Nuutila K, Tuominen H, Tapola A, Vainikainen MP, Niemivirta M. Consistency, longitudinal stability, and predictions of elementary school students' task interest, success expectancy, and performance in mathematics. Learning and Instruction. 2018; 56:73-83.
[8] Lang C, Siemens G, Wise A, Gasevic D. Handbook of learning analytics. New York, NY, USA: SOLAR, Society for Learning Analytics and Research; 2017.
[9] Nawang H, Makhtar M, Shamsudin SN. Classification model and analysis of students' performance. Journal of Fundamental and Applied Sciences. 2017; 9(6S):869-85.
[10] Tsai YS, Gasevic D. Learning analytics in higher education---challenges and policies: a review of eight learning analytics policies. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference 2017(pp. 233-42).
[11] A. Hamoud, A. Humadi, W. A. Awadh, and A. S. Hashim, "Students' success prediction based on Bayes algorithms," International Journal of Computer Applications, vol. 178, pp. 6-12, 2017.
[12] A. K. HAMOUD, "CLASSIFYING STUDENTS'ANSWERS USING CLUSTERING ALGORITHMS BASED ON PRINCIPLE COMPONENT ANALYSIS," Journal of Theoretical & Applied Information Technology, vol. 96, 2018.
[13] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," Procedia-Social and Behavioral Sciences, vol. 97, pp. 320-324, 2013.
[14] M. Berland, R. S. Baker, and P. Blikstein, "Educational data mining and learning analytics: Applications to constructionist research," Technology, Knowledge and Learning, vol. 19, pp. 205- 220, 2014.
[15] Hamoud, Alaa, Ali Salah Hashim, and Wid Akeel Awadh. "Clinical Data Warehouse: A Review." Iraqi Journal for Computers and Informatics 44.2 (2018).
[16] A. K. Hamoud and A. M. Humadi, "Student's Success Prediction Model Based on Artificial Neural Networks (ANN) and A Combination of Feature Selection Methods," Journal of Southwest Jiaotong University, vol. 54, 2019.
[17] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," in 2015 International Symposium on Educational Technology (ISET), 2015, pp. 125- 128.
[18] GUPTA, VRATIKA, PRIYANK SINGHAL, and VIPIN KHATTRI. "ANALYSIS OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING ALGORITHMS:–A STUDY."
[19] G. MeeraGandhi, "Machine learning approach for attack prediction and classification using supervised learning algorithms," Int. J. Comput. Sci. Commun, vol. 1, pp. 11465-11484, 2010.

[20] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques: Elsevier, 2011.
[21] R. B. Millar, Maximum likelihood estimation and inference: with examples in R, SAS and ADMB vol. 111: John Wiley & Sons, 2011.
[22] G. Fitzmaurice and N. Laird, "Multivariate Analysis: Discrete Variables (Logistic Regression)," 2001.
[23] Gupta, Vratika, Vinay Kumar Mishra, Priyank Singhal, and Amit Kumar. "An overview of supervised machine learning algorithm." In 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 87-92. IEEE, 2022.
[24] Ben Youssef, A., Dahmani, M., &Ragni, L. (2022). ICT use, digital skills and students' academic performance: exploring the digital divide. Information, 13(3), 129.
[25] Nakhipova, Venera, Yerzhan Kerimbekov, Zhanat Umarova, Laura Suleimenova, Saule Botayeva, Almira Ibashova, and Nurlybek Zhumatayev. "Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction." International Journal of Information and Education Technology 14, no. 1 (2024).
[26] Yağcı, Mustafa. "Educational data mining: prediction of students' academic performance using machine learning algorithms." Smart Learning Environments 9, no. 1 (2022): 11.
[27] Xianghan, Z., &Qunli, Z. (2022). An Analysis of Students ' Failing in University Based on Least Square Method and a New Logistic Regression Function. Mathematical Problems in Engineering, 2022.
[28] Geetha, R., T. Padmavathy, and R. Anitha. "Prediction of the academic performance of slow learners using efficient machine learning algorithm." Advances in Computational Intelligence 1, no. 4 (2021): 5.
[29] GE, Okereke, C. H. Mamah, E. C. Ukekwe, and H. C. Nwagwu. "A machine learning-based framework for predicting student's academic performance." Physical Science & Biophysics Journal 4, no. 2 (2020).
[30] Hashim, Ali Salah, Wid Akeel Awadh, and Alaa Khalaf Hamoud. "Student performance prediction model based on supervised machine learning algorithms." In IOP conference series: materials science and engineering, vol. 928, no. 3, p. 032019. IOP Publishing, 2020.
[31] Canagareddy, Derinsha, Khuslendra Subarayadu, and Visham Hurbungs. "A machine learning model to predict the performance of university students." In Smart and Sustainable Engineering for Next Generation Applications: Proceeding of the Second International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM 2018), November 28–30, 2018, Mauritius 2, pp. 313-322. Springer International Publishing, 2019.
[32] Mgala, M., &Mbogho, A. (2018). Prediction modelling of academic performance with logistic regression: A case of rural primary school students in Kenya The impact of engineering students' performance in the first three years on their graduation result using educational data mining 2405-8440/ 2019.
[33] M. Mustapha1 F. W Usman2 Suleiman Yusuf, "A LOGISTIC REGRESSION MODEL ON ACADEMIC PERFORMANCE OF STUDENTS IN MATHEMATICS" ISSN: 1597 – 9928 © Wilolud Journals, 2016 1 – 15.
[34] Kim, Tae-Hyun, Dong-Chul Park, Dong-Min Woo, Taikyeong Jeong, and Soo-Young Min. "Multiclass classifier-based adaboost algorithm." In Intelligent Science and Intelligent Data Engineering: Second Sino-foreign-interchange Workshop, IScIDE 2011, Xi'an, China, October 23-25, 2011, Revised Selected Papers 2, pp. 122-127. Springer Berlin Heidelberg, 2012.
[35] Dinakaran, S., and P. Ranjit Jeba Thangaiah. "Ensemble method of effective AdaBoost algorithm for decision tree classifiers." International Journal on Artificial Intelligence Tools 26, no. 03 (2017): 1750007.
[36] Strobl, Carolin, James Malley, and Gerhard Tutz. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." Psychological Methods 14, no. 4 (2009): 323.
[37] Beaulac, Cédric, and Jeffrey S. Rosenthal. "Predicting university students' academic success and major using random forests." Research in Higher Education 60 (2019): 1048-1064.
[38] Pathak, Soham, Indivar Mishra, and Aleena Swetapadma. "An assessment of decision tree based classification and regression algorithms." In 2018 3rd International Conference on Inventive Computation Technologies (ICICT), pp. 92-95. IEEE, 2018.

[39] Shetty, Shruthi H., Sumiksha Shetty, Chandra Singh, and Ashwath Rao. "Supervised machine learning: algorithms and applications." Fundamentals and methods of machine and deep learning: algorithms, tools and applications (2022): 1-16.

[40] The output is in the form of a 0 or 1. LR issues are very similar to regression problems, with the exception that regression problems are solved using continuous data, while logistic regression problems are solved using classification.

[41] Gupta, Vratika, Priyank Singhal, and Vipin Khattri. "Student Performance Using Antlion Optimization Algorithm and ANN Regression." In 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), pp. 468-471. IEEE, 2023.

[42] Bayro-Corrochano, Eduardo Jose, and Nancy Arana-Daniel. "Clifford support vector machines for classification, regression, and recurrence." IEEE Transactions on Neural Networks 21, no. 11 (2010): 1731-1746.

[43] Saluja, R., Rai, M., & Saluja, R. (2023). Original Research Article Designing New Student Performance Prediction Model Using Ensemble Machine Learning. Journal of Autonomous Intelligence, 6(1).

[44] Wiyono, S., Abidin, T., Wibowo, D. S., Hidayatullah, M. F., & Dairoh, D. (2019). Comparative study of machine learning knn, svm, and decision tree algorithm to predict students performance. International Journal of Research-Granthaalayah, 7(1), 190-196.