Evaluation of School Students Performance Using Machine Learning

N. T. Renukadevi, K. Saraswathi, E. Roshini, M. G. Lakshitha and S. Pratheeksha
Department of CT-UG, Kongu Engineering College, Erode, Tamil Nadu, India

Keywords: Support Vector Machines, Decision Trees, Random Forest, K-Nearest Neighbour, Naïve Bayes.

Abstract: In today's educational perspective, the need for data-driven insights to enhance student outcomes is increasingly recognized. In this paper we are going to develop a machine learning model to predict and evaluate student performance based on various academic and demographic factors. This system will utilize past information about students like their grades and background to provide educators better suggestions on how to assist students in choosing their academic group for the 11th grade. The dataset will be prepared by collecting information from school students through Google forms. To predict and evaluate student's performance, we apply four distinct machine learning models like Decision Trees, Random Forest, Support Vector Machines (SVM), and Logistic Regression. This research exposes the application of machine learning in guiding the selection of academic groups with promising results and significant potential in educational settings. Furthermore, the research underscores the importance of using data-driven approaches to support educators in making informed decisions and also this can assist in altering personalized interventions enhancing learning outcomes for all students.

1 INTRODUCTION

In today's fast-changing world of education, there's a growing recognition that using data to improve student outcomes is essential. Traditional methods of guiding students may not be enough to address the varied and complex needs of modern learners. Schools are trying to support student success more effectively and advanced technologies like ML are emerging as promising tools. This paper discusses a detailed study on how MLmodels developed and predict student performance, particularly to help students choose their academic groups, specifically for 11th grade. This is done by analyzing the academic and personal factors, like student's grades and socio-economic backgrounds and goals to give educators strong, data-based recommendations to better support students in making informed decisions about their future studies.

To succeed in this research, a dataset will be created by collecting information from students through Google Forms. This ensures that the data may vary and reflects the many different aspects that can show a student's performance. The data will be analyzed using four different ML methods: Random Forest, K-Nearest Neighbours, Naïve Bayes, and Adaptive Boosting. Each of these methods has its own strengths in recognizing patterns and making accuracy, which will allow for a comparison of how to evaluate student performance and select academic groups.

The research aims to show how machine learning can predict the student's performance by using historical data. This not only makes more accurate predictions about student performance but also helps educators for more personalized educational support for each student, leading them to choose academic groups for better learning outcomes. This study highlights the value of using data-driven approaches in education. This gives educators an actionable analysis based on real data; machine learning can significantly improve the decision-making processes involved in academics. Additionally, these technologies can help make educational practices, ensuring that all students get the guidance and support to succeed academically.

Thus, this research explores how machine learning can be used in education, predicting a new way to evaluate student performance and help the educators to guide the students in academic group selection. The promising results from these models suggest that machine learning has the potential to provide educational practices, providing a new level of support for both educators and students. Thus, as

Renukadevi, N. T., Saraswathi, K., Roshini, E., Lakshitha, M. G. and Pratheeksha, S.
Evaluation of School Students Performance Using Machine Learning.
DOI: 10.5220/0013622900004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 503-512
ISBN: 978-989-758-763-4

we continue to explore all these technologies, it becomes clear that data-driven approaches are necessary steps in improving educational outcomes for all students.

ML is a part of AI that focuses on developing methods for computers to learn from data and make predictions or decisions. The process starts with collecting data, which can come from various sources like databases, text, or images. This data is then pre-processed to ensure whether the data is ready for analysis. Once the data is prepared, machine learning algorithms are applied. These algorithms can handle different types of problems: supervised learning is used when the model learns from labeled data to predict outcomes, while unsupervised learning is useful for finding patterns in data that is not labeled. There's also reinforcement learning, which trains models to make decisions in a step-by-step manner mainly used in robotics and gaming. The model is trained on historical data to recognize patterns and can be used to predict new data. To ensure the working of the model, it is evaluated using specific metrics and often adjusted or fine-tuned to improve the performance. The final step involves comparing different algorithms to find the best for the problem. Machine learning's ability has made it useful in many areas like healthcare, finance, understanding language, and recognizing images etc., It helps organizations make smarter decisions based on data.

2 LITERATURE REVIEW

Esmael Ahmed (Ahmed, 2024) conducted a comparative study on the challenges faced by contemporary educational institutions in analyzing the efficiency; endowing sound education, formulating strategies for evaluating students' performance and recognizes future needs. In this study the data is collected from the Wollo University learning management system and grouped together to predict the student's end outcome based on their demographic details and performance in entrance exams, and with various factors etc., In this paper, performance of the classifiers such as SVM, DT, NB, and KNN are examined. It is observed that SVM algorithm provides better outcomes with 96% accuracy.

In research work, done by P. Krishna Reddy et al. (Reddy, Bavankumar, et al. , 2024) the data is collected with 22 attributes. This study applied an empirical method to opt for students' records and decide on important variables for analysis. The study has applied five different data mining techniques PAC, SVM, LDA, RNC and ET and subsequently compares the consequences of five ML algorithms to identify the paramount performing algorithm. And from this research SVM provides the highest accuracy of 94.86%.

Various studies have addressed the challenge of predicting student performance using machine learning techniques; Nitin Yadav et al. used UCI machinery student dataset as input and pre-processed the data to select the attributes. The study solved this problem by applying SVM, NB, C4.5, ID3 and found that SVM gives more accuracy of 88%. For future work, select the best attribute sets like Stud_Name, Gender, Previous Exam Marks, Address, Parents Education, etc. with ML algorithm to provide the high accuracy result of prediction of student performance system (Yadav and Deshmukh, 2023).

Another comparative study was done by Yawen Chenand and Linbo Zhai(Chen and Zhai, 2023). First the three datasets were selected which are oriented to Student performance, Engineering placements prediction and Student admission and also done pre-processing to select the specific features. In this subsection, the study applies seven popular ML algorithms including KNN, DT, RF, LR, SVM, NB, and ANN. RF provides highest accuracy among those seven ML algorithms.

Nuha Alruwais and Mohammed Zakariah (Alruwais and Zakariah, 2023) said that assessing acquaintance of student is crucial to determine student progress and providing feedback to improve student performance.Thisprovides feedback to improve student performance. The study used 2 different data sets, the dataset 1 includes all the attributes and the dataset 2 removes the least correlated variables to create a smaller dataset. To solve this problem, the 7 different classifiers are used, including SVM, LR, RF, DT, GBM, GNB, and MLP. Then the performance of Dataset 1 and Dataset 2 were compared. This resulted that the GBM exhibited the highest prediction accuracy of 98%.

Rosemary Vargheese et al. (Vargheese, Peraira, et al., 2022) proposes Student Performance Analysis System (SPAS) to remain track of students' results. During the implementation phase, to generate rules for prediction of students' performance, they analyzed 114 students' records by using data mining techniques for their future. This research has 6 applied four different data mining techniques SVM, DT, RF, NB. After comparison, it is concluded that the SVM has provided the highest accuracy

(81.82%) among other classification techniques. They predicted the performance of the students in two phases (i) training phase, and (ii) testing phase. The current prediction model is not dynamically updated within the system's source code. For future, they introduce a dynamic prediction model allowing that model to be retrained automatically whenever new training data is added to the system with ML algorithm for the prediction of the student performance system.

Jovana Jovic et al. Educational Data Mining has conducted an EDM approach in order to classify and predict student performance with ML techniques. ML algorithms such as LR, LDA, KNN, DT, NB and SVM have been used. The research classified that SVM shows the best results with accuracy of 88.5%. Thus, the future work will analyze a larger number of ML algorithms and try to gain a more accurate model for the prediction of student performance and support learning (Jović, Kisić, et al. , 2022).

Educational data mining has emerged as a powerful tool for predicting students' academic success and uncovering hidden relationships in educational data. Mustafa Yagci (Yağcı, 2022) presented a novel model in this study that uses machine learning (ML) algorithms to forecast undergraduate students' final exam grades. The midterm exam grades of the students are the source data for this project. The final exam grades of the students were predicted by calculating and comparing the performances of the RF, SVM, LR, NB, and KNN. In the future, it may be possible to review students' working methods and enhance their performance by projecting their achievement grades.

AlabbasHayderassesses and contrasts the accuracy, precision, recall, and prediction of two machine learning algorithms—SVM and ANN—when trained to classify binary datasets. Three distinct subquestions were used to break down each research question. In this study, the seven researchers developed two machine learning models and evaluated their efficacy. An ANN model was the second, and an RF model was the first. Future research could look at utilizing each course's degree separately rather than just the course pass rate to determine how each course's performance impacts future academic results. They used ANN, SVM and other ML classification algorithms for the prediction of students' performance (Hayder, 2022).

Charalampos Dervenis et al. (Dervenis, Kyriatzis, et al. , 2015) has completed a project on learning analytics which measures, gathers and analysis data about learners and their contexts in order to optimize learning and the environments in which it takes place. This represents a tangible step toward a more change in a society that is increased by algorithms. The algorithms such as KNN, RF and SVM were tested. The test data is used to gauge the algorithm's performance, training data served as the foundation for its creation. Every data set is run through the test data and outcome is predicted. The RF produced the beat results then followed by KNN and SVM. The study made accurate prediction for passes or fails using the model. For future behaviors, identify potential problems at an early stage or even improve inter-institutional collaboration and develop an agenda for the larger community of students and teachers.

Opeyemi Ojajuni et al. (Ojajuni, et al. , 2021) has done a study using ML to explore and to predict student academic performance by analyzing data from 1,044 students, including demographic, social, and academic information. It applies various supervised ML algorithms, such as RF, SVM, GBM, DT, LR, and DL, to classify and predict outcomes. The research involves data preprocessing and feature engineering, categorizing final grades into excellent, good, satisfactory, poor, and failure. XGBoost showed the highest accuracy of 97.12%. The study suggests that future research could investigate additional ML models and address challenges like overfitting and model deployment in real educational settings.

Poonam Sawant et al. (Sawant, Gupta, et al. , 2021) examine the impact of the COVID-19 pandemic on student performance in India, highlighting the challenges faced due to the shift from offline to online learning. Various ML algorithms, such as DT and NB, are used to analyze and predict student performance during this period. Findings indicate that students' performance decreased during the pandemic, with NB showing better accuracy than the DT. Future research aims to incorporate more advanced algorithms to enhance the performance evaluation system.

Samah Fakhri Aziz et al. explored using ML to predict student performance. It compares Gradient Boosted Decision Trees, RF, DT, and DL algorithms. This paper showed that DL is the highest accuracy in predicting student performance. The study used a dataset of 450 students with 17 attributes, and the pre-processing involved checking and handling missing values. Overall, DL was found to be the best for predicting how well students will do (Aziz, 2020).

Ali Salah Hashim et al. (Hashim, Awadh, et al. , 2020) evaluated number of supervised ML algorithms such as DT, NB, LR, SVM, KNN, Optimization and NeuralNetwork had performed in predicting student's performance on final exams. The results showed that the logistic regression gave most accurate accuracy as 68.7% for passed students and 88.8% for failed students. Data pre-processing involves data preparing, combining and cleaning the data in order to train them was the first step. The subject of second step is the most frequently used algorithm's classification performance according to ML technique.

J. Dhilipan et al. (Dhilipan, Vijayalakshmi, et al. , 2021) explores the application of ML techniques to forecast academic success in educational settings. The study utilizes various algorithms, including Binomial Logistic Regression, DT, Entropy, and KNN, to analyze student performance based on their 10th, 12th, and semester marks. In this study The Binomial Logistic Regression has achieved the highest accuracy with a rate of 97.05%. The results demonstrate that these methods can help educators identify students at risk of underperforming and guide them in improving academic outcomes Rajalaxmi et al. (Rajalaxmi, Natesan, et al. , 2019). The research suggests that in future additional features are added to our dataset to acquire better accuracy Saraswathi et al (Saraswathi, Renukadevi, et al. , 2024).

3 METHODOLOGIES USED

3.1 K-NearestNeighbors

KNN is a supervised learning algorithm used for both classification and regression tasks. Data Points are classified or predicted according on the maximum number ofclasses.The "k" value denotes the number of neighbors for the classification or regression decision.Distance metrics commonly used include Euclidean distance, whichmeasures the similarity between data points, based on the Euclidean distance calculation shown in Equation (1).KNN is a lazy-learner because it does not make assumptions based on underlying data and defers computation until predictions are needed. KNN is computationally expensive for large datasets, and its performance can be affected by noise, and its role extends to areas such as recommendation systems and anomaly detection.

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

KNN is known for its simplicity and intuitiveness, with no training phase required. It can handle non-linear decision boundaries and complex patterns without making strong assumptions about data distribution to make it suitable for multiclass problems. However, KNN is a sensitive choice to the hyper parameter "k". Though it calculates distance for all data points it is computationally expensive for large data sets. It can suffer from the "curse of dimensionality", because when the number of features increases the performance will decrease. The diagrammatic representation is shown in Figure 1.

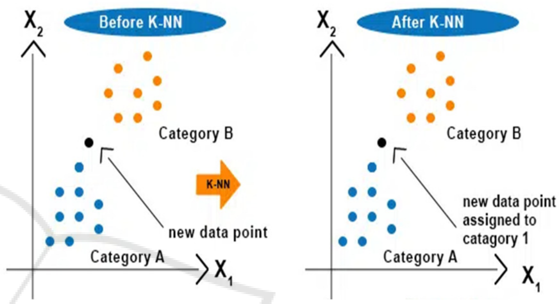

Figure 1: K - Nearest Neighbors

3.1.1 How KNN Works

Step 1: Determine the value of "k".
Step 2: Calculate the distance of all data points from existing data using the Euclidean distance measuring technique.
Step 3: Choosing the number of nearest data points to the new data based on the "k" value.
Step 4: For each data point, one of its K neighbors predicts the class.
Step 5: Counting the data points according to the class labels and resulting the majority voted label to the new data point.

3.2 Decision Tree

DT is a supervised machine learning algorithm used for both classification and regression tasks. It splits the dataset into subsets based on the most important feature, by creating a tree-like structure as shown in Figure 2, where each internal node represents a decision based on feature, each branch fits to an outcome of that decision, and each leaf node represents the final classification. The algorithm selects the best features for splitting by maximizing information gain by calculating entropy for each as defined by equation (2). DT are easy to visualize and useful for understanding the decision-making process. RF oftenleverages multiple DT for improved accuracy.

$$E(H(S)) = \sum_{i=1}^{n} pi \log_2(pi) \tag{2}$$

DT is transparent and easy to understand, offering human-readable decision rules. They handle both classification and regression tasks. DT is robust to outliers and alsoaccommodates missing value. They are exposed to overfitting, especially when the tree is deep and complex, it will be sensitive to small changes in the data which leads to different tree structures. DT may struggle with imbalanced datasets and are limited in capturing decision boundaries compared to ensemble methods. Additionally, the greedy search during tree construction may not always lead to globally optimal solutions Renukadevi et al..

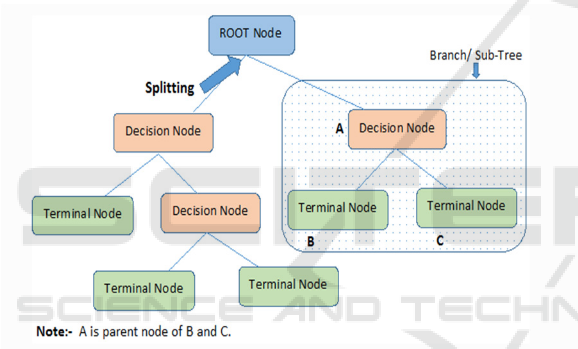

Figure 2: Decision Tree

3.2.1 Decision Tree analysis

Step 1: Find the problem.
Step 2: Build thestructure of the decision tree.
Step 3: Identify the alternatives of the decision.
Step 4: Calculate entropy for each feature.
Step 5: Use entropy to calculate information gain.
Step 6: Discover the possible outcomes.
Step 7:Examine the best decision.

3.3 Support Vector Machines

SVM is a supervised machine learning algorithm that can be applied to classification and regression problems. It functions by locating the hyperplane in a high dimensional space that divides data points into distinct classes the best, as defined by equation (3). The data points that are closest to the hyperplane are known as "support vectors", and they have an impact on the hyperplane's orientation and position which is shown in Figure 3. The goal of SVM is to maximize the margin, or the separation between each class's closest data points and the hyperplane. The margin is calculated as defined in the equation (4).With the use of kernel functions, the algorithm can be expanded to handle non-linear relationships and is efficient when processing high-dimensional data. SVM is renowned for its reliable operation and capacity to manage intricate decision boundaries across a range of applications.

$$w.x + b = 0 \tag{3}$$

$$Margin = \frac{2}{\|w\|} \tag{4}$$

SVM is effective at finding optimal hyper planes for binary and multi-class classification, and also uses kernel functions to handle both linear and non-linear decision boundaries. They are robust to outliers and offer high accuracy when properly tuned, making them suitable for complex and high-dimensional datasets.SVMs however, can be computationally demanding for large datasets and have limited interpretability, particularly when dealing with non-linear kernels. They are also prone to overfitting if not properly regularized and can struggle with noisy or overlapping classes.

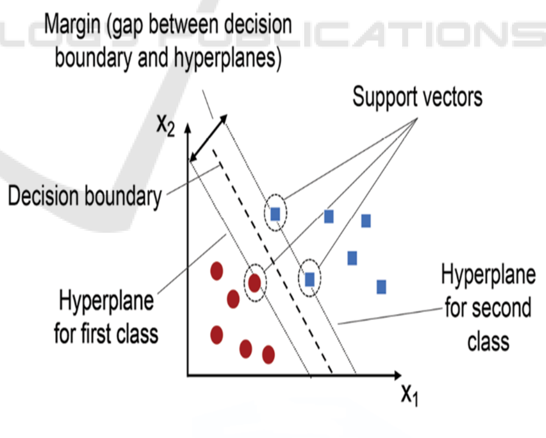

Figure 3: Support Vector Classifier

3.3.1 Steps in the SVM Algorithm

Input Data:Gatherlabeled training data using class labels and feature vectors.

Linear SVM:Choose the optimal hyperplane if the data can be divided linearly.

Non-Linear SVM:If the data cannot be separated linearly,use the kernel trick to convert it into a higher dimension.

Optimal Hyperplane:Determine which hyperplane maximizes the difference in class margins..

Optimization Problem:The process of finding the best model parameters, hyperparameters, or features to minimize or maximize the performance metrics, such as error, accuracy or loss.

Support Vectors:To define the hyperplane, find the closest support vectors.

Prediction:Determine which side of the hyperplane new data falls on and classify it accordingly.

3.4 Random Forest

RF is an ensemble learning technique based on DT. It builds multiple decision trees during training and outputs the classification or regression as defined in the equation (5) of the individual trees for a more robust and accurate result. The key idea is to introduce randomness in the tree-building process by considering a random subset of features and data points for each tree. This helps mitigate over-fitting and enhances the model's generalization ability. RF is known for its high predictive accuracy, resilience to outliers, and suitability for complex datasets. It is widely used in various applications, Including classification, regression, and feature importance analysis in machine learning.

$$\int_{rf}^{B}(x) = \frac{1}{B}\sum_{b-1}^{B} Tb(x) \tag{5}$$

Ensemble methods, like RF, improve predictive accuracy and reduce overfitting by aggregating multiple decision trees. They are robust to outliers and noisy data which handle both classification and regression tasks and can effectively manage high-dimensional data and large datasets and shown in Figure 4. Additionally, they provide feature importance scores, which aid in feature selection and interpretation. However, ensemble methods are less interpretable than a single DT. They may not perform as well as gradient boosting for certain tasks and require more memory and computational resources.

3.4.1 The steps to build a random forest are

- Understand the problem: Define the objectives
- Understand the data: Get to know the raw materials
- Data preprocessing: Prepare the data
- Data analysis: Build models
- Results interpretation and evaluation: Interpret and evaluate the results
- Data report and communication: Create a data report and communicates the results

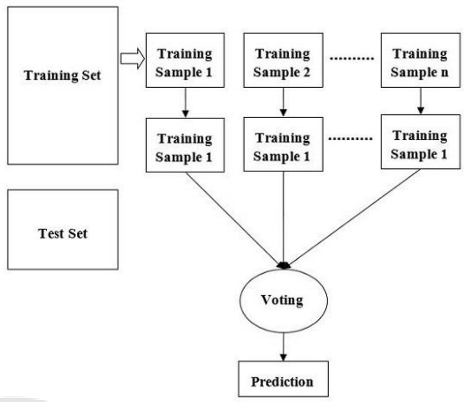

Figure 4. Random Forest

3.5 Naive Bayes

NB is a simple and fast ML algorithm used for classification tasks. It is based on Bayes' theorem, which calculates the probability of a class given in the features of the dataset. In NB, the posterior probability is used to calculate the probability of a data point belonging to a particular class as defined in the equation (6). The algorithm assumes that all features are independent of each other, so that it is called"naive". And also,Naive Bayes works well in many real-world applications, especially with large datasets.

$$P(C \mid X_1, X_2, ..., X_n) \alpha P(C) x \prod_{i=1}^{n} P(X_i \mid C) \tag{6}$$

NB is fast and efficient, especially with large datasets and it performs well for tasks like text classification and spam detection. It’s easy to implement and works well with a small amount of data. However, the main limitation is the assumption that all features are independent, which may not hold true values in many cases, potentially reducing accuracy. It also struggles when features are closely related.

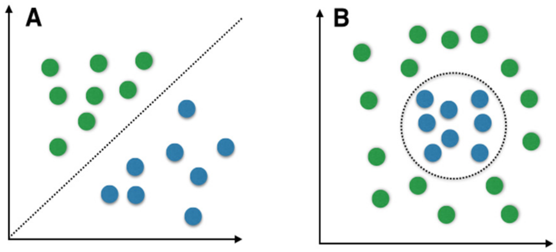

Figure 4: Naive Bayes Diagram

3.5.1 The steps of the Naive Bayes classifier are

Training: Use the training data to estimate the parameters of a probability distribution.

Prediction: For new test data, calculate the posterior probability of that sample belonging to each class.

Classification: Classify the test data according to the largest posteriorprobability.

3.6 Boosting Algorithm

Boosting is a machine learning technique that improves the accuracy of models by combining the strengths of multiple weak models to create a stronger overall model. It works by training models sequentially, where each new model focuses on correcting the errors made by the previous ones.

3.6.1 Common types of boosting algorithms include

Ada-Boost -This adjusts the weights of incorrectly classified data points.

Gradient Boosting - This optimizes performance by minimizing errors through gradient descent.

XGBoost–isan efficient version of Gradient Boosting known for its speed and performance on large datasets.

Boosting algorithms are powerful for improving model accuracy by combining multiple weak models into a strong one. They are effective for complex datasets and can handle various types of data. However, boosting can be sensitive to noisy data and outliers which can lead to overfitting if not carefully managed. It also requires more computational resources and time due to its sequential nature and tuning the model can be complex.

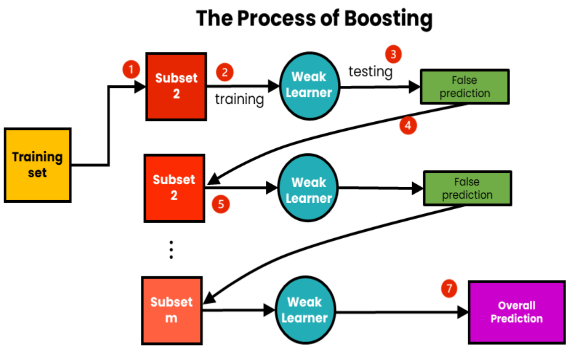

Figure 5. Boosting Algorithm

3.7 Gradient Boosting

Gradient Boosting is a popular boosting algorithm that combines multiple weak models to create a strong predictive model. It works by iteratively training decision trees to predict the residual errors of the previous iteration, using gradient descent to minimize the loss function. The process starts with an initial simple model, and then, at each iteration, a new decision tree is trained to predict the difference between the observed output and the predicted output from the previous iteration. The predictions from each tree are weighted and combined to update the overall model. The gradient descent step adjusts the weights to minimize the loss function, typically mean squared error or logistic loss, ensuring that each subsequent tree focuses on correcting the errors of the previous ones. This sequential process continues until a specified number of trees is reached or a stoppingcriterion is met, resulting in a highly accurate and robust predictive model.

In this system, we are predicting and evaluating the performance of students based on their academic factors using ML algorithms. Block diagram is shown in Figure 6. We have collected data from the school students, upto 170 data for our project.Our research mainly focuses to support educators in making informed decisions and also assist in altering personalized interventions enhancing learning outcomes for all students.

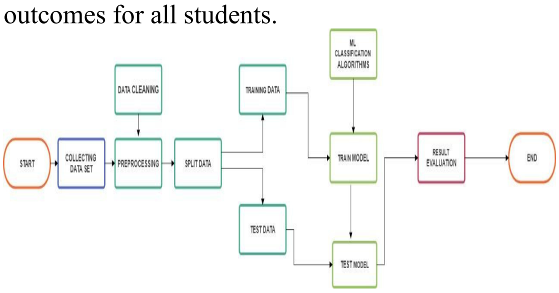

Figure 6. Block Diagram

4 RESULTS AND DISCUSSION

In this research work, various ML algorithms were implemented for predicting the performance of the student. We collect the information from the school students through google form. The dataset was split into two parts: 80% for training the model and 20% for testing it. This is a common practice in machine learning to make sure the model works well with new data.

4.1 Environment Setup

The special tools and programs were used to conduct the experimentation, including Google Colab for data visualization, and Microsoft Excel for data handling. Python was chosen due to its easy-to-learn syntax and the availability of libraries like Numpy, Pandas, Scikit-learn, and Sklearn. Numpy calculates mean values; Pandas fetches data from files, creates data frames, and handles data frames. Scikit-learn, also known as Sklearn, contains machine learning tools like classification and so on.

4.2 Data Pre-Processing

The dataset was cleaned by removing missing values, extracting features and selecting key features 25. Here is an elaboration on each step:

- Removing Missing Values - Data often has missing or NAN values, It will affect the accuracy of the dataset
- There are several techniques for handling missing data such as:
- Removal - If the amount of missing values is small, rows or columns with missing values can be dropped.
- Imputation - Missing values can be filled in with statistical measures such as the mean, median, or mode, or even using more sophisticated methods like regression or interpolation.
- Feature extraction – It involves the process of creating new features from existing data. These new features can have better patterns in data and it will improve the model performance.

There are different techniques it includes:

Time-related Features - If you have date-time data, you can extract features like the day of the week, month, or hour.

Textual Features - For text data, you might extract features like word counts, or use techniques like TF-IDF or word embeddings.

Statistical Features - Generating features based on statistical properties like mean, variance, or correlation can add valuable information.

4.2.1 Selecting Key Features

Correlation Analysis - It identifies the features and it is highly related with the target variable. It helps to find the outcome more.

Variance Threshold - Low variance features do not help in providing any useful information.

Feature Importance - Algorithms like decision trees or random forests can ranked as a importance feature, because it will help to improve the prediction by showing which algorithm is useful for the model.

Dimensionality Reduction – In this process we eliminate the unwanted features without losing much information. We can easily improve the accuracy By using these steps, the dataset is prepared for analysis or model building, resulting in a cleaned data.

In this study, we applied several Machine Learning algorithms, including KNN, Decision Tree (DT), SVM, Naive Bayes (NB), Gradient Boosting (GB), and Random Forest Classifier (RFC), to predict student performance based on data collected from school student. Here we use factors such as Q_TOTAL, H_TOTAL, AVERAGE as the feature selection subset and RESULT is the target to predict the accuracy of the algorithm. The dataset includes 170 students, with their average scores ranging from 198.5 to 483.5.The average score across all students is approximately 357.2with most students scoring between 302.5 and 431.Students were categorized into four groups based on their results are Above Average: 59 students (average score: 394.0),Average: 49 students (average score: 317.6),Excellent: 33 students (average score:468.8),Below Average: 28 students (average score: 217.5). In order to choose a classifier for predicting the final exam outcome (whether the course was passed or failed), a 10-fold cross-validation approach was conducted. Cross validation approach splits the training dataset into 10 groups of approximately equal size, trains the model on nine groups and tests the model on the tenth group in ten iterations. The outcomes of the experiments are summarized using classifier accuracy that was calculated as the average accuracy after ten cross validation iterations.

Gradient boosting is a popular boosting algorithm in machine learning used for classification and regression tasks. Boosting is one kind of ensemble Learning method which trains the model sequentially and each new model tries to correct the previous model. It combines several weak learners into strong learners. We have used this algorithm for comparative analysis and we that It has the accuracy of 96%.

Table 1: Accuracy measures of ML algorithms

ALGORITHM	ACCURACY
KNN	88%
DT	91%
SVM	99.23%
NB	84%
GB	96%
RFC	94%

Accuracy is a metric that measures how often a model correctly predicts the outcome. It is the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. Accuracy is commonly used for classification tasks. KNN gives the accuracy as 88%, In DT has the accuracy of 91%, In SVM we have implemented Normalization and we get the accuracy as 99.23%, NB has the accuracy of 84%, GB has the accuracy of 88.24%, RF has the accuracy of 94%. Among the algorithms tested, the Support Vector Machine has the most accuracy, with an 99.23% prediction accuracy, compared to the other models.

REFERENCES

E. Ahmed, “Student Performance Prediction Using Machine Learning Algorithms,” Applied Computational Intelligence and Soft Computing, vol. 2024, 2024, doi: 10.1155/2024/4067721.

P. K. Reddy, S. Bavankumar, Y. P. Mohaideen, and G. Kishore, “Students’ performance evaluation using machine learning algorithms”, Material Science and Technology, 2024, vol.23, pp.310 – 322.

N. R. Yadav and S. S. Deshmukh, “Prediction of Student Performance Using Machine Learning Techniques: A Review,” 2023, pp. 735–741. doi: 10.2991/978-94-6463-136-4_63.

Y. Chen and L. Zhai, “A comparative study on student performance prediction using machine learning,” Educ Inf Technol (Dordr), vol. 28, no. 9, pp. 12039–12057, Sep. 2023, doi: 10.1007/s10639-023-11672-1.

N. Alruwais and M. Zakariah, “Evaluating Student Knowledge Assessment Using Machine Learning Techniques,” Sustainability (Switzerland), vol. 15, no. 7, Apr. 2023, doi: 10.3390/su15076229.

R. Vargheese, A. Peraira, A. Ashok, and B. Johnson, “Students’ Performance Analysis Using Machine Learning Algorithms”,International Journal of Research in Engineering and Science, 2022, vol.10, no.6, pp.1804-1809.

J. Jović, E. Kisić, M. R. Milić, D. Domazet, and K. Chandra, “Prediction of student academic performance using machine learning algorithms, Proccedings of the Thirteenth International Conference on e-Learning, 2022.

M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

Alabbas Hayder,“Predicting Student Performance Using Machine Learning: A Comparative Study Between Classification Algorithms” , DigitalaVetenskapligaArkivet, 2022.

C. Dervenis, V. Kyriatzis, S. Stoufis, and P. Fitsilis, “Predicting Students’ Performance Using Machine Learning Algorithms,” in ACM International Conference Proceeding Series, Association for Computing Machinery, Sep. 2022. doi: 10.1145/3564982.3564990.

O. Ojajuni et al., “Predicting Student Academic Performance Using Machine Learning,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2021, pp. 481–491. doi: 10.1007/978-3-030-87013-3_36.

P. Sawant, S. Gupta, Y. Sharma, and A. Singh, “Classification Approach for Evaluating Students Performance in Covid 19 Pandemic,” Int J Eng Adv Technol, vol. 10, no. 4, pp. 110– 113, Apr. 2021, doi: 10.35940/ijeat.d2374.0410421.

S. F. A. Aziz, “Students’ Performance Evaluation Using Machine Learning Algorithms,” College Of Basic Education Research Journal, vol. 16, no. 3, pp. 977–986, Jul. 2020, doi: 10.33899/berj.2020.166006.

A. S. Hashim, W. A. Awadh, and A. K. Hamoud, “Student Performance Prediction Model based on Supervised Machine Learning Algorithms,” in IOP Conference Series: Materials Science and Engineering, IOP Publishing Ltd, Nov. 2020. doi: 10.1088/1757-899X/928/3/032019.

J. Dhilipan, N. Vijayalakshmi, S. Suriya, and A. Christopher, “Prediction of Students Performance using Machine learning,” IOP Conf Ser Mater Sci Eng, vol. 1055, no. 1, p. 012122, Feb. 2021, doi: 10.1088/1757-899x/1055/1/012122.

Rajalaxmi, R. R., Natesan, P., Krishnamoorthy, N., & Ponni, S. (2019). Regression model for predicting engineering students academic performance. International Journal of Recent Technology and Engineering, 7(6S3), 71-75.

N. T. Renukadevi, S. Nanthitha, K. Saraswathi, S. Shobika and R. T. Karthika, "WhatsApp Group Chat Analysis by using Machine Learning," *2023 International*

Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023, pp. 340-346, doi: 10.1109/ICSCDS56580.2023.10104961.

K. Saraswathi, N. T. Renukadevi, K. G. Akshaya and S. Kanishka, "Hyper Parameter Optimization in Machine Learning For Enhancing Loan Sanction Processes," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10724586.