Student performance prediction

Dechamma MP, Madhu Shree M, Surya thej S, Guru Prasad BA
2nd Sem MSc Data science, School of Computer Applications

ABSTRACT:

Predicting student performance using machine learning models can help identify at-risk students and optimize educational interventions. By analyzing various factors such as demographic information, previous academic records, and socio-economic background, accurate predictions can be made, enabling early intervention and personalized support to improve student outcomes.

I. INTRODUCTION

Machine learning is a field of study and application within artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed. It involves the creation of mathematical models and algorithms that automatically improve their performance through experience, without human intervention. In essence, machine learning allows computers to learn from data and make predictions or take actions based on patterns and insights derived from that data. Education plays a crucial role in shaping the future of individuals and societies. With the advancements in technology and the availability of vast amounts of educational data, machine learning techniques have emerged as powerful tools to analyse and predict student performance. By leveraging various data points such as academic records, socio-economic factors, and learning behaviour, machine learning models can provide valuable insights to educators, policymakers, and institutions to enhance student success and tailor personalized interventions. Machine learning algorithms have the ability to identify patterns and relationships within large datasets that humans might not easily discern. By training models on historical data, they can learn to make accurate predictions about future outcomes, including student performance. These predictions can assist educators in understanding which factors are most influential in determining academic success, enabling them to intervene and provide targeted support to students who may be at risk of falling behind. There are several key factors that machine learning models often consider when predicting student performance. These include past academic achievements, attendance records, demographic information, socioeconomic status, and various other indicators of student engagement and behaviour.

II. LITERATURE REVIEW

The author Punlumjeak.W at all (2017) [1] in paper have proposed that, the researchers utilized a large student dataset as big data to develop a prediction model for classifying students' performance on the Microsoft Azure platform. They employed feature selection methods, applied a classification mining technique, and evaluated the model's performance. Although the overall accuracy was high, potential conflicts in the confusion matrix and the presence of data imbalance will be addressed in future research.

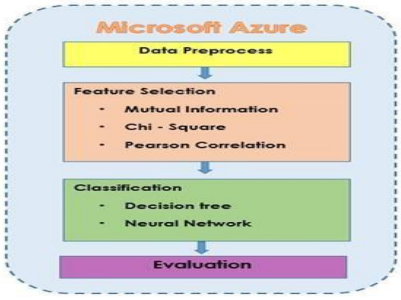


Figure 1. Propose Framework

The author Ahammad. K at all (2021) [2] of this paper have proposed that, predicting student performance is becoming more challenging because it is influenced by various factors beyond just scores in previous exams. The proposed models aim to identify at-risk students early and take preventive measures to improve their chances of success, thus enhancing the overall quality of education in the institution.

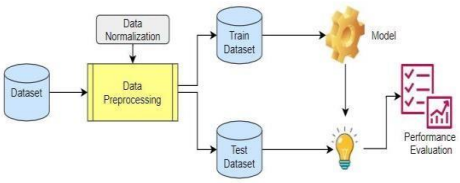


Figure2. Frequency of each GPA class in dataset

The author Taras. H at all (2005) [3] of this paper have proposed that, the link between obesity and school performance has been a topic of research, and while the number of articles on this subject may be limited, there have been consistent findings suggesting a negative impact on school performance among overweight or obese children. One study, for example, suggests that obese children and adolescents may miss more school days.

The author of Singh. R at all (2020) [4] of this paper presents a model for predicting the performance accuracy of students using machine learning techniques and ensemble methods [4]. The results demonstrate the effectiveness of the Boosting ensemble technique in achieving high prediction accuracy. The practical implications of this study include the ability to identify non-performing students early and provide targeted support to improve their academic performance. By enhancing the quality of higher education, these findings can have a positive impact on educational institutions and students alike.

The author Naomi. J.F at all (2021) [5] of this paper have presented an investigation on early prediction of students' academic performance using linear regression models. The researchers considered six independent parameters that were recorded during classroom proceedings and examined their relationship with the end-semester exam scores. the researchers concluded that the multiple linear regression model outperformed the simple linear regression model in early prediction of student performance. This implies that considering all six independent parameters individually in the MLR model provided a more accurate prediction of students' academic performance compared to using a single aggregate value in the SLR model.

The authors Mohammadi. M at all (2019) [6] of this study focused on predicting the Grade Point Average (GPA) of tertiary students in Vietnam using machine learning techniques. Factors such as personal characteristics (gender and living location), university entry scores, gap year, and academic grades in the first and second year were considered. The research found that all students were accurately classified, with MLP and Naïve Bayes performing well in terms of true positive rate and recall for most classes. These findings can be applied to other higher education institutions in Vietnam to predict students' final academic results and provide necessary support. The developed model can also be integrated into a university's management information system to offer early warnings for atrisk students. Future work will address imbalanced datasets to improve the quality of predictions. The authors express gratitude to the Ministry of Education and Training and Thuong Mai University for their support and provision of data and devices for this research.

The author Ismail. L at all (2021) [7] have Predicted the student performance is crucial in education for improving institutional performance, admission requirements, and course selection. The proposed model utilized supervised machine learning algorithms to predict student grades and status, achieving high accuracy with Logistic Regression. Factors such as data cleanliness, feature domain, feature number, dataset size, and class domain affected accuracy. Reducing predicted values and increasing dataset size improved accuracy. ROC and AUC analysis confirmed Logistic Regression as the best algorithm for predicting student status and grades.

The author Dhilipan. J at all (2021) [8] have done the research to address poor academic performance in computer science at Al-Muthanna University, four classification models were developed: Artificial Neural Network (ANN), Naïve Bayes, Decision Tree, and Logistic Regression. These models were compared using the ROC index and classification accuracy. The ANN model achieved the highest ROC index of 0.807 and an accuracy of 77.04%. The decision tree model revealed that specific attributes, including Computer GradesCourse1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and Residency, influenced the classification process.

The author Albreiki. B at all (2021) [9] of this paper focuses on applying machine learning algorithms (MLAs) to predict student results and identify outstanding and potentially poor-performing students. The research methodology, based on surveying Computer Science students in Kolkata, can be applied to both full-time and distance education courses, including web-based learning. The training set consists of 309 instances, while the testing set has 104 instances. Decision Trees (DTs), specifically C4.5, were found to be the most suitable algorithm for generating production rules. The prediction algorithm's efficiency was evaluated using FMeasure and Kappa Statistic, with average F-Measure values of 0.79 for training and 0.66 for testing, likely due to the small testing dataset. The paper suggests

improvements such as distinguishing between different grade ranges, considering students with missing attributes, and enhancing prediction efficiency through techniques like Combining Multiple Classifiers (CMC) and genetic algorithms.

The author Alamri.R at all (2021) [10] of this paper had the main objective of this systematic review is to identify the state-of-the-art in explainable student performance models. We defined five research questions to explore the student performance measure, predictors used, explainable methods, and evaluation metrics. Through a systematic literature review of the past five years, we found that most studies focused on predicting student outcomes per course as a multi-class problem. Socio-economic features and pre-course performance were the top predictors, while decision trees and rule-based learning algorithms were commonly used. However, there is a lack of research exploring state-of-the-art explainable methods and rich predictors like elearning analytics across different levels of student performance. Another limitation is the absence of evaluation metrics for model explainability, making it difficult to compare models. Overcoming this limitation requires investigating state-of-the-art metrics for assessing explainability. This study highlights the importance of explainable models in education, identifies gaps in the literature, and suggests future research directions.

The author Agrawal. H at all (2015) [11] of this work automates student performance prediction using the RF classification algorithm. RF algorithm is widely used due to its versatility and model-free nature. The results demonstrate accurate prediction of student performance across multiple classes. RF algorithm outperforms the SVM algorithm in terms of prediction quality. This methodology can benefit teachers and students by improving learning outcomes. The model enables early intervention for poor and average performing students.

The author Vijayalakshmi. V at all (2019) [12] have Predicted the student performance is crucial in education for improving outcomes. Educational data mining involves applying data mining concepts and algorithms, including machine learning, in this field. Our proposed system for student performance prediction was trained and tested using various machine learning algorithms such as Decision Tree (C5.0), Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Deep Neural Network in R Programming. Comparing the results, the Deep Neural Network achieved the highest accuracy of 84%.

The author Rai. S at all (2021) [13] have done a work on this systematic review aims to identify the stateof-the-art in explainable student performance models. Five research questions were formulated to investigate the output, input, model, and evaluation metrics used in these models. The review of literature from the past five years revealed that most studies focused on predicting student outcomes per course, with socio-economic features and pre-course performance as the top predictors. Decision trees and rule-based learning algorithms were commonly employed in these studies. However, there is a lack of research utilizing state-of-the-art explainable methods and rich predictors like e-learning analytics.Evaluation metrics for model explainability were also not widely adopted, hindering comparisons among models. Future work should address these limitations and emphasize the importance of explainable models in the educational context.

The author Chauhan. N at all (2019) [14] in this paper uses various machine learning techniques, including artificial neural networks, decision trees, and data mining, to predict students' academic performance. The papers demonstrate the accuracy and reliability of these techniques in predicting students' performance and identifying struggling students. The choice of machine learning technique depends on the specific application, and ensemble methods can be used to improve the performance of classifiers.so they have used various machine learning methods like multiple linear regression, knearest neighbor, random forest, support vector machine to analyse the performance of the students.

The author Al-Shehri, H at all (2017) [15], in this paper "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor" presents two prediction models for estimating student performance in the final examination. the empirical studies showed that SVM slightly outperformed KNN for predicting student grades, achieving a higher correlation coefficient. The paper provides insights into the use of SVM and KNN algorithms for student performance prediction and highlights the importance of exploring different models.

The author Ofori. F at all (2020) [16] purposed this literature review is to explore the potential of machine learning algorithms in predicting students'

performance and enhancing the overall learning process. The authors systematically analyze a range of relevant research articles, highlighting the various machine learning algorithms employed and their effectiveness in predicting student performance. The review covers key aspects related to the use of machine learning in education, including data collection techniques, preprocessing methods, feature selection, model development, and evaluation metrics. The authors also discuss the impact of various factors such as student demographics, learning materials, and teaching methods on the prediction accuracy. The findings of this literature review demonstrate the significant potential of machine learning algorithms in predicting student performance. The authors highlight the importance of accurate prediction in identifying students at risk of poor performance, allowing educators to intervene and provide targeted support. Additionally, the review emphasizes the potential of machine learning algorithms in personalizing the learning experience, optimizing curriculum design, and facilitating adaptive learning systems. Overall, this paper contributes to the existing literature by providing a comprehensive overview of the application of machine learning algorithms in predicting students' performance and improving learning outcomes. The insights presented can guide educators and researchers in implementing effective machine learning techniques to enhance educational practices.

In this study the author Sokkhey. P at all (2020) [17] demonstrate the effectiveness of hybrid machine learning algorithms in predicting academic performance. By leveraging various techniques, the authors successfully improve the accuracy and reliability of the prediction models, making a valuable contribution to the field of educational data analysis and prediction of performance of the students.

References	Data Sets	Methods used	Overall view
[1]	Student data of Rajamangala University of Technology Thanyaburi, Pathumthani, Thailand	Big Data Analytics Microsoft Azure, Feature Selection	The result of this experiment was evaluated by the confusion matrix and the overall accuracy with ten-fold cross validation. Mutual information in the feature selection method with neural network classifier gave the best overall accuracy at 90.60%
[2]	Student examination data set	Naive Bayes, KNN, SVM, MLP, XGBoost	A large dataset from different schools that contain student results from more academic years can give a better understanding of student's academic success prediction.
[3]	The National Coordinating Committee on School Health and Safety data set.	BMI, CDC	pure health benefits, perhaps schools' adherence to stricter physical activity and nutrition recommendations will occur before we fully understand the connection between achievement and obesity
[4]	Academic performance evaluation data set	Data Mining, Machine Learning; K-Nearest Neighbor Classifier, Extra Tree Classifier, Ensemble Technique.	The best accuracy among these different machine learning classifiers is 86.83% from Naive Bayesian and 91.76% in boosting ensemble technique.
[5]	student's academic performance data set	linear regression models	MLR outperforms SLR for the early prediction of student performance
[6]	Student's performance level data set	Data Mining, Machine Learning, K Nearest Neighbor, Naive Bayes, Decision Tree	KNN, DT and Naive Bayes classifiers were used on the dataset of 230 students of Kabul University to predict their GPA as high, medium and low
[7]	Student's performance data set	Decision Tree, Naive Bayes, Artificial Neural Network,SVM, Random Forest	Dataset having fewer observations, SVM linear, SVM polynomial, and NB outperforms the other models under study, whereas for a dataset having a large number of observations, DT and RF outperforms the other models under study
[8]	student achievement data set	Regression, Decision tree, Entropy and KNN classifier	Regression, Decision tree, Entropy and KNN classifier are used. This process can help the instructor to decide easily about performance of the students and schedule better method for improving their academics.
[9]	student learning environment data set	NB, BN, DT, CR, PART, ADTree, J48, and RF	data mining and machine learning techniques have been proposed for analyzing and monitoring massive data giving rise to a whole new field of big data analytics.
[10]	Student performance data set	DT, Rule learning algorithm	decision trees and rule-based learning algorhythms were the common machine learning methods used in studies.
[11]	Academic organization data set	Artificial intelligence, machine learning, neural networks	They confirmed that the performance of neural networks increases with increase in dataset size.
[12]	Student 's Performance Kaggle dataset	Data mining, Decision Tree, K-Nearest Neighbor, Neural Network, Random Forest, Support Vector Machine	They compared the results of six algorithms out of which Deep Neural Network outperformed with 84% as accuracy.
[13]	student performance university data set	Machine learning, Random Forest, SVM	Accuracy of the random forest (RF) classifier is more than the other classification method such as support vector machine (SVM).
[14]	Academic performance data set	Multiple Linear Regression,Decision Tree Regression,K-nearest neighbor,Random Forest, SVM,	By analyzing the results from the models, it became apparent that the Multiple Linear Regression Model gives the optimal solution
[15]	Student's final examination data set	KNN, SVM	Based on the results, we can conclude that both SVM and KNN regression would suit this type of problem
[16]	Student education performance data set.	Decision Tree,Artificial Neural Network,Logistic regression,Naive Bayes,SVM,Random Forest,K-Nearest Neighbor,Multiple regression,	It is important to accurately rank machine models based on their prediction capabilities in predicting students' performance prediction and subsequent decision making.
[17]	student performance in mathematics subjects data set	k-fold cross-validation; principal component analysis, NB, RF.	By combining the baseline models with principal component analysis, and evaluated by k-fold crossvalidation, the proposed hybrid models produced a high performance which shows itself as a potential algorithm for solving prediction and classification problem.

Table 2. Overall view

III. Methodology

In order to perform student performance prediction, we are required to collect data from relevant sources such as educational institutions or learning management systems. This data undergoes various steps of pre-processing, which aims to make it more suitable for machine analysis and prediction compared to its original form.

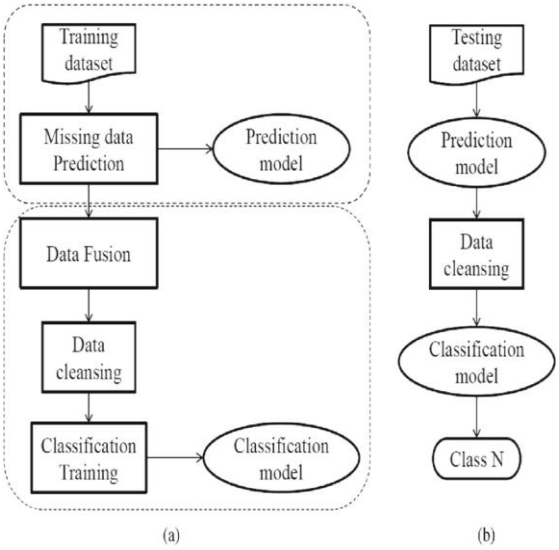


Figure 3. Methodology

A. Dataset

This dataset can be used for analysing the relationships between these variables and academic performance. For example, it could be used to investigate whether there are gender or race/ethnicity-based disparities in scores, or to examine the impact of parental education or test preparation on student performance.

The dataset provided includes information about students' gender, race/ethnicity, parental level of education, lunch type, test preparation course, math score, reading score, and writing score. Here is a breakdown of the variables:

Gender: Indicates the gender of the student (male or female).

Race/Ethnicity: Represents the racial or ethnic group to which the student belongs, categorized into groups A, B, C, D, and E.

Parental Level of Education: Describes the highest level of education achieved by the student's parents, such as high school, some college, associate degree, bachelor's degree, or master's degree.

Lunch: Indicates the type of lunch the student receives, categorized as "standard" or "free/reduced."

Test Preparation Course: Specifies whether the student completed a test preparation course or not, categorized as "completed" or "none."

Math Score: Represents the score achieved by the student in the math subject.

Reading Score: Represents the score achieved by the student in the reading subject.

Writing Score: Represents the score achieved by the student in the writing subject.

B. Data Pre-processing

Data preprocessing plays a critical role in predicting student performance. It involves several steps aimed at enhancing the quality and suitability of the data for subsequent analysis and modeling. Here are some common preprocessing steps for predicting student performance:

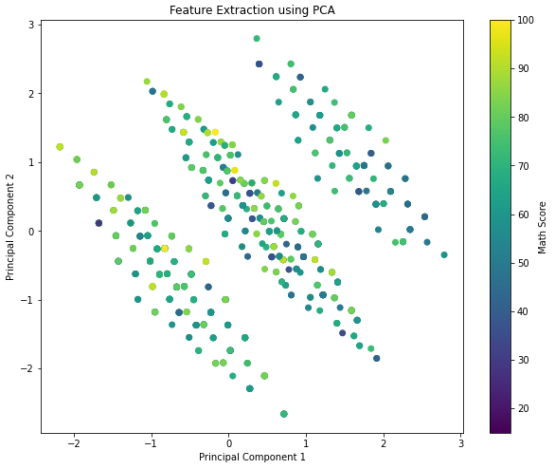
- **Handling Missing Values:** Identify missing values in the dataset and decide how to handle them. Missing values can be imputed using techniques such as mean or median imputation, or more advanced methods like regression imputation or multiple imputation.
- **Encoding Categorical Variables:** Convert categorical variables (e.g., gender, race/ethnicity, parental level of education) into numerical representations that can be understood by machine learning models. This can be achieved through one-hot encoding, label encoding, or ordinal encoding.
- **Feature Scaling:** Scale numeric features to a common range to prevent certain features from dominating others during model training. Common scaling techniques include standardization.
- **Handling Outliers:** Identify and handle outliers in the data. Outliers can be treated by removing them if they are due to data entry errors or by applying techniques like data transformation to minimize their impact on the model.
- **Feature Engineering:** Create new features or transform existing ones to improve their

relevance and predictive power. This can involve creating interaction terms, polynomial features, or deriving domain-specific features.

- **Train-Test Split:** Split the dataset into training and testing sets to evaluate the model's performance on unseen data. The training set is used to train the model, while the testing set is used to assess its performance.
- **Feature Selection:** Select the most relevant features to reduce dimensionality and improve model interpretability and performance. This can be done through techniques like correlation analysis, feature importance from tree-based models, or using domain knowledge to prioritize specific features.

C. Feature extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.



Plot1. Feature extraction using PCA

The scatter plot visualizes the relationship between the extracted principal components (PC1 and PC2) and the math scores of the students in the dataset. Each data point in the plot represents a student, and its position on the plot is determined by the values of PC1 and PC2.

The principal components are derived from the original features of the dataset using PCA, which is a dimensionality reduction technique. PCA finds the directions (principal components) along which the data varies the most, and these components capture the most significant information in the data. By plotting the students' positions based on PC1 and PC2, we can observe patterns or clusters in the data. The position of a student on the plot provides information about their relative values along PC1 and PC2. The scatter plot can help identify groups or patterns of students with similar math scores.

Additionally, the color of each data point represents the math score of the corresponding student. The color mapping provided by the colorbar on the right side of the plot allows us to interpret the math scores based on the color scale. For example, darker shades may indicate higher math scores, while lighter shades may represent lower scores.

Overall, the scatter plot provides a visual representation of the relationship between the extracted features (PC1 and PC2) and the math scores. It helps identify any underlying structures or trends in the data and can provide insights into how the extracted features are related to the math scores of the students.

D. Model selection

Logistic Regression: logistic regression is applied to student performance prediction, it aims to classify students into different performance categories based on their features. In simple words, it tries to find a relationship between the student's features and the probability of belonging to a specific performance category. Logistic regression works by analyzing the data to estimate the coefficients of the logistic regression equation. This equation calculates the probability of a student belonging to a certain performance category. The equation takes the form of:

Probability = 1 / (1 + exp(-(Intercept + Coefficient1 * Feature1 + Coefficient2 * Feature2 + ...)))

The coefficients represent the impact of each feature on the probability of belonging to a specific performance category. The intercept represents the baseline probability when all features are zero.

During the training process, logistic regression learns the optimal values of the coefficients that maximize the likelihood of the observed performance categories given the features.

Once the logistic regression model is trained, we can use it to predict the performance category of new students by inputting their feature values into the equation. The model will calculate the probability of belonging to each performance category and classify the student into the category with the highest probability.

Random Forest: Random Forest is an ensemble learning method used for classification and regression tasks. It combines multiple decision trees to make predictions. In scikit-learn, the Random Forest algorithm is implemented through the 'RandomForestClassifier' class for classification tasks.

Unlike some other algorithms, Random Forest does not have a single mathematical formula that represents the entire ensemble. Instead, it works by creating a collection of decision trees, where each tree is trained on a random subset of the data and features.

During prediction, each decision tree in the ensemble independently makes a prediction based on its trained subset of data and features. For classification tasks, the final prediction is determined by aggregating the individual predictions of the trees using majority voting. The class with the most votes across all trees is selected as the final prediction. For regression tasks, the final prediction is usually the average or median of the predictions made by the individual trees. The strength of Random Forest lies in its ability to reduce overfitting and improve prediction accuracy. Each decision tree in the ensemble learns from a different subset of the data, reducing the impact of outliers and noise. Additionally, Random Forest provides an estimate of feature importance, indicating the relative contribution of each feature in the ensemble's predictions.

Decision tree: decision tree is applied to student performance prediction, it aims to create a model that can predict the performance of students based on certain features or attributes. The decision tree algorithm mimics the decision-making process similar to how humans make decisions. In this context, let's consider a dataset of students with features such as gender, race/ethnicity, parental level of education, lunch type, and test preparation course, along with their scores in math, reading, and writing.

The decision tree algorithm analyzes the data and creates a flowchart-like structure. Each internal node of the tree represents a decision based on a specific feature, and each branch represents the outcome of that decision. The leaf nodes represent the final prediction or the student's performance level. During the training process, the decision tree algorithm learns the optimal decision rules by finding the most informative features and splitting the data accordingly. It aims to create homogeneous subsets of data at each node, where students within the same subset have similar performance levels. Once the decision tree model is trained, we can use it to predict the performance of new students by providing their feature values as input. The model will follow the decision rules and traverse the tree to reach a leaf node, which will provide the predicted performance level for the given students.

Support Vector Machine: When SVM is applied to a student prediction dataset, it aims to classify students into different categories or predict their performance based on certain features. Let's consider an example where we have a dataset of students with features like gender, parental level of education, test preparation course, and previous scores in math, reading, and writing.

Using SVM, we want to build a model that can predict whether a student will perform well or poorly based on these features. The model will learn from the data and create a decision boundary or hyperplane that separates students who are likely to perform well from those who are likely to perform poorly. SVM will analyze the data and find the best hyperplane that maximizes the margin, which is the distance between the decision boundary and the nearest data points. This margin allows SVM to generalize well to new, unseen students. Once the SVM model is trained, we can use it to predict the performance of new students by inputting their feature values. The model will classify them into the appropriate category based on their position relative to the decision boundary.

K-Nearest Neighbors: K-Nearest Neighbors is applied to student performance prediction, it aims to predict the performance of a student based on the performance of their nearest neighbors in the dataset. In simple words, KNN works by finding the K nearest neighbors to a given student based on their feature values. These neighbors are the students in the dataset whose feature values are most similar to the given student. To make a prediction for the performance of the given student, KNN looks at the performance levels of its K nearest neighbors. It assigns the predicted performance level to the most common performance level among the neighbors. For example, if a majority of the neighbors have a high-performance level, the given student is predicted to have a high-performance level as well.

The value of K determines the number of neighbors to consider. A higher value of K means more neighbors

are considered, which can provide a more stable prediction but may be less sensitive to local variations. On the other hand, a lower value of K can capture local patterns more accurately but may be more susceptible to noise.

E. Model Evaluation

The confusion matrix is a widely used and suitable method for assessing the performance of a classifier. It provides a comprehensive representation of the classifier's predictions and their alignment with the actual outcomes. The table below illustrates a generalized form of the confusion matrix.

Applying this technique allows us to derive a comprehensive and standardized evaluation framework. It encompasses various parameters to assess and compare systems or processes effectively. These parameters include:

- **Accuracy:** Accuracy of a classifier measures the correctness of its predictions, and it can be calculated by dividing the number of correct predictions by the total number of predictions.

$$\text{accuracy(a)} = \frac{tp+tn}{tp+tn+fp+fn}$$

- **Precision:** Precision measures the proportion of true positive predictions out of all positive predictions made by the classifier. It can be calculated by dividing the number of true positives by the sum of true positives and false positives.

$$\text{precision(p)} = \frac{\text{tp}}{tp+fp}$$

- **Recall:** Recall is a metric that measures the proportion of actual positive cases correctly identified by a classifier, reflecting its true positive rate. It indicates the classifier's ability to find all positive instances.

$$\text{recall(r)} = \frac{tp}{\text{tp+fn}}$$

- **F1 score:** The F1 score is a measure of a classifier's performance that combines both recall and precision into a single metric. The formula for recall is:

$$\text{F1 score} = \frac{2.\text{p.r}}{p-r}$$

IV. Student Performance Analysis with Python

Pandas: Pandas is a popular open-source data analysis and manipulation library for Python. It provides powerful tools for working with structured data, including data frames, and offers a wide range of functions for data cleaning, transformation, and analysis.

Sklearn: Scikit-learn (sklearn) is a Python library that provides a comprehensive set of tools for machine learning and statistical modeling. It offers a wide range of algorithms for classification, regression, clustering, and dimensionality reduction tasks, along with utilities for data preprocessing and model evaluation.

	Predicted class1	Predicted class2
Actual class 1	True positive(tp)	False negative(fp)
Actual class 2	False positive(fp)	True positive(tn)

Seaborn: Seaborn is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating aesthetically pleasing and informative statistical graphics, making it easier to explore and visualize data in a visually appealing manner.

Matplotlip.pyplot: It is a sub-module of the Matplotlib library that provides a MATLAB-like plotting interface in Python. It allows for the creation of various types of plots, including line plots, scatter plots, bar plots, and histograms, among others.

A. Setting Up Environment for Student performance Analysis Using Python

To perform sentiment analysis using Python, the following components need to be downloaded and installed correctly: •
Python 2.6 or above in the desired location.

• NumPy

• Pandas

• Scikit-learn library

By installing these components, you can create an environment suitable for Student performance analysis tasks on exams data using Python.

B. Training the model

The training process for the student performance analysis model involves several steps. Firstly, the dataset is prepared by handling missing values, encoding categorical variables, and splitting it into features and the target variable. The data is then divided into training and testing sets to assess the model's generalization ability. Linear regression is selected as the algorithm for predicting the total score, and feature scaling is applied if necessary. The model is trained using the training data, learning the relationship between features and the target variable. Model evaluation is performed using the testing data, calculating metrics such as accuracy, precision, recall, and F1 score. These metrics measure the model's prediction performance. The results provide insights into the model's effectiveness in predicting the total score. Overall, the training process involves data preparation, splitting, algorithm selection, feature scaling, model training, evaluation, and interpretation of the results.

IV. RESULTS

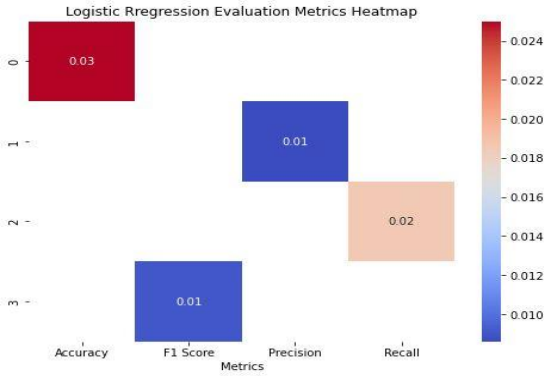


Fig 1: Logistic Regression

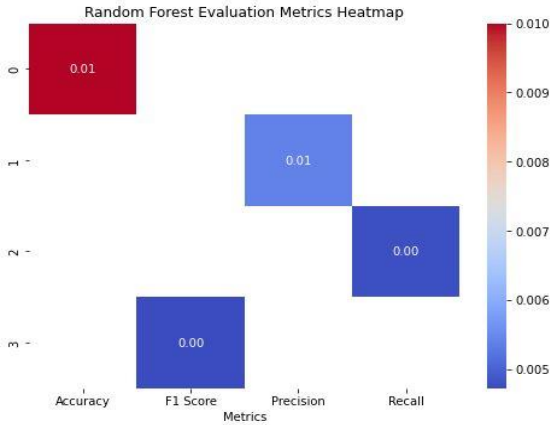


Fig 2: Random Forest

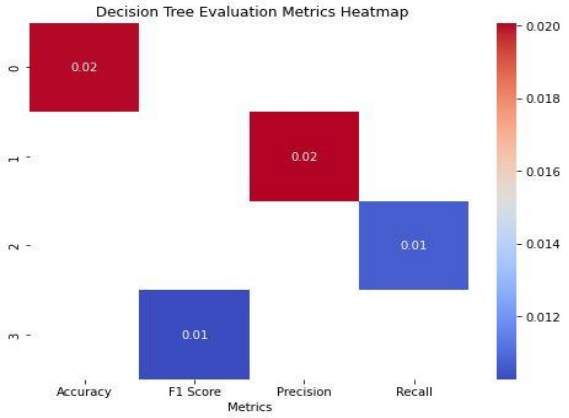


Fig 3: Decision Tree

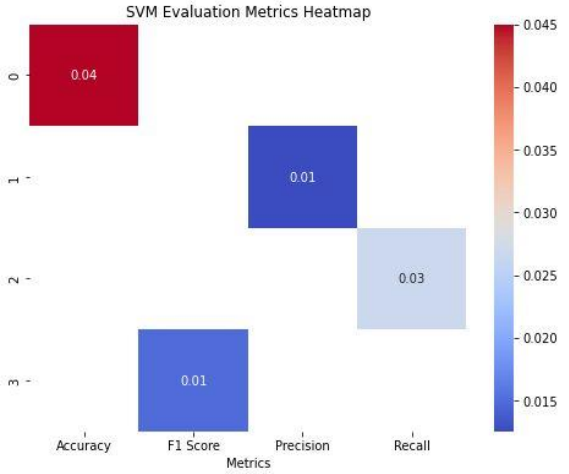


Fig 4: Support Vector Machine

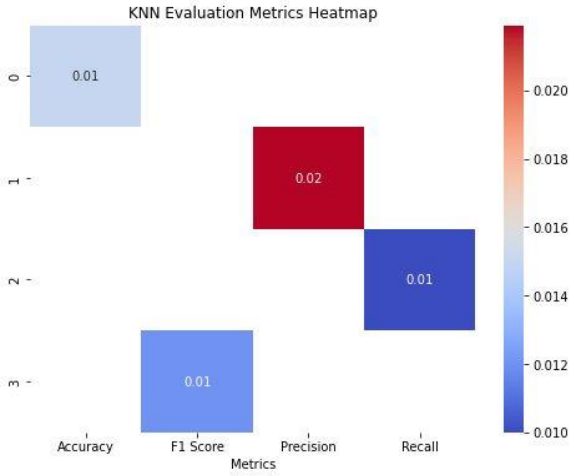


Fig 5: K – Nearest Neighbor

From Above Five classification models have been created and tested using five machine learning techniques, fully connected feed, Logistic Regression, Random Forest, Support Vector Machine, K – Nearest Neighbor and Decision Tree. Above Figures shows the accuracy and the performance measures for each model as well as the confusion matrices.

V. CONCLUSION

This paper investigates the effectiveness of various feature selection, dimensionality reduction, and classification techniques for student performance prediction using the "exams.csv" dataset. The study examines the performance measures utilized to evaluate the classifiers and compares the advantages and disadvantages of different feature selection, dimensionality reduction, and classification approaches. The paper categorizes the feature selection techniques based on their application in data mining tasks, search strategy, and evaluation criteria. It highlights that filtering techniques exhibit high efficiency and performance in identifying optimal feature subsets compared to wrapper and embedded techniques. The review emphasizes the utilization of hybrid techniques to eliminate noisy, redundant, and insignificant features in the "exams.csv" dataset for student performance prediction. In terms of classification techniques, the study explores the application of logistic regression, decision tree, random forest, K-nearest neighbors (KNN), and support vector machines (SVM) for student performance prediction. It discusses that the classification techniques to demonstrate superior success rates and low computation time for accurate prediction of student performance. These techniques offer significant improvements in the diagnosis and classification accuracy of student performance, assisting educational professionals, teachers, and stakeholders in effective decision-making.

REFERENCES

[1] Punlumjeak, W., Rachburee, N., & Arunrerk, J. (2017). Big data analytics: Student performance prediction usingfeature selection and machine learning on microsoft azure platform. *Journal of Telecommunication, Electronic and Com*

[2] Ahammad, K., Chakraborty, P., Akter, E., Fomey, U. H., & Rahman, S. (2021). A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. *International Journal of Computer Science and Information Security (IJCSIS)*.

[3] Taras, H., & Potts-Datema, W. (2005). Obesity and student performance at school. *Journal of school health*.

[4] Singh, R., & Pal, S. (2020). Machine learning algorithms and ensemble technique to improve prediction of students performance. *IJATCSE*.

[5] Naomi, J. F., Shivaani, V., Abhirami, V. S. D., & Thiruvarasu, T. (2021, March). Scrutinizing Students Performance using Machine learning. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE.

[6] Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019, February). Comparative study of supervised learning algorithms for student performance prediction. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE.

[7] Ismail, L., Materwala, H., & Hennebelle, A. (2021). Comparative Analysis of Machine Learning Models for Students' Performance Prediction. In *Advances in Digital Science: ICADS 2021*. Springer International Publishing.

[8] Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021, February). Prediction of students performance using machine learning. In *IOP conference series:Materials science and engineering*. IOP Publishing.

[9] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*.

[10] Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: a systematic review. *IEEE Access*.

[11] Agrawal, H., & Mavani, H. (2015). Student performance prediction using machine learning. *International Journal of Engineering Research and Technology*.

[12] Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting student's performance using machine learning algorithms. *International Journal of Intelligent Systems and Applications*.

[13] Rai, S., Shastry, K. A., Pratap, S., Kishore, S., Mishra, P., & Sanjay, H. A. (2021). Machine learning approach for student academic performance prediction. In *Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020)*. Springer Singapore.

[14] Chauhan, N., Shah, K., Karn, D., & Dalal, J. (2019, April). Prediction of student's performance using machine learning. In *2nd International Conference on Advances in Science & Technology (ICAST)*.

[15] Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017, April). Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE)*. IEEE.

[16] Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. *Journal of Information and Technology*.

[17] Sokkhey, P., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *Int. J. Adv. Comput. Sci. Appl*.