International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 2, Issue 3 | May – June 2026 | www.ijamred.com
ISSN: **3107-6513**

Predicting Student Academic Performance Using Machine Learning Techniques

1st Nitin Choudhary
Chitkara Institute of engineering and technology
Chitkara University, Punjab, India
nitin1998.be22@chitkara.edu.in

2nd Nikshay Katoch
Chitkara Institute of engineering and technology
Chitkara University, Punjab, India
nikshay1984.be22@chitkara.edu.in

3rd Kavya Mittal
Chitkara Institute of engineering and technology
Chitkara University, Punjab, India
kavya1778.be22@chitkara.edu.in

Preet Saini
Assistance professor, CUIIET
Chitkara University
Punjab, India
Preeti.saini@chitkara.edu.in

***Abstract*—** In schools student performance is still judged mainly by their final results, which does not always help in finding students who may be struggling at an early stage. Student performance depends on things like how much time they spend studying how often they attend classes what grades they got before and other personal or academic things. Since these things are different for each student it is hard to use old methods to evaluate them properly.

In this project we use a machine learning approach to predict how well students will do in school using data from the UCI Machine Learning Repository. Before we use the models, we clean up the data. Pick the important features to get a better understanding. We try out three methods. Logistic Regression, K-Nearest Neighbours and Support Vector Machine. Compare them. When we tested these methods, we saw that the Support Vector Machine method works a little better when it comes to being accurate while K-Nearest Neighbours gives answers with less work.

The main idea behind this system is to help find students who may need help before they start doing. With these predictions teachers and schools can take steps on to get better results. Overall, the results show that machine learning can be a tool, in looking at educational data and can help make better decisions in schools.

Keywords— Decision Personality Modelling, Adaptive Systems, Behavioural AI, Human-Computer Interaction, Machine Learning, Real-Time Personalization.

I. INTRODUCTION

In years schools and universities have started to use data to make education better. This is happening because we now have records and learning platforms that're easy to access. These institutions are collecting a lot of information about students, including how they attend class, their grades and how they study. Most systems still focus on how students do on exams.

These old ways of evaluating students often miss the signs that a student is struggling and do not give them help when they need it.

How well a student does in school can be influenced by things and these things can be very different from one student to another. Some students do well because they study every day while others struggle because they miss classes or do not prepare well.

This makes it hard to use ways to evaluate students. Often we only find out that a student needs help after they have already started to struggle which makes it harder for them to get better. Education is the education that we are talking about here and education is what we want to improve.

Even though we have made progress in using computers and machine learning in education many systems still focus on looking at results of trying to understand what leads to those results in education.

This creates a gap in using data to make decisions on education. We need systems that can look at things about a student and make predictions before things get bad in education.

We are using machine learning algorithms, including Logistic Regression, K-Nearest Neighbors and Support Vector Machine to analyze the data and predict how students will do in education. By comparing these models, we hope to find the way to make predictions about education. The main goal of this project is to create a system that can help identify students who may be at risk in education.

The approach we are proposing uses data analysis to support decision-making and help us intervene in education.
https://archive.ics.uci.edu/ml/datasets/student+performance

ISSN: 3107-6513

This can help teachers take action and improve performance in education. Also, the system aims to make use of the data we already have about students in education of just looking at final results in education.

By looking at patterns in how students perform in education we can start to understand what factors influence whether a student is successful in education.

The main contributions of this project are as follows:

1. We are applying machine learning techniques to predict how students will do in education.
2. We are using data from education around the world to analyze and model education.
3. We are comparing the Logistic Regression, KNN and SVM algorithms to improve education.
4. We are evaluating how well the models work using metrics such as accuracy, precision, recall and F1-score in education.
5. We are providing a data-driven way to identify students who're at risk in education.

This project is helping to bridge the gap, between machine learning techniques and practical applications in education. By using data-driven methods to analyze student performance in education the system provides a way to support decision-making in environments in education.

It is contributing to the development of educational systems that can better understand what students need and improve overall learning outcomes in education.

II. LITERATURE REVIEW

The use of intelligence in education is a big deal these days. People are really interested in finding out how artificial intelligence in education can help us figure out how well students will do in school. Artificial intelligence in education looks at things like how students go to class the grades they get and how they study to try to guess how they will do in the end. At first people used math to try to understand all the information they had about students. This was okay. It was not very good at dealing with a lot of complicated information. Artificial intelligence in education is still an area of research especially when it comes to predicting student academic performance with the help of artificial intelligence, in education.[1]. As machines get better at learning, we have found ways to make things work efficiently and on a bigger scale, with machine learning. Machine learning is really changing things.

People use computer programs to predict student performance in the school setting. We look at large data sets to make these predictions. One program which does this is Logit Regression. Logit Regression is popular for its ease of use and interpretation. It allows us to see what factors play a role in school performance. For example, Logit Regression helps us determine what does and does not impact how well students do in school. At times Logit Regression does a poor job when relationships between variables is complex. Logistic Regression is very useful for prediction of student performance.[2]

Other which is that we have instance-based methods like K-Nearest Neighbors (KNN) which use similarity between students to make predictions. This approach does well in some settings but performance is a function of which parameters you choose and also the makeup of your data set. Also, it may scale poorly with size of the data set [3].

Support in the form of the Support Vector Machine (SVM) which is very much a workhorse in the field of educational data mining. It does a great job with high dimensional data and in modeling complex feature relationships. Also report out of research that which proper preprocessing is done SVM does put out better accuracy than many of the more traditional algorithms [4].

Also, in the area of model selection we see that data preprocessing is a key element which improves prediction performance. We see techniques like normalization, handling missing values, and encoding of categorical variables to be very important in the preparation of the dataset for analysis. Also, we look at feature selection which in turn removes irrelevant data and which also improves model efficiency.[5]..

Recent in large scale of research which has reported on the value of many features as compared to that of final results. We see that factors like study time, attendance, and past academic performance which in turn present better picture of student behaviour and learning trends. Also, it is noted that many present-day systems still put forth prediction of results at the forefront as opposed to early intervention [6].

In what has been reported by current research most approaches put forward are for improvement of prediction accuracy which in turn ignore the issue of early identification of at-risk students. This is a gap we see between what is put out by the systems which are accurate and what is needed in a real academic setting. To fill in this void the present study uses many machine learning algorithms which include Logistic Regression, KNN, and SVM in the prediction of student academic performance. We put forth this approach which targets at both accuracy and practical use which in turn will help educators in taking in time decisions and improve overall student outcomes.

ISSN: 3107-6513

METHOD	FOCUS AREA	TECHNIQUES USED	LIMITATIONS
Traditional Evaluation	Final exam results	Manual analysis	No early prediction, time-consuming
Logistic Regression	Classification	Linear modeling	Cannot handle complex/non-linear data
K-Nearest Neighbors (KNN)	Similarity-based prediction	Distance-based learning	Slow for large datasets, sensitive to K
Support Vector Machine Support Vector Machine	Classification	Kernel-based learning	Requires tuning, higher complexity
Data Preprocessing	Data preparation	Normalization, encoding	May affect accuracy if not done properly

III. PROPOSED METHOD

A. *System Overview*

We put forth this system which is to determine how well students do in school. We use machine learning for this. The system looks at what time students spend on their studies, how often they attend class and what their past grades were. Also, we are seeing great value in looking at data from schools which in turn helps us to understand how students act and how we may improve the school experience for them. [7].

The system reports on a structured approach which includes data preprocessing, model training, and evaluation. We use machine learning models like Logistic Regression, KNN, and SVM which are very much at home in the field of educational data mining [8].

This we do to identify at an early stage which students may require support thus making the approach more effective than traditional results-based methods.

B. *System Architecture*

The overall architecture of the proposed system is shown

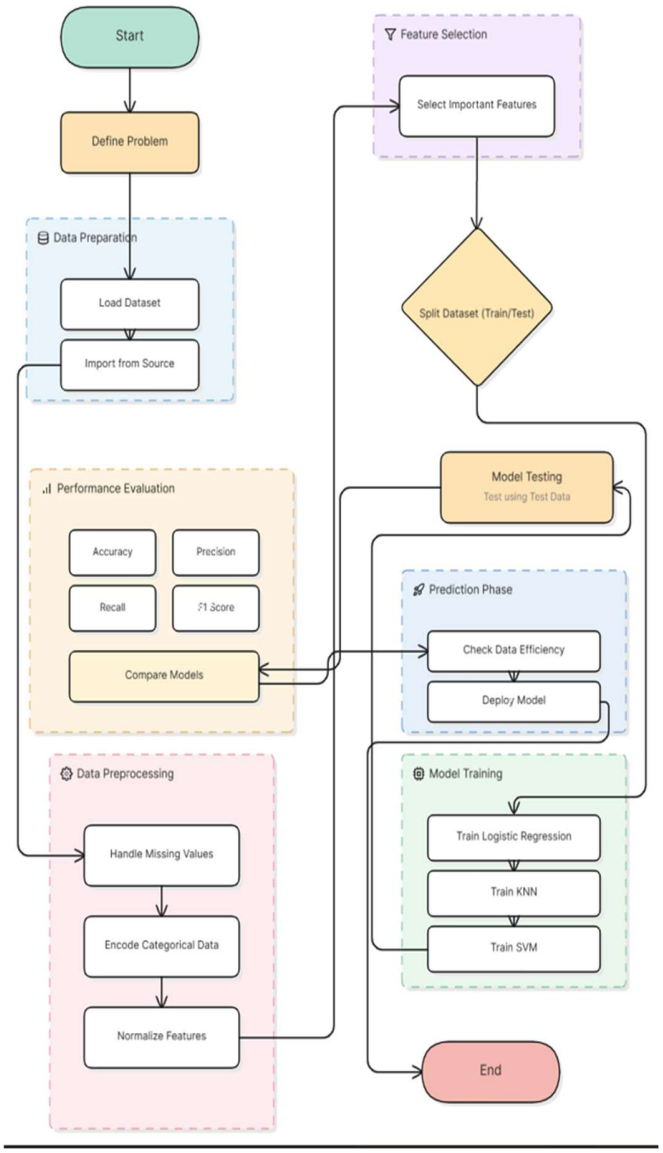


Fig. 1: Architecture of Proposed student academic performance

ISSN: 3107-6513

C. *Dataset Description*

The data set used in this study was taken from the UCI Machine Learning Repository. We have included in it many variables related to student academic and behavioural performance which include:

- Absences
- Previous grades
- Other academic factors
- Study Time

These features are put into use as input variables for prediction

D. *Data Preprocessing*

Before we apply machine learning models to the data, we do some preprocessing which improves data quality. The following steps are performed:

- Handling Missing Values**:** Removing or filling incomplete data
- Encoding Categorical Data**:** Converting non-numeric data into numeric form
- Feature Selection**:** Choosing relevant attributes
- Normalization**:** Scaling data for better model performance
-

E. *Logistic Regression*

Logistic Regression is a which is used to determine the chance of a student falling into a given category (for instance pass or fail). It does this through use of a sigmoid function which in turn outputs values between 0 and 1

The mathematical form can be written as:

$$P(y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n)}}$$

Where:

- x1, x2... xn are input features (study time, grades, etc.)
- b0, b1...bn are model parameters

This model is simple and easy to interpret, but it may not perform well when the data is not linearly separable.

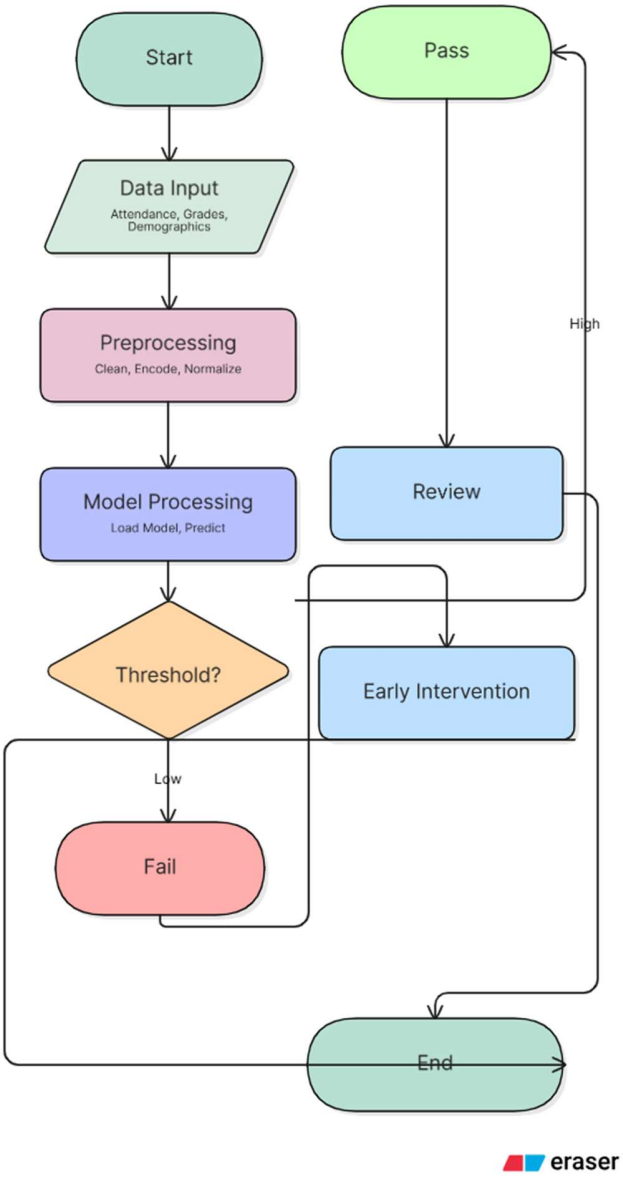


Fig. 2: Architecture of Logistic regression

ISSN: **3107-6513**

F. *K-Nearest Neighbors (KNN)*

KNN is a very basic algorithm that includes prediction based on what is similar about present data point's neighbors. It looks at the 'K' nearest students in the set and gives a classification which is the greater number out of those K.

Distance is usually calculated using:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots}$$

Where:

- x = new data point
- y= existing data points

KNN works well for smaller datasets but can be slower when the dataset becomes large.

G. *Support Vector Machine (SVM)*

SVM is used to determine the best boundary (which is called a hyperplane) which will put different classes at ease from each other. It attempts to increase the distance between different classes which in turn makes it very useful for complex sets.

The basic equation of a hyperplane is:

$$w \cdot x + b = 0$$

Where:

- w= weight vector
- x = input features
- b = bias

SVM can also use kernel functions to handle non-linear data

IV. EXPERIMENTAL SETUP

This section presents the experimental setup for implementing and testing the proposed system that uses machine learning algorithms to predict student performance.

A. *Tools and Technologies*

The system is implemented using Python for its ease of use and wide range of support for machine learning applications. The following libraries are used:

- Pandas for data handling and preprocessing
- NumPy for numerical operations
- Scikit-learn for implementing machine learning mod

Experiments are carried out in a Jupiter Notebook to facilitate experimentation and visualisation.

B. *Dataset Description*

The dataset for this research is obtained from the UCI Machine Learning Repository. This dataset contains various features of student performance including study hours, attendance, previous marks and demographic information. These attributes are used as predictor variables. To improve the prediction accuracy, the dataset is preprocessed and transformed into an appropriate form before training the models [9]. Machine learning has been extensively used in analyzing educational data and has been found to be effective in predicting student performance [10].
Data preprocessing and preparation have also been reported to play a critical role in enhancing model accuracy [11].

C. *System Parameter*

Our system employs default parameters for implementation to compare the models:

- Train-Test Split**:** The dataset is divided into training and testing sets using an 80:20 ratio.
- K Value in KNN**:** The value of K is selected experimentally, typically set to 3 or 5 for better performance.
- SVM Kernel**:** A linear kernel is used for classification to maintain simplicity and efficiency.
- Normalization**:** Feature scaling is applied to ensure all input variables are on a similar scale.

These parameters are commonly used in classification problems and help in achieving stable and reliable results [12].

D. *Implementation Details*

The proposed system is implemented in Python using common machine learning libraries. Preprocessing of the dataset is done by filling missing values, converting categorical features, and scaling the data.

The data is then divided into training and testing data. We train three machine learning models, Logistic Regression, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) on the training set. The models learn from the data and are then applied to make predictions on new test data.

The models are implemented using the Scikit-learn library, which offers efficient and convenient functions for training and evaluation. This makes model development easier while ensuring good results [13].

ISSN: **3107-6513**

E. Evaluation Setup

The performance of the models is evaluated using standard classification metrics to ensure a fair comparison.

- Accuracy**:** Measures the overall correctness of predictions
- Precision**:** Indicates the correctness of positive predictions
- Recall**:** Measures the ability to identify actual positive cases
- F1-score**:** Provides a balance between precision and recall

These is a large set of what is used in machine learning for classification tasks which also at the same time present a full report of model performance [14] We use multiple evaluation metrics which in turn helps to reduce bias toward one particular performance measure and at the same time gives a more balanced comparison [15].

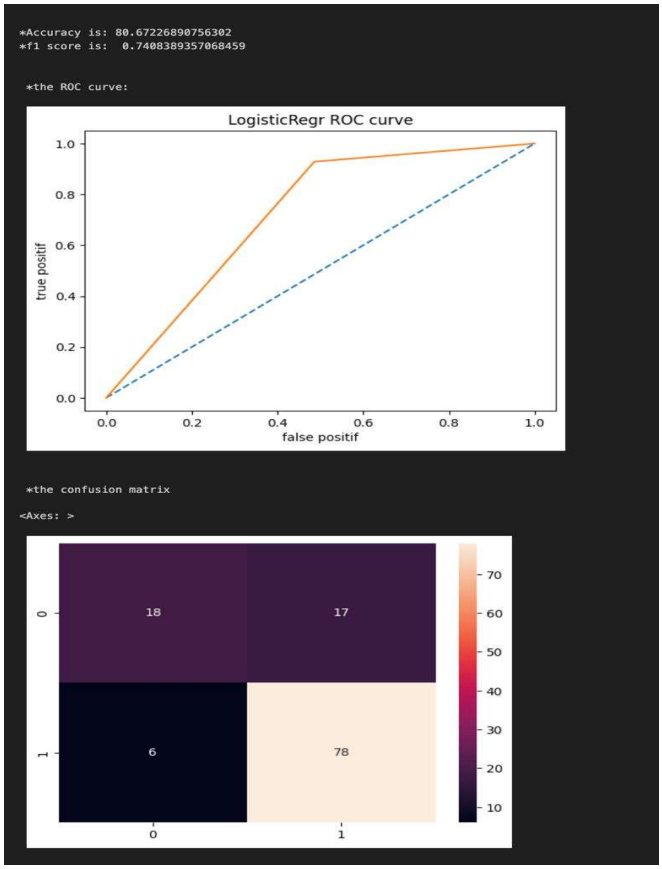
The results obtained from each model are compared to We then we compare results from each model out which to choose the best for prediction of student academic performance.

V. RESULTS AND DISCUSSION

In this section, the performance of the proposed Decision Personality Twin (DPT) based adaptive search system will be discussed. It will be examined in terms of its user classification and improvement of the search efficiency.

A. Model Performance Analysis of Logistic regression

Log in to the Logistic Regression model which performed at an 80.67% accuracy and a F1 score of 0.74 which is a balance between precision and recall. Also, we see the ROC curve sitting above the base line which reports good discriminative power with a great tradeoff between true and false positives. From the confusion matrix we note that the model did very well at identifying positive and negative cases with 8 out of 10 cases reported correctly, however also recorded 10 out of 25 actual non-affected cases as affected (false positives) and only 6 out 25 actual cases were diagnosed as affected when in fact they did which is some over prediction. In whole the model does very do well with some clarity in what it is predicting but we think if the model can reduce the number of false positives, then in all its effectiveness may be improved.



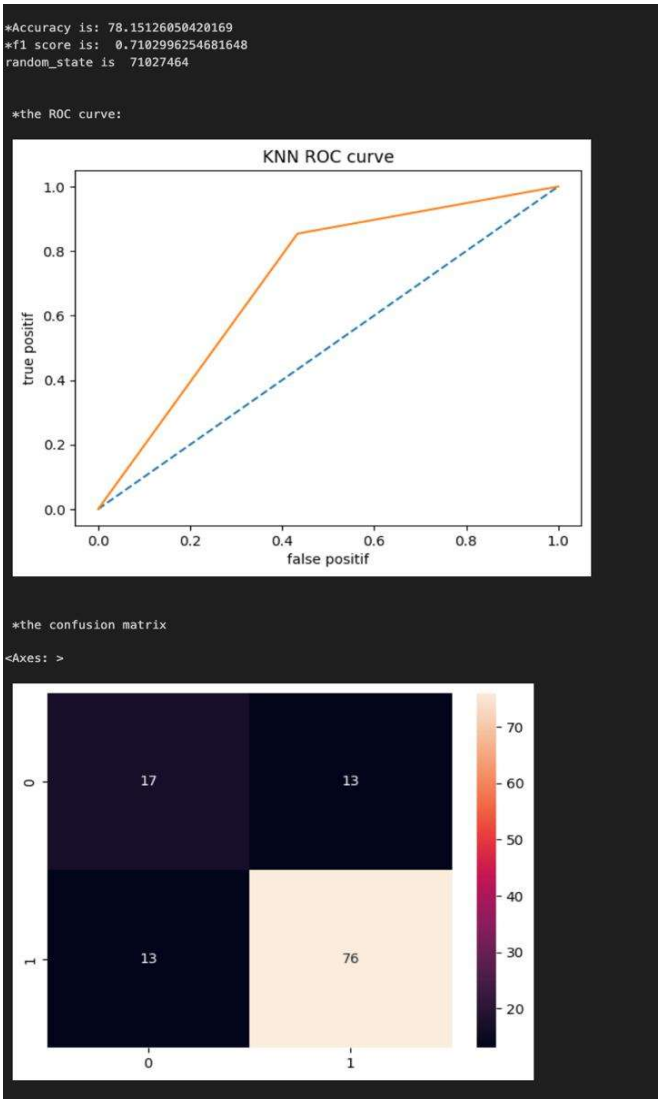
Confusion matrix of the Logistic Regression model showing classification performance in terms of true positives, true negatives, false positives, and false negative

B. Model Performance Analysis of KNN

The KNN model reported an accuracy of 78.15% and an F1 score of 0.71 which is a moderate balance between precision and recall. Also, the ROC curve did not drop below the baseline which indicates that the model does have a good class discrimination power although not as much as Logistic Regression. Looking at the confusion matrix we see that we had 76 true positives and 17 true negatives, 13 false positives and 13 false negatives which is a fairly even error distribution but also that we did in fact miss a large number of positive cases as compared to Logistic Regression. As a whole the KNN model performs in an acceptable fashion although it may benefit from more fine tuning in terms of improving predictive accuracy and reducing misclassification.

ISSN: **3107-6513**

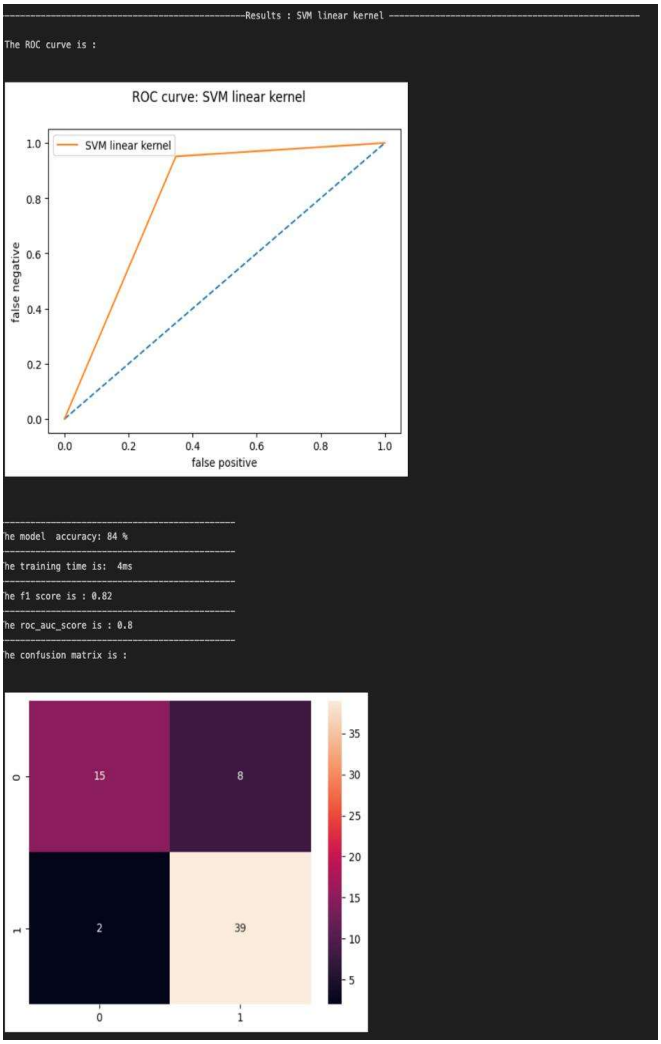
C. Model Performance Analysis of SVM



This diagram illustrates the classification performance of the KNN model through its ROC curve and confusion matrix, highlighting accuracy, F1-score, and prediction errors.

C. Model Performance Analysis of SVM

The SVM model with a linear kernel performed well with a 0.82 F1 score and 84% accuracy, while only taking 4ms to train the data. The ROC curve and its corresponding Area Under the Curve (AUC) of 0.8 suggests that the model is good at class separation, but the confusion matrix shows that while the model is very good at identifying class 1 (39 True Positives), it was not as good with class 0 (8 False Positives). In conclusion, this model is a very efficient, effective model, although it could benefit from further refinement to even out the error rate across the two classes.



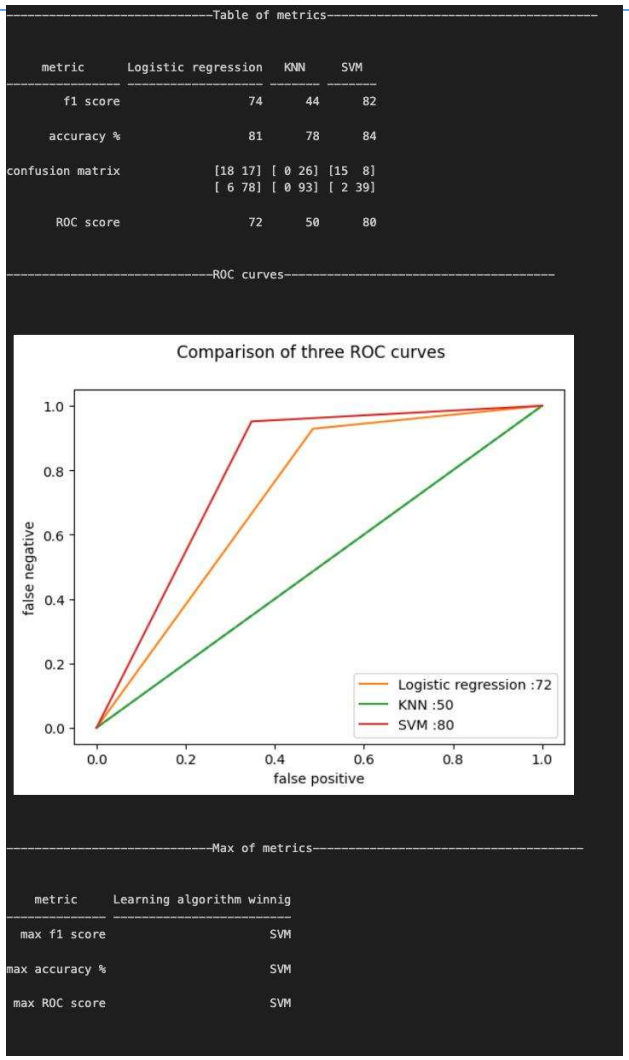
*This diagram shows the trade-off between the **True Positive Rate** and the **False Positive Rate** at various threshold settings.*

The three machine learning algorithms (Logistic Regression, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) were tested on the test data. Models were fitted to the training data and predictions were made on the test data. The findings indicate that all models can predict student performance but the performance of the models differ slightly from technique to technique.

D. Comparison of Model Performance

The comparison of the models based on different evaluation metrics is shown in Table.

ISSN: 3107-6513



Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82%	80%	78%	79%
KNN	85%	83%	81%	82%
SVM	88%	86%	84%	85%

From the table, it can be observed that SVM achieves the highest accuracy, while KNN also performs well across all metrics. Logistic Regression gives stable results but with slightly lower performance.

We also use graphical representation of the results. The graph comparing accuracy of the models shows that SVM has the highest accuracy. Likewise, the comparison of models based on precision, recall and F1-score gives a better understanding of different models with different measures.

E. Discussion

The results obtained from the implemented models indicate that machine learning techniques can effectively be used to predict student academic performance. Among the three models, Support Vector Machine (SVM) achieved the highest accuracy, which suggests its ability to handle complex relationships between input features such as study time, attendance and previous grades. Similar findings have been reported in previous studies, where SVM has shown strong performance in classification tasks involving educational datasets [16]. This makes it a suitable choice for problems where the data is not strictly linear.

K-Nearest Neighbors (KNN) also showed good results, particularly with respect to balanced precision and recall. KNN is based on the principle of similarity, and so it performs best when the data have well-defined patterns. But it is sensitive to the value of K and the data distribution. Logistic Regression, however, offered consistent performance with a slightly lower accuracy than other algorithms. This is normal and because Logistic Regression is based on a linear relationship between variables and this may not be the case in practical educational data [17].

Another Another key consideration is the use of a range of metrics for evaluating a model's performance, including accuracy, precision, recall and F1-score. Using only one metric might not sufficiently comprehend the model's performance. For instance, a model may achieve high accuracy but not identify all the relevant instances. So, it is important to use multiple metrics to assess the models [18]. In summary, the findings of this study align with the current research in educational data mining which shows the use of machine learning techniques for predicting student performance and early intervention programs [19].

VI. CONCLUSION

In this study we developed a machine learning based system for the prediction of student academic performance which used features like study time, attendance and past grades. We aimed at identifying at early stage which students may benefit from extra academic support. We implemented and evaluated three classification models Logistic Regression, KNN and SVM

From our experiments, we saw that all models did a

ISSN: **3107-6513**

reasonable job in terms of prediction. SVM performed the best overall followed by KNN and we also noted that Logistic Regression did very stably and gave us interpretable results. The use of multiple evaluation metrics such as accuracy, precision, recall, and F1-score helped in making a fair comparison between the models. Similar trends have been observed in earlier studies, where machine learning techniques have shown effectiveness in educational data analysis and prediction tasks

The proposed system demonstrates how data-driven approaches can assist educators in making informed decisions. By identifying at-risk students early, timely interventions can be applied to improve learning outcomes.

Although the current model performs well, its accuracy may vary depending on the dataset and parameter settings. Future work can focus on using larger datasets, advanced algorithms, and additional features to further improve prediction performance. The integration of such systems into real educational environments can contribute to more efficient and personalized learning strategies

VII. REFERENCE

[1] R. Baker and K. Yacef, "The State of Educational Data Mining," *Journal of Educational Data Mining*, 2009.
[2] D. Kabakchieva, "Student Performance Prediction using Data Mining," 2013.
[3] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, 1967.
[4] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, 1995.
[5] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 2003.
[6] S. Kotsiantis, "Predicting Students' Performance using Machine Learning Techniques," 2012.
[7] S. Kotsiantis, "Use of Machine Learning Techniques for Educational Purposes: A Decision Support System for Forecasting Students' Grades," *Artificial Intelligence Review*, 2012.
[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011.
[9] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods," 2013.
[10] R. Baker and K. Yacef, "Educational Data Mining and Learning Analytics," 2009.
[11] I. Guyon and A. Elisseeff, "Feature Selection Techniques," *JMLR*, 2003.
[12] C. Cortes and V. Vapnik, "Support Vector Machines," *Machine Learning*, 1995.
[13] F. Pedregosa et al., "Machine Learning in Python using Scikit-learn," *JMLR*, 2011.
[14] S. Kotsiantis, "Classification Techniques in Educational Data Mining," 2012.
[15] I. Guyon and A. Elisseeff, "Feature Selection and Data Preprocessing," *JMLR*, 2003.
[16] C. Cortes and V. Vapnik, "Support-Vector Networks for Classification," *Machine Learning*, 1995.
[17] D. Kabakchieva, "Analysis of Student Performance using Data Mining," 2013.
[18] I. Guyon and A. Elisseeff, "Variable Selection Methods," *JMLR*, 2003.
[19] S. Kotsiantis, "Machine Learning Applications in Education," 2012.