Visvesvaraya Technological University
Belagavi, Karnataka-590 018

A MINI PROJECT REPORT
On

‘STUDENT ACADEMIC PERFORMANCE PREDICTION’

Submitted
In partial fulfilment requirements for the award of the Degree

of
BACHELOR OF ENGINEERING
IN
INFORMATION SCIENCE AND ENGINEERING

By

PRATHEEKSHA KN	**4NM21IS111**
PRATHVI HEGDE	**4NM21IS113**

Under the Guidance of

Abhishek S. Rao
Assistant Professor

Department of Information Science and Engineering

NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that Ms. **Pratheeksha K N (**4NM21IS111) and Ms. **Prathvi Hegde(**4NM21IS113) has satisfactorily completed the Machine Learning Mini Project work entitled "**STUDENT ACADEMIC PERFORMANCE PREDICTION**" of Third Year, Bachelor of Engineering in Information Science and Engineering at NMAMIT, Nitte in the academic year 2023 - 24.

Project Guide
Mr. Abhishek S. Rao
Assistant Professor

Head, Dept. of ISE
Dr. Ashwini B
Associate Professor

ABSTRACT

The goal of this project is to predict the grades of students based on several features such as weekly study hours, attendance, notes, and reading habits. The grades are categorized into four classes: A, B, C, and Fail. By using algorithms like K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest, the project aims to build models that can accurately classify students into these grade categories. This predictive model can be beneficial for educators and administrators to identify at-risk students early on and implement targeted interventions to improve their academic performance and overall success.

TABLE OF CONTENTS

1. INTRODUCTION

In today's educational landscape, understanding and predicting student performance play a crucial role in fostering academic success and providing personalized support. With the advancement of machine learning techniques, it has become possible to develop predictive models that can anticipate student grades based on various input features.

This project focuses on predicting student grades using machine learning algorithms and features such as weekly study hours, attendance, notes, and reading habits. The grades are categorized into four classes: A, B, C, and Fail, representing different levels of academic achievement. By leveraging algorithms like K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest, this project aims to create accurate and reliable models to assist educators and institutions in identifying students who may need additional support or interventions to enhance their academic outcomes.

1.1. Features and Functionalities

1. Data Collection: Our project collects data related to student performance, including factors such as weekly study hours, attendance records, notes quality, reading habits, and possibly demographic information. This data forms the basis for building comprehensive profiles of students.
2. Feature Engineering: We extract and engineer features such as weekly study hours, attendance percentages, quality of notes taken, and reading habits from the collected data. These features are crucial for predicting student grades accurately.
3. Machine Learning Models: Our project employs various machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest, to predict student grades based on the engineered features. These models help in categorizing students into grade classes such as A, B, C, and Fail.
4. Evaluation Metrics: We use evaluation metrics such as Accuracy, Confusion Matrix, Precision, Recall, F1-Score, and possibly ROC curves to assess the performance of each predictive model. These metrics provide insights into the model's effectiveness in predicting student grades across different classes.
5. Scalability: Our project is designed to handle varying amounts of student data and can accommodate future expansions such as incorporating additional features or exploring new machine learning models for enhanced prediction accuracy. This scalability ensures that the predictive capabilities of our system can adapt to changing educational requirements and student behaviors.

2. LITERATURE SURVEY

Utilizing data from 8 years of student intakes spanning from July 2006/2007 to July 2013/2014, this study applied Decision Tree, Naïve Bayes, and Rule Based classification techniques to predict academic performance based on demographics, previous records, and family background. The Rule Based approach yielded the highest accuracy at 71.3%, offering valuable insights into identifying and profiling students' success in their first semester.[1]

The research methodology focused on predicting academic performance through classification methods like decision trees, random forests, and support vector machines, utilizing attendance, assignments, quizzes, and exam scores as features. Leveraging a dataset comprising these academic indicators, the study underscored the importance of employing classification techniques to improve student outcomes within educational contexts.[2]

The paper presents a methodology addressing issues like class imbalance and data hi-dimensionality in student academic performance prediction, employing phases including data preprocessing and ensemble methods. Utilizing a dataset from the UCI Machine Learning Repository encompassing student grades, demographic, social, and school-related factors, the proposed ensemble model, integrating the Decision Tree (J48) classifier, achieved a notable accuracy rate of 95.78%, showcasing its efficacy in enhancing predictive capabilities.[3]

The study at Dilla University utilized educational data mining techniques on a dataset spanning 5 years of undergraduate student data from the Department of Horticulture, employing RapidMiner software for preprocessing and analysis. Integrating university placement and course grade data, the research highlighted the effectiveness of data mining in predicting and enhancing student academic performance, emphasizing the significance of course-specific factors and academic status in determining success.[4]

The research employed machine learning algorithms like Naïve Bayes, Neural Network, Support Vector Machine, and Decision Tree classifier to analyze students' academic performance using transcript data from a university. While Naïve Bayes demonstrated the highest reliability for class predictions, the study suggests the need for additional dataset exploration and information about students' performance to enhance prediction accuracy.[5]

3. ANALYSIS AND REQUIREMENT SPECIFICATION

In our data analysis phase for predicting student grades, we will explore a variety of demographic and academic factors such as age, gender, weekly study hours, attendance records, notes quality, and reading habits to understand their impact on academic performance. Using statistical methods, we will uncover correlations and relationships between these features and student grades. Machine learning algorithms including Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machines, K-Nearest Neighbors, and Random Forest will be employed to develop predictive models. These models will help us identify the most significant factors influencing student grades and predict which students are likely to achieve higher grades based on their characteristics and academic behaviors. This analysis and modeling approach will provide valuable insights into student success factors and enable targeted interventions to improve overall academic outcomes.

3.1. Purpose

The purpose of this analysis is to develop a predictive model that can forecast whether an individual will adopt a newly launched game. By analyzing various demographic and behavioral factors, such as age, gender, gaming interests, social media engagement, and gaming habits, the aim is to identify patterns that can indicate the likelihood of game adoption. This predictive model can be valuable for game developers and marketers to target their audience effectively andoptimize their marketing strategies.

3.2. Scope

The scope of our project involves collecting and analyzing data on student performance, including factors like study hours, attendance, notes quality, and reading habits. We'll engineer features from this data to improve predictive accuracy and develop machine learning models using algorithms such as KNN, SVM, Logistic Regression, Decision Tree, and Random Forest. These models will be evaluated using various metrics to ensure reliability. Our project aims to provide insights for educators to enhance student success strategies, with a focus on scalability for future expansion and adaptability to changing educational needs and behaviors.

3.3. Functional Requirements

- Data collection from Students.
- Data preprocessing to handle missing values, encode categorical variables, and scale features.

- Feature selection to identify the most relevant grade the student belongs to.
- Model training using machine learning algorithms such as Decision Trees, SVM,Logistic Regression, KNN and Random Forest.
- Model evaluation using performance metrics such as accuracy, precision, recall, and ROCAUC.
- Deployment of the predictive model for real-time predictions.

3.4 Non-Functional Requirements

3.4.1 Hardware Requirements:

Processor	Intel Pentium or above
RAM	2GB above
Hard disk	10GB or above

3.4.2 Software Requirements:

Operating System	Windows
Language	Python
Integrated Development Environment (IDE)	Jupyter Notebook or Anaconda for code development and experimentation
Libraries	Pandas, NumPy, scikit-learn for data manipulation and modeling

4. DESIGN

4.1. FlowChart

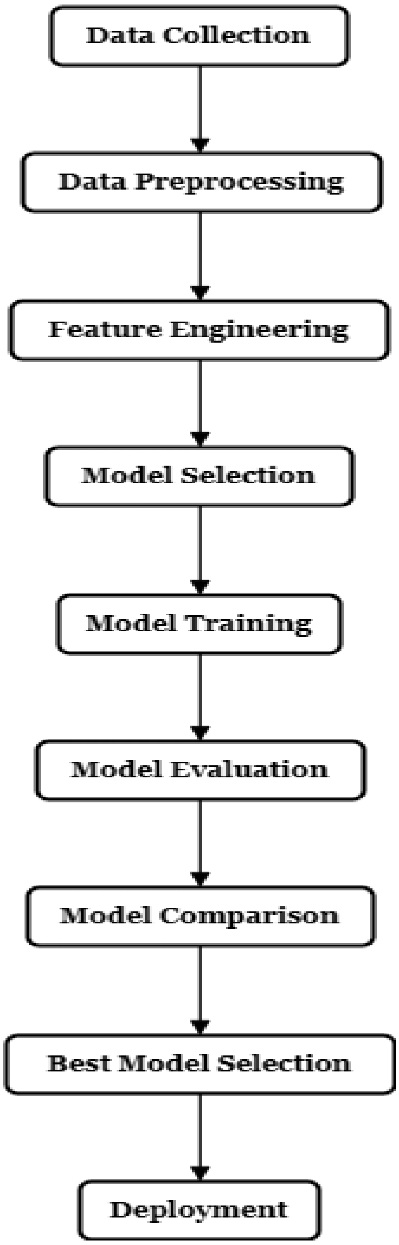


Fig. 1. FlowChart

5. IMPLEMENTATION

5.1. Data Collection

The project commenced with manual data collection from various sources, including hostelites, friends, cousins, and family members. Initially, the dataset comprised information related to student performance prediction, such as academic records, study habits, extracurricular activities, socio-economic background, and personal interests. This diverse dataset formed the basis for subsequent preprocessing and analysis phases in the machine learning project.

5.2. Data Preprocessing

Our project commenced with the collection of a data set comprising 140 instances. To enhance the robustness of our analysis, we did SMOTE analysis. Subsequently, we embarked on the crucial step of data preprocessing. This involved addressing missing values and encoding categorical variables using techniques such as LabelEncoder. Furthermore, to ensure the integrity of our model evaluation, we partitioned the dataset into a 80% training set and a 20% test set, facilitating robust training and evaluation of our machine learning models.

5.3. Feature Engineering

We extract relevant features from the dataset such as weekly study hours, attendance, notes, reading habits. These features are then used to train our models.

5.4. Model Training

In our project aimed at predicting student academic performance, we employed various machine learning techniques to gain insights from our data. We initiated our analysis with Random Forest, which allowed us to discern the impact of factors such as study hours, note-taking habits, and reading frequency on students' academic achievements. Subsequently, Support Vector Machine (SVM) was utilized to identify intricate patterns within the data, shedding light on how different attributes influence students' grades.

Following SVM, we delved into K-Nearest Neighbors (KNN), which operates by identifying similarities among students in our dataset to make predictions. By focusing on individuals with comparable study habits and academic behaviors, KNN revealed hidden trends that influence academic performance, offering valuable insights into why certain students may excel academically compared to others.

Additionally, Decision Trees were employed to partition our data into smaller subsets based on specific characteristics. This segmentation facilitated a clearer understanding of which factors hold the most significance in determining students' academic outcomes. Decision Trees provided a comprehensive overview of how variables such as age and study interests contribute to academic success.

By utilizing an ensemble of decision trees, Random Forest allowed us to analyze the importance of different features such as study hours, note-taking habits, and reading frequency in predicting academic outcomes

In our project, Logistic Regression excelled in predicting student grades by effectively categorizing them into the defined classes of A, B, C, or fail. Its simplicity and interpretability allowed us to comprehend the relationship between input variables such as study hours, attendance, and Notes taking habits with the likelihood of achieving different grade outcomes.

5.5. Model Evaluation

The trained models are evaluated using metrics such as accuracy, confusion matrix, sensitivity, specificity, and ROC AUC. This step helps us assess the performance of each model and select the best-performing one.

5.6. Predictive Analysis

Once the models are trained and evaluated, they are used to make predictions on new data. This allows us to predict whether the Students are likely to belong to the grades A,B,C or Fail.

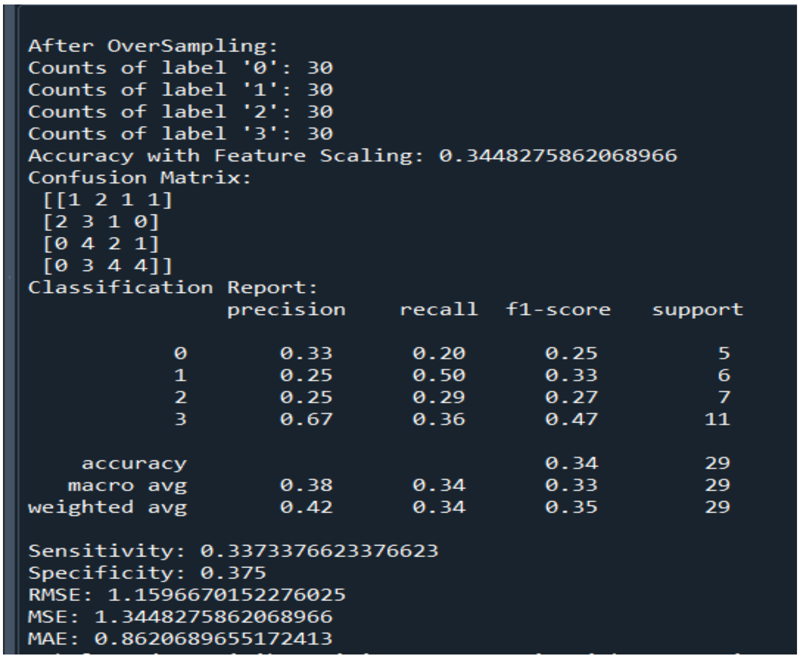
5.7. Performance Comparison

Finally, we compare the performance of each model to determine which one provides the most accurate predictions.

6. RESULT

6.1. Sample Output (Snapshots)

The fig.3. with a score of 34.4%, SVM demonstrates predictive capabilities for student grades. The accompanying classification reports and confusion matrices provide a thorough assessment of its performance.



```
After OverSampling:
Counts of label '0': 30
Counts of label '1': 30
Counts of label '2': 30
Counts of label '3': 30
Accuracy with Feature Scaling: 0.3448275862068966
Confusion Matrix:
 [[1 2 1 1]
 [2 3 1 0]
 [0 4 2 1]
 [0 3 4 4]]
Classification Report:
              precision    recall  f1-score   support

           0       0.33      0.20      0.25         5
           1       0.25      0.50      0.33         6
           2       0.25      0.29      0.27         7
           3       0.67      0.36      0.47        11

    accuracy                           0.34        29
   macro avg       0.38      0.34      0.33        29
weighted avg       0.42      0.34      0.35        29

Sensitivity: 0.3373376623376623
Specificity: 0.375
RMSE: 1.1596670152276025
MSE: 1.3448275862068966
MAE: 0.8620689655172413
```

Fig. 3. SVM Performance

The fig.4. demonstrates SVM's accurate classification across various thresholds, addressing class imbalances effectively.

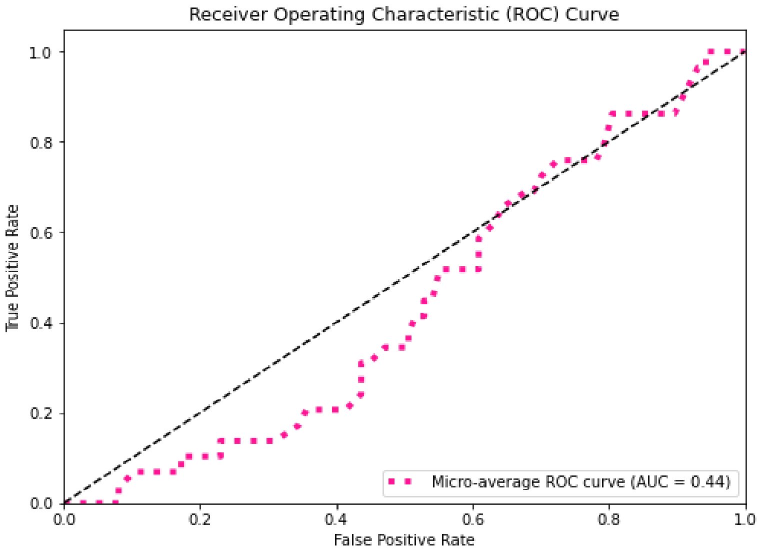
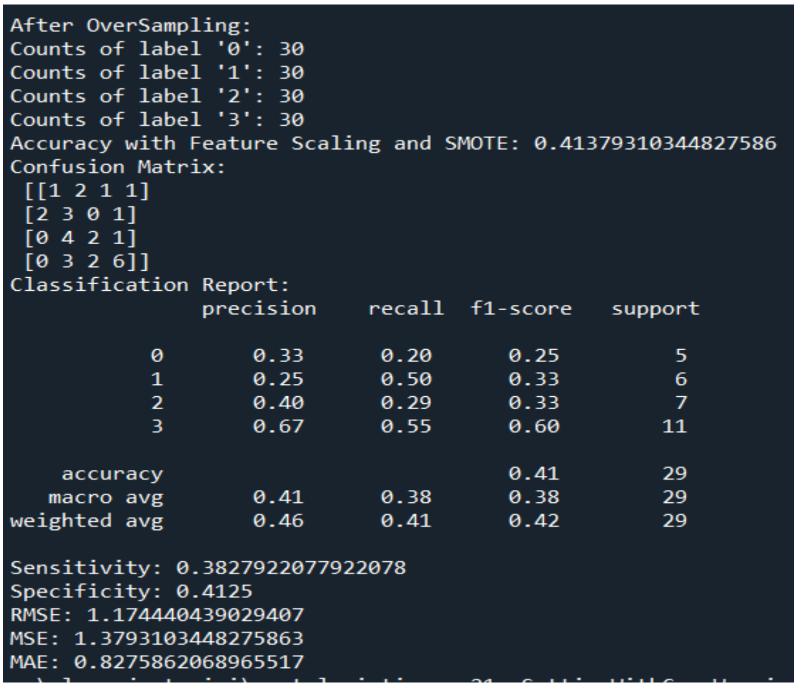


Fig. 4. ROC curve for SVM

The fig.5. displays an accuracy of 41.37%, showcasing Logistic Regression's predictive performance in game adoption. Detailed metrics like classification reports and confusion matrices offer insights into its effectiveness.



```
After OverSampling:
Counts of label '0': 30
Counts of label '1': 30
Counts of label '2': 30
Counts of label '3': 30
Accuracy with Feature Scaling and SMOTE: 0.41379310344827586
Confusion Matrix:
 [[1 2 1 1]
 [2 3 0 1]
 [0 4 2 1]
 [0 3 2 6]]
Classification Report:
               precision    recall  f1-score   support

           0       0.33      0.20      0.25         5
           1       0.25      0.50      0.33         6
           2       0.40      0.29      0.33         7
           3       0.67      0.55      0.60        11

    accuracy                           0.41        29
   macro avg       0.41      0.38      0.38        29
weighted avg       0.46      0.41      0.42        29

Sensitivity: 0.3827922077922078
Specificity: 0.4125
RMSE: 1.174440439029407
MSE: 1.3793103448275863
MAE: 0.8275862068965517
```

Fig. 5. Logistic Regression Performance

The fig.6. illustrates discrimination ability, balancing sensitivity, and specificity for game adoption prediction.

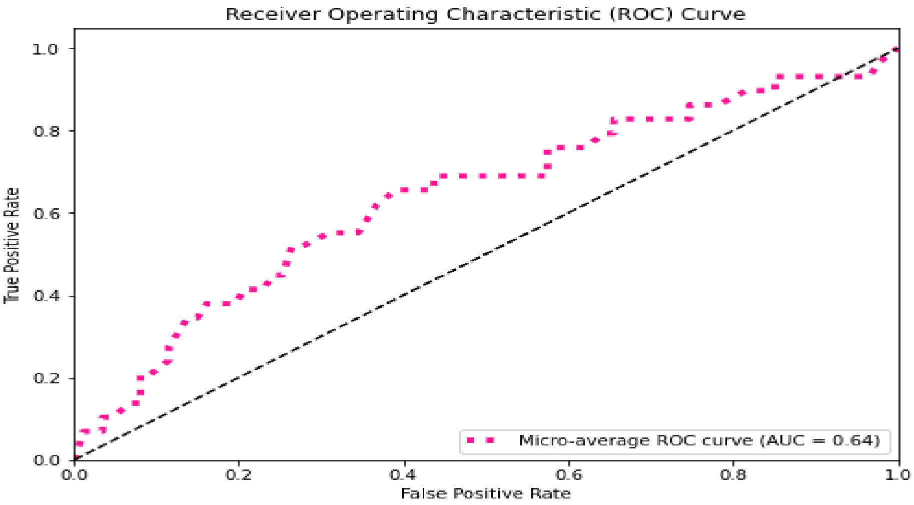
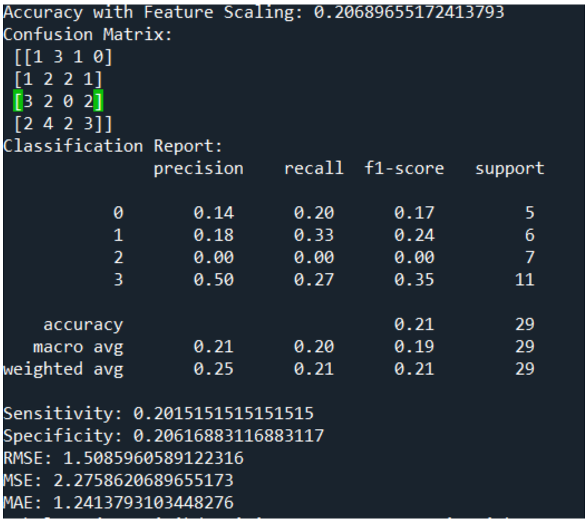


Fig. 6. ROC curve for Logistic Regression

The fig.7. displays an accuracy of 20.06%, Decision Tree exhibits robust performance in game adoption prediction. Comprehensive classification reports and confusion matrices offer valuable insights.



```
Accuracy with Feature Scaling: 0.20689655172413793
Confusion Matrix:
 [[1 3 1 0]
 [1 2 2 1]
 [3 2 0 2]
 [2 4 2 3]]
Classification Report:
               precision    recall  f1-score   support

           0       0.14      0.20      0.17         5
           1       0.18      0.33      0.24         6
           2       0.00      0.00      0.00         7
           3       0.50      0.27      0.35        11

    accuracy                           0.21        29
   macro avg       0.21      0.20      0.19        29
weighted avg       0.25      0.21      0.21        29

Sensitivity: 0.2015151515151515
Specificity: 0.20616883116883117
RMSE: 1.5085960589122316
MSE: 2.2758620689655173
MAE: 1.2413793103448276
```

Fig. 7. Decision Tree Performance

The fig.8. showcases Decision Tree's accuracy in diverse data distributions for game adoption forecasts.

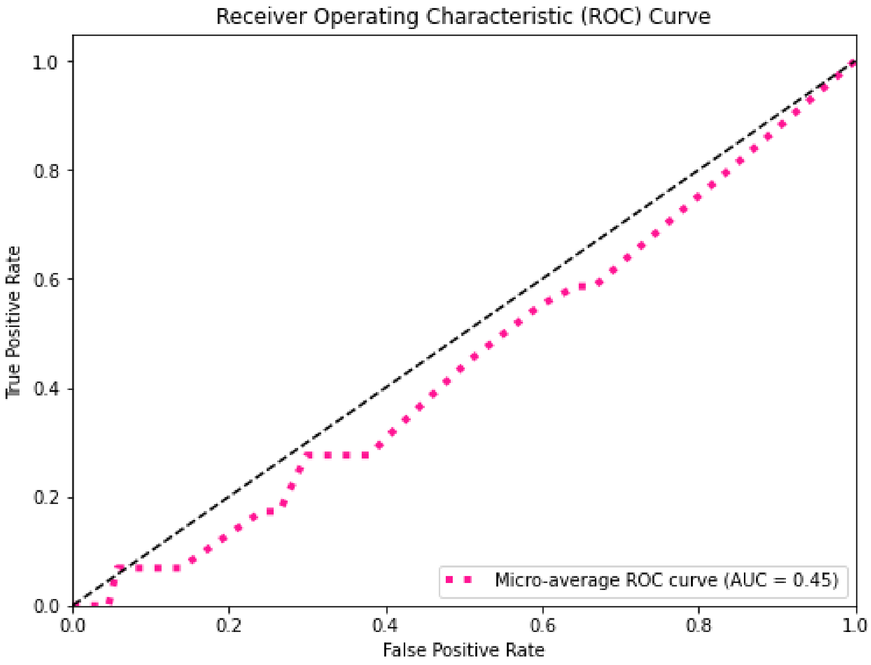
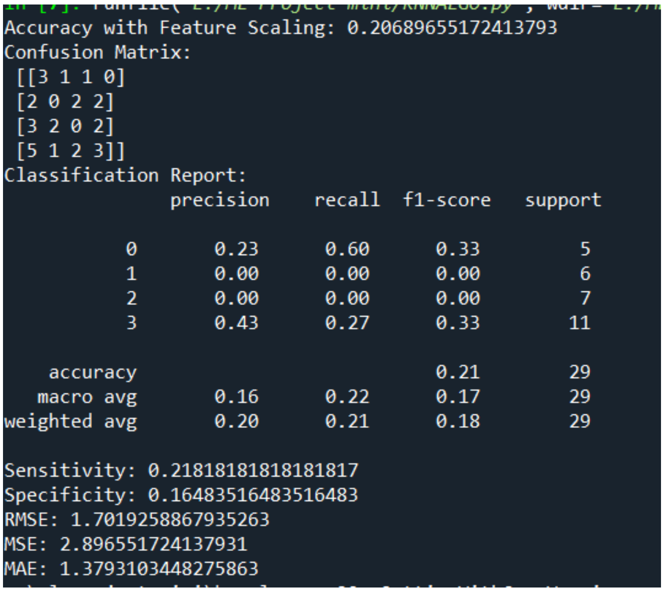


Fig. 8. ROC curve for DT

KNN achieves an impressive accuracy of 20.68% as shown in fig.9., highlighting its proficiency in predicting game adoption. Detailed classification reports and confusion matrices enhance understanding.



```
Accuracy with Feature Scaling: 0.20689655172413793
Confusion Matrix:
 [[3 1 1 0]
 [2 0 2 2]
 [3 2 0 2]
 [5 1 2 3]]
Classification Report:
               precision    recall  f1-score   support

           0       0.23      0.60      0.33         5
           1       0.00      0.00      0.00         6
           2       0.00      0.00      0.00         7
           3       0.43      0.27      0.33        11

    accuracy                           0.21        29
   macro avg       0.16      0.22      0.17        29
weighted avg       0.20      0.21      0.18        29

Sensitivity: 0.21818181818181817
Specificity: 0.16483516483516483
RMSE: 1.7019258867935263
MSE: 2.896551724137931
MAE: 1.3793103448275863
```

Fig. 9. K-NN Performance

The fig.10. highlights KNN's strong discrimination, leveraging proximity for precise classification in grade prediction.

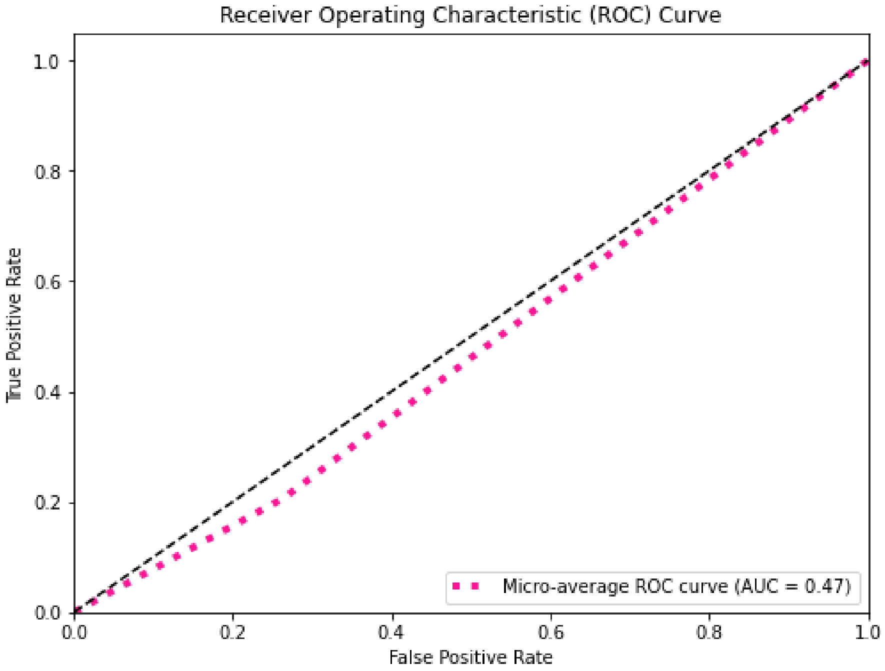
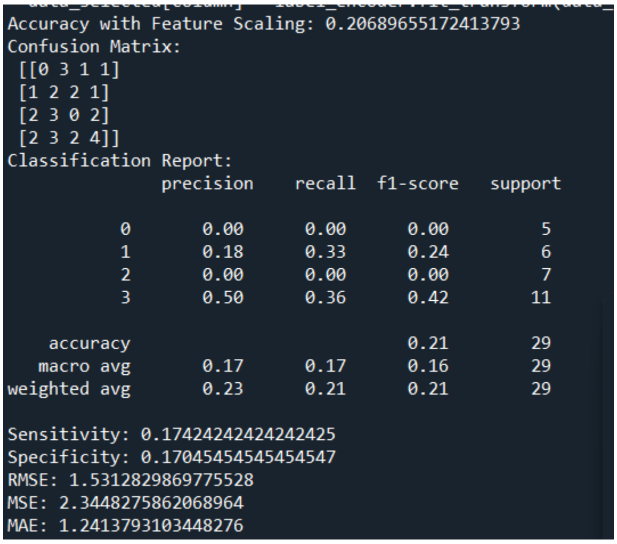


Fig. 10. ROC curve for K-NN

Random Forest achieves outstanding accuracy at 20.68% as shown in fig.13., on par with KNN and Decision Tree. Comprehensive classification reports and confusion matrices underline its effectiveness.



```
Accuracy with Feature Scaling: 0.20689655172413793
Confusion Matrix:
 [[0 3 1 1]
 [1 2 2 1]
 [2 3 0 2]
 [2 3 2 4]]
Classification Report:
               precision    recall  f1-score   support

           0       0.00      0.00      0.00         5
           1       0.18      0.33      0.24         6
           2       0.00      0.00      0.00         7
           3       0.50      0.36      0.42        11

    accuracy                           0.21        29
   macro avg       0.17      0.17      0.16        29
weighted avg       0.23      0.21      0.21        29

Sensitivity: 0.17424242424242425
Specificity: 0.17045454545454547
RMSE: 1.5312829869775528
MSE: 2.3448275862068964
MAE: 1.2413793103448276
```

Fig. 13. Random Forest Performance

The fig.14. highlights Random Forest's robustness to noise and complex patterns in game adoption prediction.

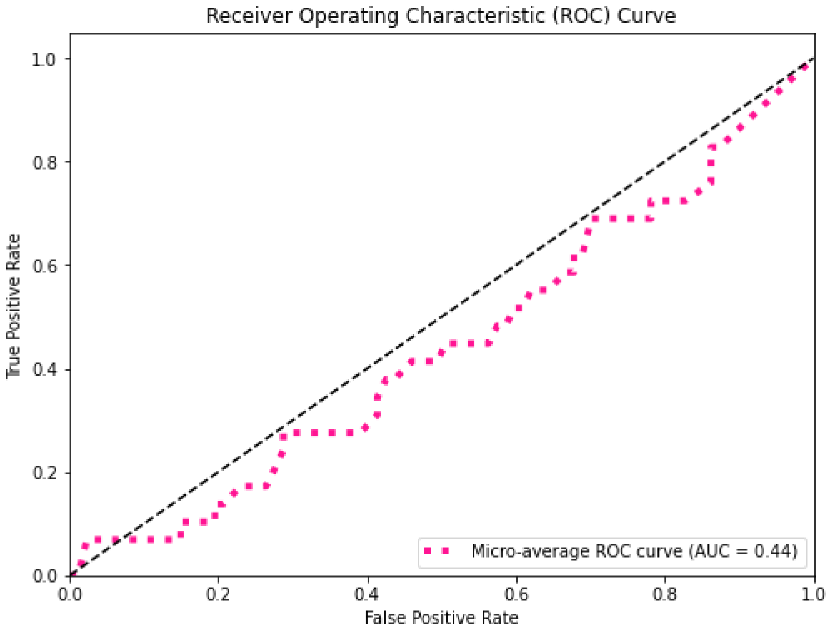


Fig. 14. ROC curve for Random Forest

Summarizing the results that we have obtained, the analysis of various machine learning algorithms yielded insightful findings regarding their effectiveness in predicting grades.
SVM achieved an accuracy score of approximately 34.48%.
KNN demonstrated predictive capabilities with an accuracy score of 20.68%. Random Forest exhibited robust performance with an accuracy score matching that of KNN, standing at
20.68%. Decision Tree achieved a accuracy score of 20.06%. Logistic Regression emerged as a highly accurate model, achieving an impressive accuracy score of 41.37%. The evaluation metrics, including confusion matrices and ROC curve analyses, provided comprehensive insights into the predictive performance of these models. These results offer valuable guidance for teachers in the field of education to make informed decisions regarding guidance and counselling for students with less grades.

7. CONCLUSION

In conclusion, our student performance prediction project leverages machine learning algorithms and data analysis techniques to accurately forecast student grades based on factors like study hours, attendance, and notes quality. By collecting and analyzing relevant data, we engineer features and develop predictive models using algorithms such as KNN, SVM, Logistic Regression, Decision Tree, and Random Forest. The project's scope includes creating a scalable system adaptable to changing educational needs, providing valuable insights to educators, and identifying at-risk students for targeted interventions, thereby enhancing student success strategies and improving academic outcomes. This proactive approach fosters a positive learning environment, contributes to student retention, and emphasizes ethical considerations in predictive analytics, ensuring responsible use of student data for fair and transparent decision-making. Overall, our project signifies a significant step forward in leveraging data-driven insights to optimize student learning experiences and outcomes.

7.1. Future Scope of the Project

- Integration with other educational data systems for comprehensive student profiling and analysis
- Implementing real-time prediction capabilities to provide timely interventions for struggling students
- Incorporating natural language processing (NLP) techniques to analyze qualitative data such as notes quality and feedback
- Exploring ensemble learning techniques to improve model performance and prediction accuracy Developing a user-friendly dashboard or interface for educators to access and interpret prediction results easily
- Conducting longitudinal studies to assess the long-term impact of predictive analytics on student outcomes
- Collaborating with educational researchers and policymakers to integrate insights into curriculum development and policy decisions
- Expanding the scope to include predictive analytics for student engagement, retention, and career readiness beyond academic grades.

REFERENCES

[1] Ahmad, Fadhilah, Nur Hafieza Ismail, and Azwa Abdul Aziz (2015) The Prediction of Students’ Academic Performance Using Classification Data Mining Techniques.

[2] Hashmia Hamsa, Simi Indira Devi, Jubilant J Khizhakkethottam (2016) Student academic performance prediction model using decision tree and fuzzy genetic algorithm.

[3] Imran, Muhammad, Shahzad Latif, Danish Mehmood, and Muhammad Saqlain Shah (2019) Student Academic Performance Prediction using Supervised Learning Techniques.

[4] Berhanu, Fiseha (MSC), and Addisalem Abera (2015) Students’ Performance Prediction based on their Academic Record.

[5] Sudais, Muhammad , Muhammad Safwan, Maryam Aisha Khalid, and Shaheer Ahmed (2019) Students’ Academic Performance Prediction Model Using Machine Learning.

[1]