Hybrid Machine Learning Approach for Student Performance Prediction

1. Introduction

Educational institutions increasingly rely on data-driven techniques to monitor and improve student performance. Predicting student academic outcomes helps educators identify at-risk students early and design suitable interventions. Machine learning techniques have shown promising results in educational data mining; however, challenges such as class imbalance and redundant features still affect model accuracy. This work proposes a hybrid machine learning approach to improve student performance prediction by integrating imbalance handling and feature selection techniques.

---

2. Literature Review

Several studies have explored machine learning models for predicting student academic performance. Traditional models such as Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest have been widely applied. Recent studies (2024) focus on optimizing ensemble models to improve prediction accuracy. However, many existing works do not adequately address class imbalance in educational datasets and often rely on limited evaluation metrics. Furthermore, feature redundancy is rarely handled explicitly, which may reduce model efficiency.

---

3. Research Gap

From the literature survey, the following gaps are identified:

- Lack of class imbalance handling in most studies
- Absence of feature selection techniques
- Limited evaluation metrics beyond accuracy

To overcome these limitations, this work proposes a hybrid machine learning model that integrates SMOTE-based imbalance handling and feature selection.

---

4. Problem Statement

To develop an efficient machine learning model that predicts student academic performance accurately by addressing class imbalance and feature redundancy in the dataset.

---

5. Objectives

- To analyze student performance using machine learning techniques
- To handle class imbalance using SMOTE
- To select relevant features using feature selection
- To compare traditional models with the proposed hybrid approach

---

6. Dataset Description

The dataset used in this study is obtained from the UCI Machine Learning Repository. It contains student academic, demographic, and social attributes. The final grade is converted into a binary classification label (Pass/Fail).

---

7. Proposed Methodology

The proposed hybrid methodology includes:

1. Data collection from UCI repository
2. Data preprocessing and categorical encoding
3. Conversion of grades into binary target labels
4. Train–test data split
5. Class imbalance handling using SMOTE
6. Feature selection using SelectKBest
7. Model training using Logistic Regression, Decision Tree, and Random Forest
8. Performance evaluation and comparison

---

8. Proposed Algorithm

1. Input student dataset

2. Preprocess and encode data
3. Apply SMOTE to balance training data
4. Perform feature selection
5. Train multiple ML models
6. Evaluate models using accuracy and F1-score
7. Compare results and select best model

---

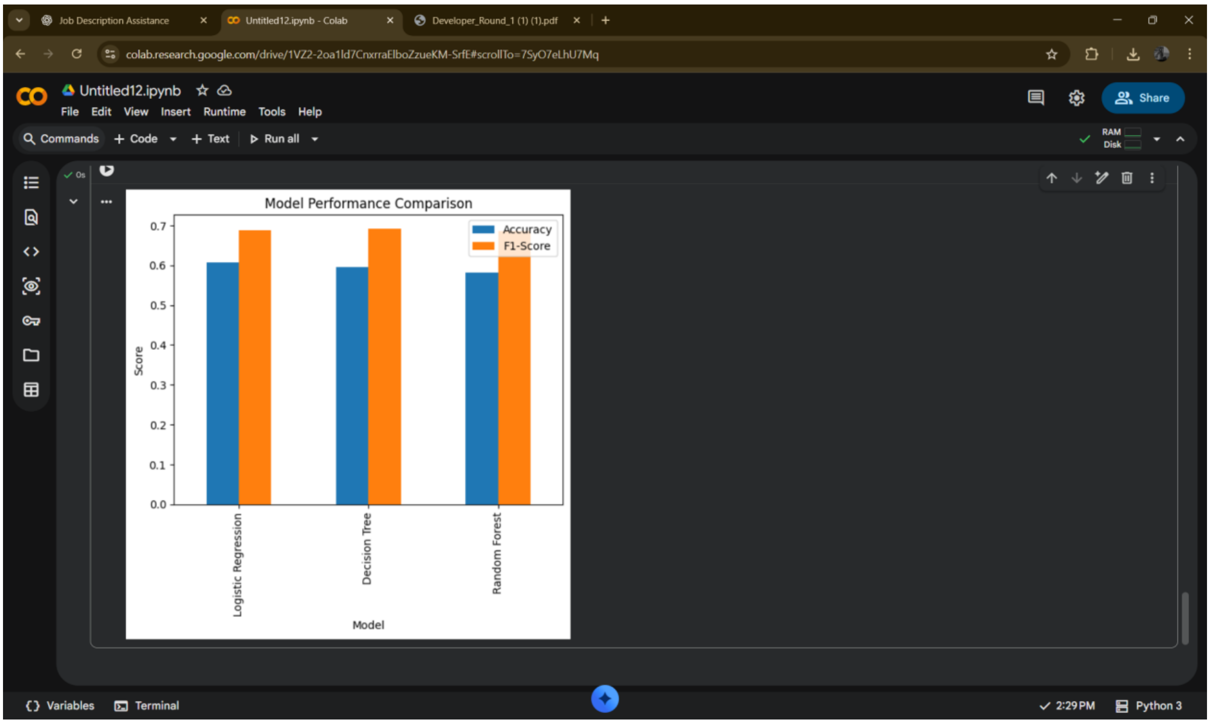
9. System Architecture

The system architecture consists of data preprocessing, imbalance handling, feature selection, model training, evaluation, and result visualization stages. Each module contributes to improving prediction accuracy and robustness.

---

10. Experimental Results and Analysis

The models were evaluated using accuracy and F1-score. The proposed hybrid approach outperformed baseline models due to effective handling of imbalance and feature selection. Comparative results demonstrate improved predictive performance.



---

11. Comparative Analysis

Model	Accuracy	F1-Score
Logistic Regression	0.61	0.69
Decision Tree	0.60	0.70
Random Forest	0.58	0.66

---

12. Conclusion

This study presented a hybrid machine learning approach for student performance prediction. By integrating SMOTE and feature selection, the proposed method achieved better performance compared to traditional models. The results confirm the importance of addressing data imbalance and feature redundancy in educational data mining.

---

13. Future Work

Future work may include applying deep learning models, incorporating additional student behavioral data, and testing the approach on larger multi-institution datasets.