Navigating Educational Paths: Optimizing Student Prediction Models via Feature Selection Analysis

Overview

- Introduction
- Motivation
- Project Goals
- Literature Review
- Dataset Description
- Methodology
- Results
- Future Plans
- Our Team

INTRODUCTION

1.Educational Data Mining (EDM) utilizes data mining techniques to analyze educational data, aiming to enhance educational practices.

2. It involves collecting data from learning platforms to gain insights and improve capabilities. This includes tracking student progress and monitoring course performance.

3. Predicting and understanding student performance is crucial in today's education landscape, empowering educators to identify learning patterns and help students to improve.

4. This study focuses on using machine learning for predicting academic performance, filling a vital gap in educational research.

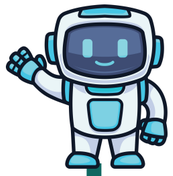
Motivation

In the Indian education system, due to financial constraints, many students may need to drop out of school or college prematurely. In many cases, students are forced to drop out of school due to various factors such as family obligations, or lack of academic support.

This predictive system can help in identifying the underlying reasons behind students' academic struggles, whether it be learning disabilities, socio-economic challenges, or lack of access to resources.

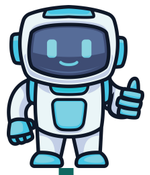
This insight enables targeted interventions and guidance to help students overcome obstacles and achieve better outcomes, ultimately fostering their educational advancement and socioeconomic mobility.

Project Goals



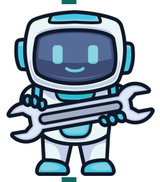
Firstly

Examining student performance data to discern trends and forecast future academic outcomes.



Secondly

Investigating regression models for forecasting student performance using background factors.



Thirdly

Conducting correlation analysis to uncover the relationships between performance across various subjects.

Literature Review

Related research works include:

01

Gupta, S., Kishan, B., & Gulia, P. (2024). Comparative analysis of predictive algorithms for performance measurement

Regressions used:

1. Logistic Regression
2. Linear Regression(LR)
3. Polynomials Regression
4. Random Forest Regression(RFR)
5. Decision Tree Regression
6. LASSO Regression

02

Mustapha, S. M. F. D. S. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods.

Regressions used:

1. Linear Regression (LR)
2. Support Vector Regressor (SVR)
3. Random Forest Regressor (RFR)
4. Gradient Boosted Regressor (GBR)

03

Pratama, D., & Husnayaini, I. (2022). Program for international student assessment (pisa) analysis of asian countries using k-mean clustering algorithms

Regression used:

Cluster analysis-mean algorithm (k-mean algorithm)

How is our analysis better than the previous ones ?

- Our prediction system stands out, making it preferable in certain scenarios because we use a method called **feature selection**, where we pick out the top 15 most important features to make our predictions better.
- This helps increase the accuracy of our predictions by about **2-3%**. It also makes our system faster because it has fewer things to consider.
- By focusing on the most important data, we make sure our predictions are based on the stuff that matters the most.
- This not only proves that our feature selection method works well but also shows how useful it is for making our predictions more accurate and reliable. The increase in accuracy after selecting features confirms that picking out the most useful ones helps our system work better and understand the data more effectively.

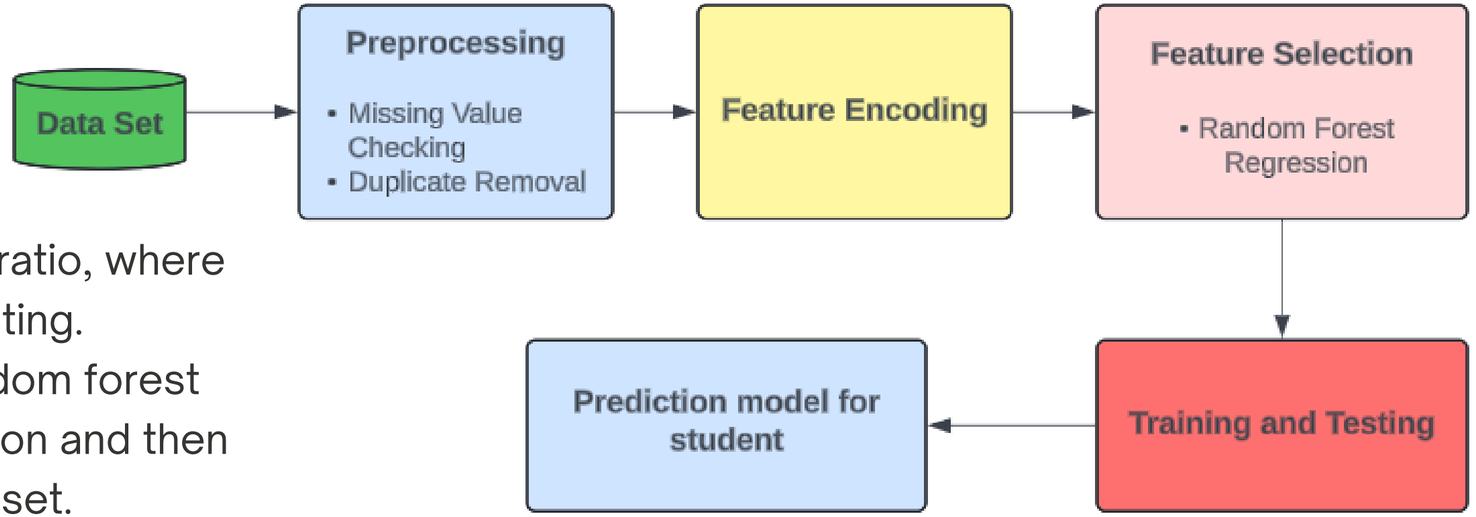
Dataset Description

In this study, we examined recent real-world data from two Portuguese secondary schools. The data came from two sources: mark reports and questionnaires. Given the limited information in the mark reports, which included only grades and number of absences, the data was supplemented with information from questionnaires. This provided various demographic, social, and school-related attributes (e.g. student’s age,father’s job and mother’s education)

Methodology

Here, we are using a **Random Forest Regressor** for feature selection.
Here's how it works:

- The data set was split in **80% - 20%** ratio, where 80% is for training and 20% is for testing.
- On the training data, we applied random forest classifier to perform feature extraction and then visualize the top features in the dataset.
- Then, we have evaluated the model using mean absolute error, mean square error and determined the r2 score from it.
- After that, we have applied nine regression models on the training dataset and evaluated that which of them gives us the best prediction accuracy.



Advantages

- High Predictive Accuracy,
- Handles Large Datasets,
- Robustness to Outliers and Missing Values

Results

Comparison of results with and without feature selection

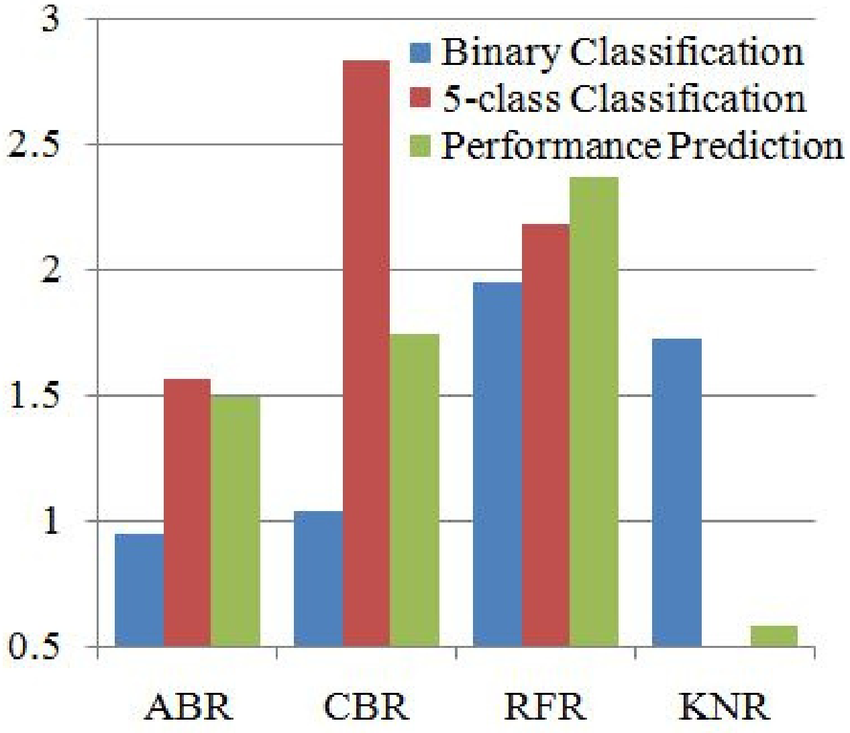
Input Setup	Mathematics				Portuguese			
	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Binary Classification	67.2	65.1	60.92	60.01	41.68	44.02	45.59	50.27
5-class Classification	62.14	64.7	62.85	53.21	38.00	45.34	48.62	43.31
Performance Prediction	80.44	81.06	80.92	77.86	71.73	80.11	83.62	77.33

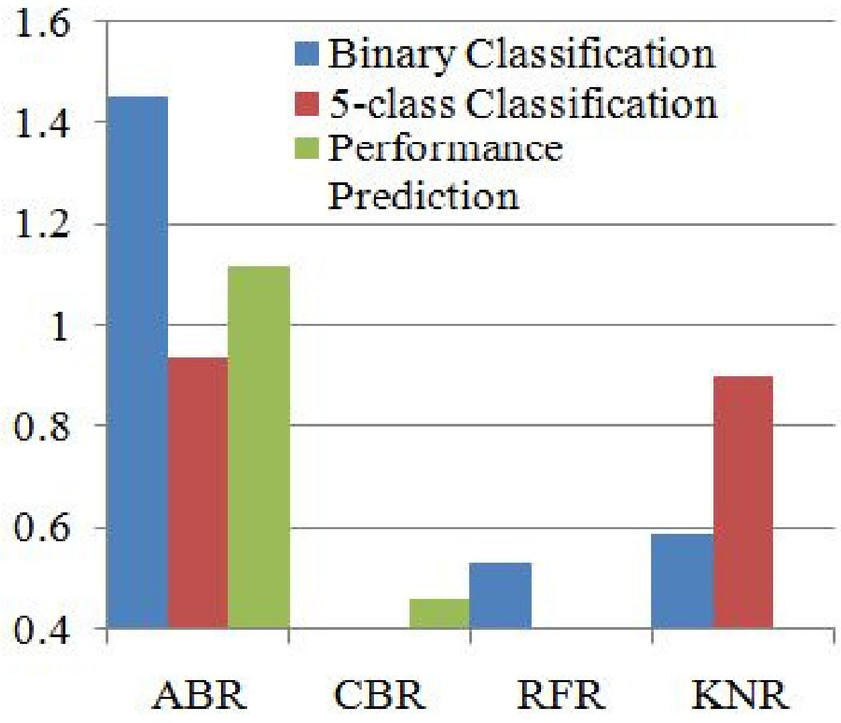
Performance comparison of different setups without utilizing features selection, with accuracy values expressed in percent (%)

Input Setup	Mathematics				Portuguese			
	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Binary Classification	68.15	66.14	62.87	61.74	43.13	43.86	46.12	50.86
5-class Classification	63.70	67.54	65.04	53.22	38.94	45.25	44.67	44.21
Performance Prediction	81.93	82.81	83.29	78.45	72.85	80.57	79.91	75.35

Performance comparison of different setups utilizing top 15 features, with accuracy values expressed in percent (%)

Results with feature selection





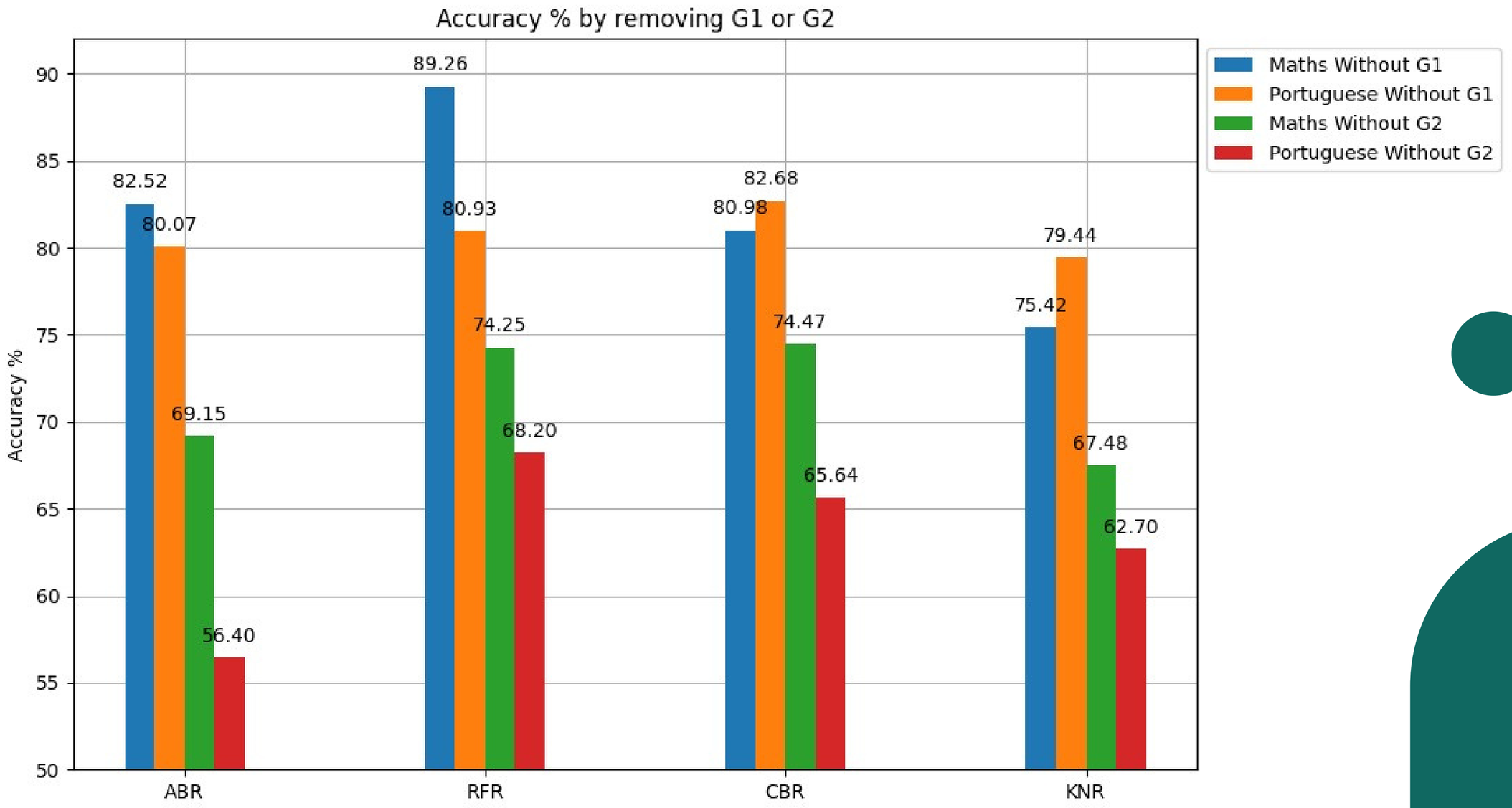
Accuracy increased after feature selection in case of (a) Mathematics and (b) Portuguese datasets.

Comparison of results without some key features

Input Setup	Mathematics				Portuguese			
	ABR	CBR	RFR	KNR	ABR	CBR	RFR	KNR
Without G1 only	78.52	79.26	80.98	75.42	70.07	77.93	77.68	74.44
Without G2 only	69.15	74.25	74.47	67.48	56.40	68.20	65.64	62.70

Performance comparison of different setups with and without some features. Accuracy values expressed in percent (%)

Result in % after removing G1 and G2



Future Plans

We're seeking to enhance our student prediction system by employing additional feature selection techniques categorized either as filters or wrappers.

We're making efforts to improve the accuracy and reliability of our student prediction system.

We're additionally focused on expanding our dataset with diverse information and integrating advanced technology to create a more user-friendly interface.

Our Team

Team Tech Strivers
M. Sc. Computer Science, Sem II
Asutosh College, Kolkata

Mousumi Banerjee
MSc 1st year

Priyanka Dhar
MSc 1st year

Piyali Naha
MSc 1st year

Supriyo Saha
MSc 1st year

THANK YOU !

For watching this presentation