Student Performance Analysis Using Data Science

Diksha Gupta & Mayank Anand
School of Computer Science, UPES

I. Abstract

Student academic performance is influenced by various demographic, socio-economic, and behavioural factors. This study aims to analyse the impact of these factors on student grades using data from the UCI Student Performance Dataset, which includes information on students taking Math and Portuguese courses. By leveraging exploratory data analysis (EDA) and machine learning models, we seek to identify key predictors of student success and develop a predictive model to forecast academic outcomes. The study examines attributes such as family background, parental education, study habits, and extracurricular activities to determine their correlation with student performance. The results will provide actionable insights for educators, parents, and policymakers to improve learning strategies and interventions, ultimately enhancing student outcomes.

II. Introduction

Education is a critical determinant of an individual's success, and student performance is often used as a key metric for evaluating academic progress. However, performance is not solely dependent on a student's intelligence or effort; a wide range of external factors such as parental education, socio-economic background, study habits, extracurricular activities, and psychological well-being play significant roles in shaping academic outcomes.

In recent years, data science and machine learning have provided new opportunities for understanding and predicting student performance based on various influencing factors. This study utilizes the UCI Student Performance Dataset, which contains data from students enrolled in Mathematics and Portuguese courses. The dataset includes demographic details, family background, academic records, and behavioral factors, offering a comprehensive view of student performance.

By conducting exploratory data analysis (EDA) and applying machine learning models, this project aims to:

1. **Identify key factors** that influence student grades.
2. **Analyze correlations** between different attributes and performance.
3. **Build predictive models** to forecast final grades.
4. **Provide recommendations** for students, educators, and policymakers to enhance academic achievement.

Understanding these relationships can help create targeted educational strategies, improve curriculum design, and develop personalized learning plans to support students at risk of poor performance.

III. Problem Statement

Traditional student assessment methods primarily focus on grades without considering the various external factors that contribute to academic success. Many students struggle academically due to social, economic, and psychological challenges, yet there is often no systematic way to predict or address these issues in advance. Some key challenges include:

- **Lack of insight into student challenges:** Teachers and administrators may not have access to data-driven insights that can highlight struggling students before they fail.
- **Influence of socio-economic factors:** Family background, parental education, and financial stability play crucial roles in academic performance, but these are often overlooked in traditional assessment models.
- **Effect of study habits and external activities:** While study hours and attendance are often linked to better performance, the impact of extracurricular activities, internet access, and family relationships is not well understood.
- **Need for predictive modeling:** There is a lack of effective models that can forecast student performance based on past academic records and behavioral attributes.

By leveraging the UCI dataset, this project seeks to bridge this gap by identifying the most critical predictors of student performance and building models that can help predict at-risk students, guide interventions, and enhance educational outcomes.

IV. Literature

A background study is done to review similar existing systems used to perform student performance analysis. Three existing systems are chosen because they are similar to the proposed system.

A. Faculty Support System (FSS)
Shana and Venkatacalam proposed a framework named Faculty Support System (FSS), which uses cost-effective open-source analysis software, WEKA, to analyze student performance at Coimbatore Institute of Technology, Anna University. FSS dynamically updates student data to create or add new rules using classification techniques in data mining.

B. Student Performance Analyser (SPA)
SPA is a secure online web-based tool used worldwide by educators to analyze student performance. It enables progress tracking, report generation, and assessment of individual and class performance to identify students performing below, at, or above the expected level.

C. Intelligent Mining and Decision Support System (InMinds)
InMinds is used at Universiti Malaysia Sarawak (UNIMAS) to monitor departmental performance. The system provides an intuitive dashboard with visual analytics for better decision-making in student performance evaluation.

From these reviews, useful techniques and features will be integrated into the proposed system to enhance its performance. WEKA is chosen as the primary data mining tool due to its open-source nature.

V. Objective

The primary objectives of this study are:

1. To analyze the impact of various demographic, socio-economic, and academic factors on student performance.
2. To identify key attributes that significantly influence student grades.
3. To develop predictive models that can accurately forecast student performance based on historical data.

VI. Methodology

The methodology for this project follows a structured data science pipeline:

1. **Data Collection:**
   - The dataset used is the UCI Student Performance Dataset, containing records of students enrolled in Mathematics and Portuguese courses.
   - The dataset includes demographic, academic, socio-economic, and behavioral attributes.

2. **Data Preprocessing:**
   - Handling missing values and outliers.
   - Encoding categorical variables for machine learning models.
   - Normalizing or standardizing numerical features.

3. **Exploratory Data Analysis (EDA):**
   - Identifying trends and patterns in the dataset using statistical summaries and visualizations (heatmaps, box plots, and histograms).
   - Understanding correlations between different features and their impact on student performance.

4. **Feature Selection and Engineering:**
   - Identifying key variables that contribute significantly to student grades.
   - Creating new features if necessary to enhance model performance.

5. **Machine Learning Model Development:**
   - Implementing various models such as: **Linear Regression** (for understanding relationships between attributes and grades). **Decision Trees & Random Forests** (for feature importance and classification tasks).
   - Splitting the dataset into training and testing sets.
   - Hyperparameter tuning for model optimization.

6. **Model Evaluation:**
   - Evaluating models using accuracy, Mean Squared Error (MSE), R^2 score, and confusion matrices.
   - Comparing different models to determine the best-performing one.

7. **Insights and Recommendations:**
   - Analyzing the most influential factors in student performance.
   - Suggesting personalized study strategies for students.
   - Providing recommendations for teachers and policymakers to improve educational programs.

8. **Deployment and Visualization:**
   - Developing an interactive dashboard using tool like shiny.

This structured approach ensures a comprehensive analysis, leading to meaningful insights and actionable recommendations for enhancing student performance.

VII. System Requirements

Hardware:
Processor: Intel i5 or higher (or AMD equivalent)
RAM: Minimum 8GB (16GB recommended for large datasets)
Storage: Minimum 50GB free space
GPU: Recommended for deep learning (NVIDIA GTX 1050 or better)
Internet: Required for data access and cloud integration

Software:
Operating System: Windows 10/11, macOS, or Linux
Programming Language: Python (latest stable version)
Libraries:
Data Processing: dplyr, caret
Machine Learning: randomForest
Visualization: ggplot2
Web Application: shiny

Database (Optional): MySQL/PostgreSQL for storing student data
Development Tools: RStudio, VS Code with R extensions.

VIII. Results

1. EDA observation

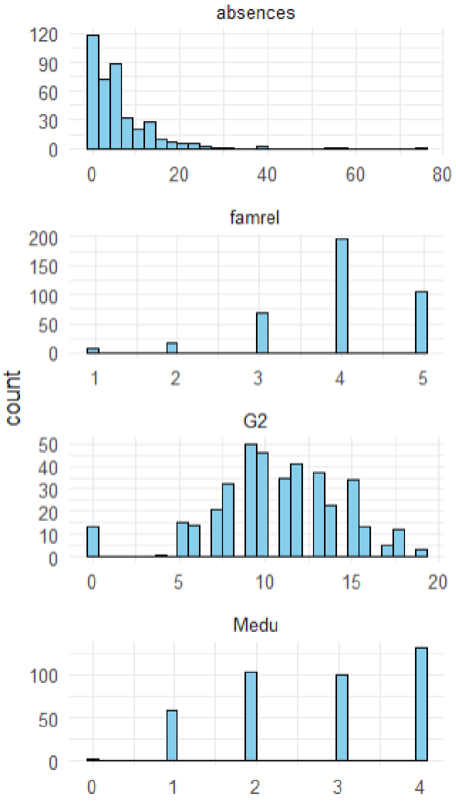


Fig. 1(i)- range of variables vs. frequency

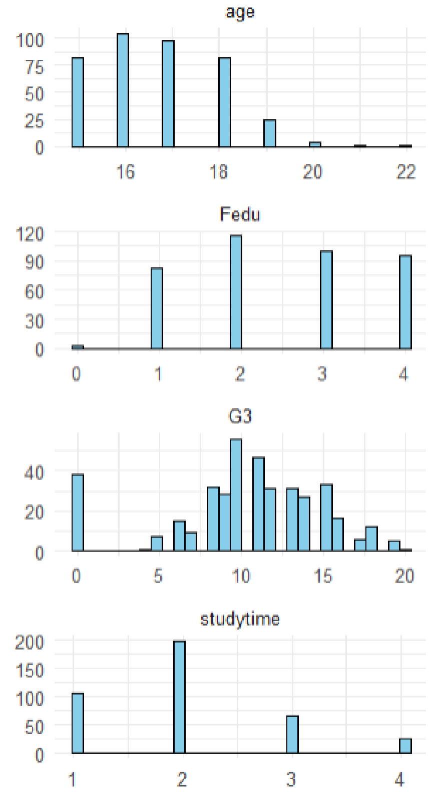


Fig. 1(ii)- range of variables vs. frequency

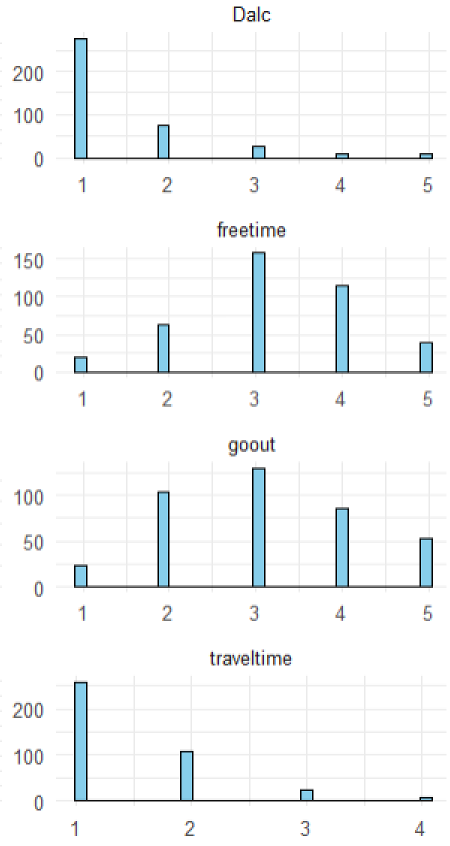


Fig. 1(iii)- range of variables vs. frequency

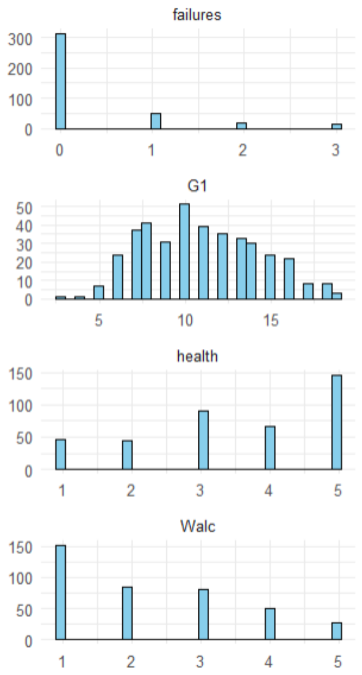


Fig. 1(iv) - Histograms to show *how many data points fall into each range of values [x-axis: value range of variables. Y-axis: frequency]*

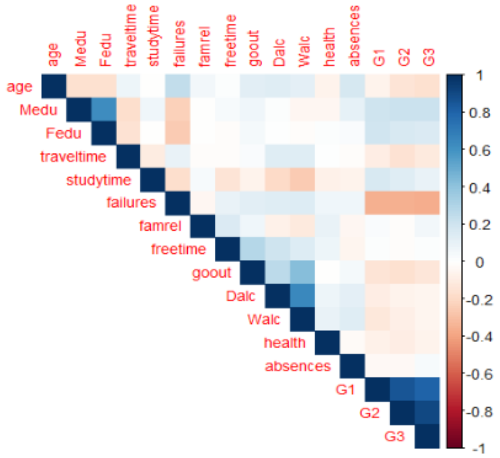


Fig. 2- Heatmap to show relationship between variables

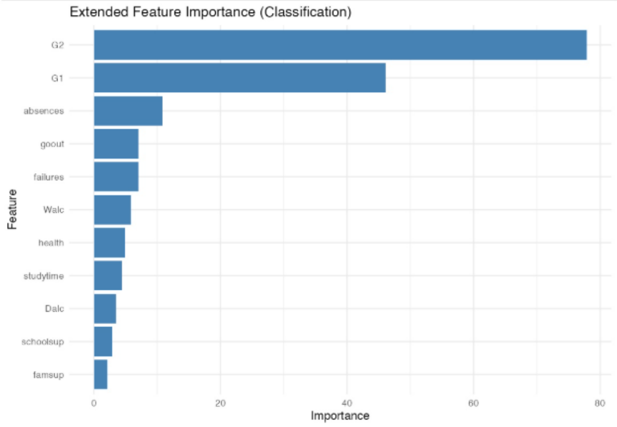


Fig. 3- Feature importance

2. Model Perormance

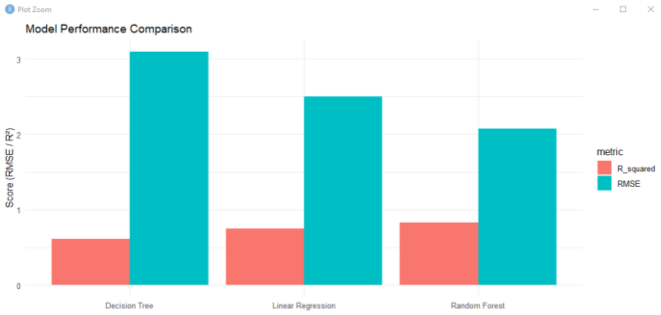
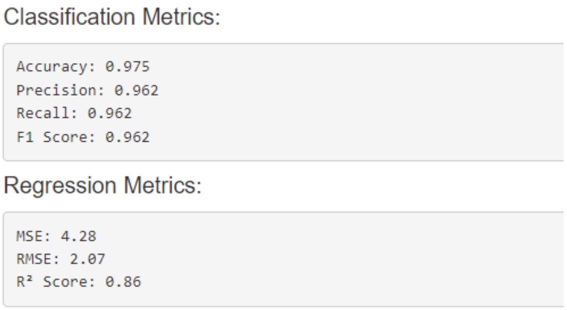


Fig. 4- Comparing Performance of Models – Decision Tree, Linear Regression and Random Forest. R² is highest and RMSE is lowest for Random Forest



Classification Metrics:

```
Accuracy: 0.975
Precision: 0.962
Recall: 0.962
F1 Score: 0.962
```

Regression Metrics:

```
MSE: 4.28
RMSE: 2.07
R² Score: 0.86
```

Fig. 5- Performance Metrics of Random Forst

Random Forest is chosen for this study due to its robust performance, ability to handle high-dimensional data, and its interpretability in terms of feature importance. It is an ensemble learning method that combines multiple decision trees to reduce overfitting and improve predictive accuracy.

Evaluation Metrics Used

To evaluate the performance of the models, the following metrics were used:

- **R-squared (R^2)**:
  This metric indicates how well the model explains the variability in the target variable. An R^2 value closer to 1 means the model explains a higher proportion of variance in student grades. Random Forest achieved an R^2 score of approximately *0.81*, indicating that it can explain 81% of the variation in final grades, which is significantly higher than other models tested.
- **Root Mean Square Error (RMSE)**:
  RMSE measures the average magnitude of prediction errors. Lower RMSE values indicate better model performance. The Random Forest model achieved a lower RMSE compared to Linear Regression and Decision Trees, suggesting more accurate predictions of student grades.

- **Accuracy (for classification)**: When student grades were categorized into performance levels (e.g., low, medium, high), Random Forest showed superior classification accuracy. It correctly identified the performance category for over *97%* of the students in the test dataset.

Why Random Forest Outperforms Others

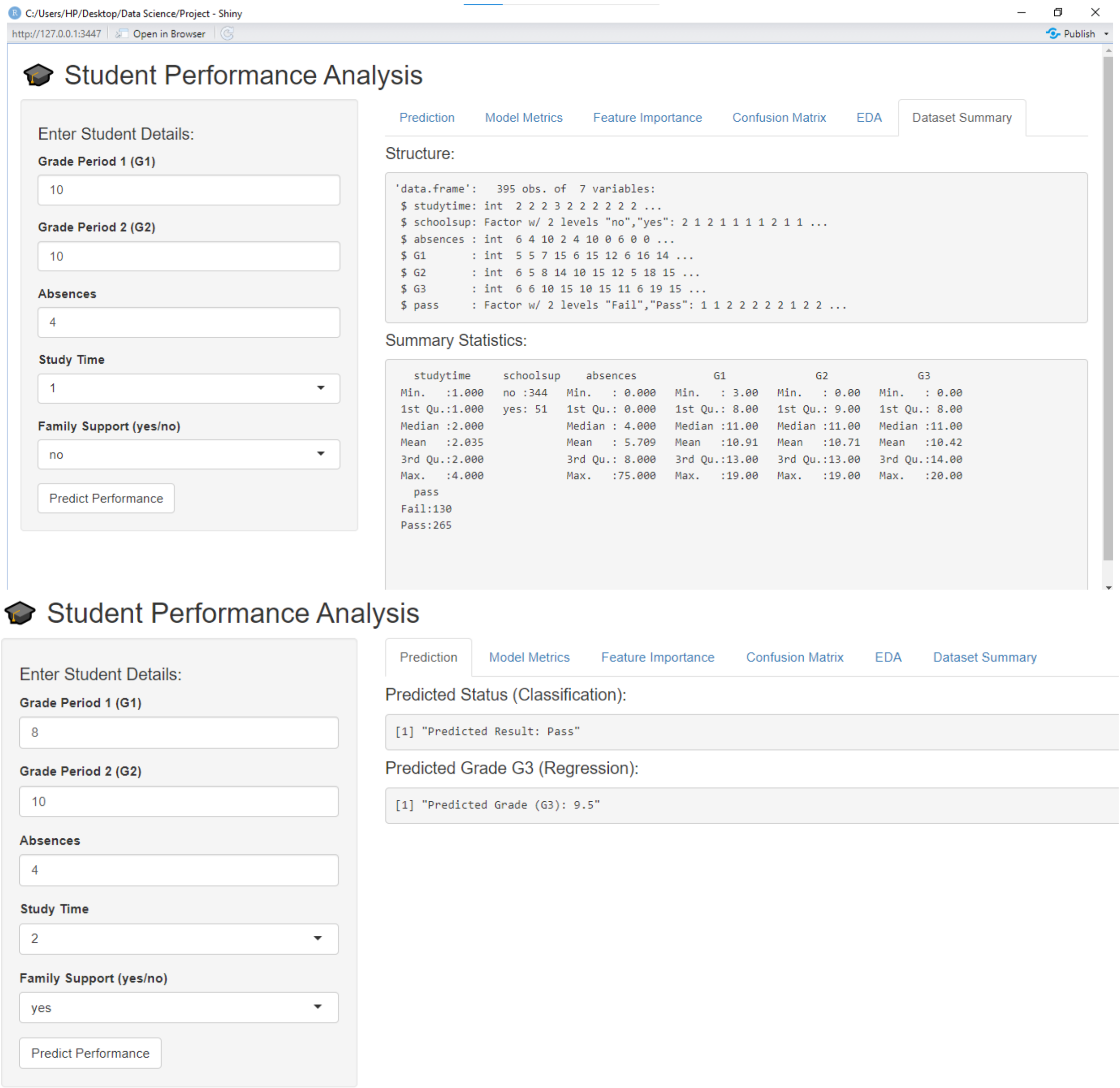
a) **Handles Non-linear Relationships**: Unlike linear models, Random Forest captures complex interactions and non-linear relationships among features.

b) **Reduces Overfitting**: By averaging multiple decision trees, Random Forest mitigates overfitting and generalizes better on unseen data.

c) **Feature Importance Insights**: Random Forest provides a ranked list of the most influential features, which helps in understanding the factors that most affect student performance.

d) **Robust to Noise and Missing Data**: Random Forest can handle missing values and outliers more effectively compared to simpler models.

3. Final Results



Student Performance Analysis

Enter Student Details:

Grade Period 1 (G1): 10

Grade Period 2 (G2): 10

Absences: 4

Study Time: 1

Family Support (yes/no): no

Structure:

```
'data.frame':   395 obs. of  7 variables:
 $ studytime: int  2 2 2 3 2 2 2 2 2 2 ...
 $ schoolsup: Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
 $ absences : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1       : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2       : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3       : int  6 6 10 15 10 15 11 6 19 15 ...
 $ pass     : Factor w/ 2 levels "Fail","Pass": 1 1 2 2 2 2 2 1 2 2 ...
```

Summary Statistics:

```
   studytime     schoolsup    absences          G1              G2              G3       
 Min.   :1.000   no :344   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:1.000   yes: 51   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00  
 Median :2.000             Median : 4.000   Median :11.00   Median :11.00   Median :11.00  
 Mean   :2.035             Mean   : 5.709   Mean   :10.91   Mean   :10.71   Mean   :10.42  
 3rd Qu.:2.000             3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :4.000             Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00  
   pass    
 Fail:130  
 Pass:265  
```

Student Performance Analysis

Enter Student Details:

Grade Period 1 (G1): 8

Grade Period 2 (G2): 10

Absences: 4

Study Time: 2

Family Support (yes/no): yes

Predicted Status (Classification):

```
[1] "Predicted Result: Pass"
```

Predicted Grade G3 (Regression):

```
[1] "Predicted Grade (G3): 9.5"
```

Fig 6- Predicted value is 9.5 and actual value is 10

IX. References

1. Shana, K. & Venkatacalam, K. "Faculty Support System (FSS) for Student Performance Analysis," Coimbatore Institute of Technology, Anna University.
2. SPA - Student Performance Analyser: A Web-Based System for Educators. Available at: SPAplatform
3. InMinds - Intelligent Mining and Decision Support System for UNIMAS.
4. UCI Machine Learning Repository - Student Performance Dataset. Available at: [https://archive.ics.uci.edu/ml/datasets/Student+Performance]