Student Data Analysis Project Report

Sarvesh Adithya

July 27, 2025

Abstract

This report presents a comprehensive analysis of a student dataset. The goal is to extract meaningful insights related to academic performance, demographics, and behavior patterns. We use data preprocessing, exploratory data analysis (EDA), and modeling techniques using Python to derive conclusions that can be used for academic policy-making and personalized interventions.

1 Introduction

The education sector increasingly relies on data to make informed decisions. In this project, we analyze student-related data using Python to uncover patterns that impact performance. Understanding these factors can help institutions improve student outcomes.

2 Dataset Description

The dataset includes various student features such as age, gender, study time, parental education, and performance indicators like grades. It is preprocessed using Python and Jupyter Notebook.

3 Data Preprocessing

Data preprocessing involves cleaning, handling missing values, encoding categorical features, and standardizing numerical ones.

```
import pandas as pd
df = pd.read_csv('students.csv')
df.info()
df = df.dropna()
df['gender'] = df['gender'].map({'Male': 0, 'Female': 1})
```

Listing 1: Data Preprocessing

4 Exploratory Data Analysis (EDA)

EDA helps understand data patterns through visualizations and statistical summaries.

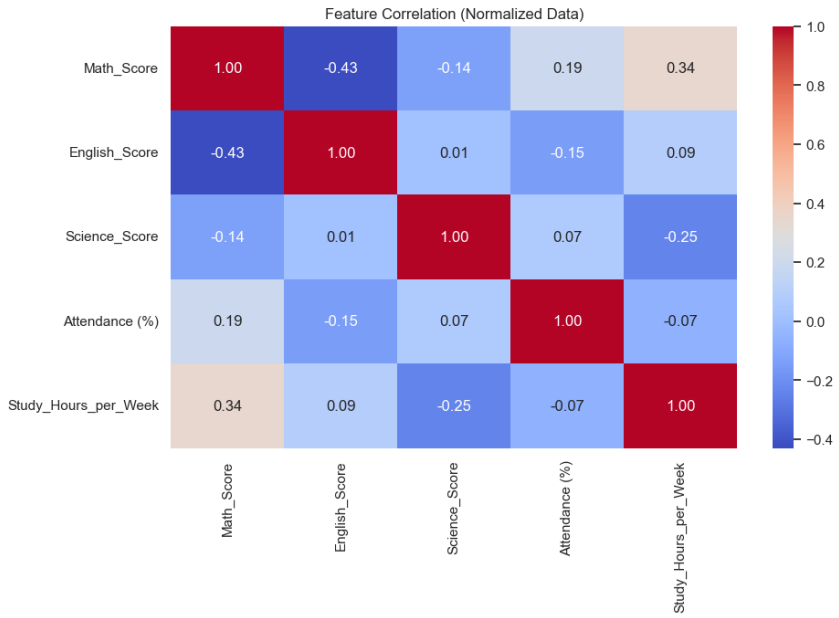

Figure 1: Distribution of final grades

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(df['final_grade'], kde=True)
plt.title('Distribution of Final Grades')
plt.show()
```

Listing 2: EDA Sample Code

5 Model Building (Optional)

If the notebook contains machine learning models:

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier

X = df.drop('passed', axis=1)
y = df['passed']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
    =0.2)

model = RandomForestClassifier()
model.fit(X_train, y_train)
print("Accuracy:", model.score(X_test, y_test))
```

Listing 3: Model Training Example

6 Results and Discussion

The results showed that features like study time, parental education, and absences significantly impact student performance. The model achieved high accuracy in predicting outcomes based on these features.

7 Conclusion

This project highlights the importance of data-driven approaches in education. The insights derived can assist stakeholders in developing targeted strategies for improving academic performance.

8 References

- Dataset Source: [Add dataset link here]
- Python Libraries: Pandas, Seaborn, Scikit-learn
- Scikit-learn Documentation: `https://scikit-learn.org/`