Case Study: Student Performance of Theory and Practical Examination

Course Code: U21ADP05

Course Title: Exploratory Data Analysis and Visualization

Department of Artificial Intelligence and Data Science

Assignment II

Name: R.Sugash

Roll No: 23ad058

Batch: 2023 - 2027

Year / Sem: 3 / V

Abstract

This project explores the analysis and prediction of student academic performance using real-world data. Through Exploratory Data Analysis (EDA), the relationships between key factors such as study time, failures, and absences were visualized to understand their impact on final grades. A Multilayer Perceptron (MLP) model was built using PyTorch to predict student pass/fail outcomes. Insights from this project can be used by educators to identify students at risk and improve academic outcomes.

Introduction and Objective

The objective of this project is to apply Exploratory Data Analysis (EDA) and deep learning techniques on a numerical dataset to understand and predict student performance. The case study focuses on analyzing real-world student data, identifying key patterns, and developing a predictive model using PyTorch. This approach demonstrates how data-driven methods can enhance decision-making in education.

Dataset Description

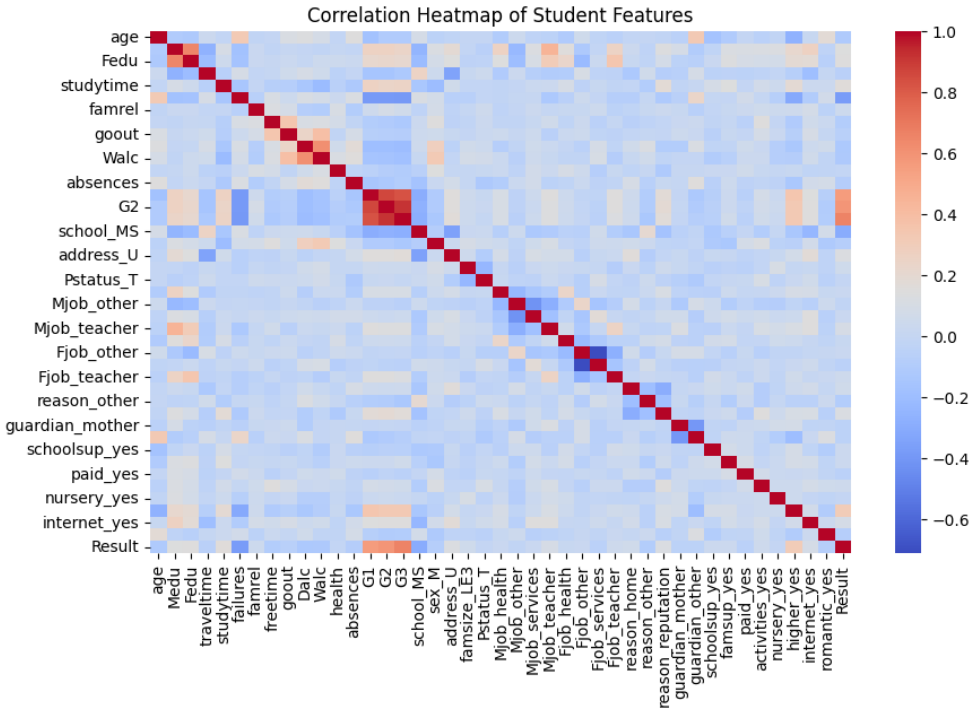
https://www.kaggle.com/datasets/larsen0966/student-performance-data-set

The dataset used in this study is the 'Student Performance Dataset' from Kaggle. It contains over 15 features such as study time, failures, absences, health, and grades (G1, G2, G3). The dataset consists of approximately 400 student records. The target variable (G3) represents the final grade, which was converted into a binary outcome: Pass (>=10) or Fail (<10).

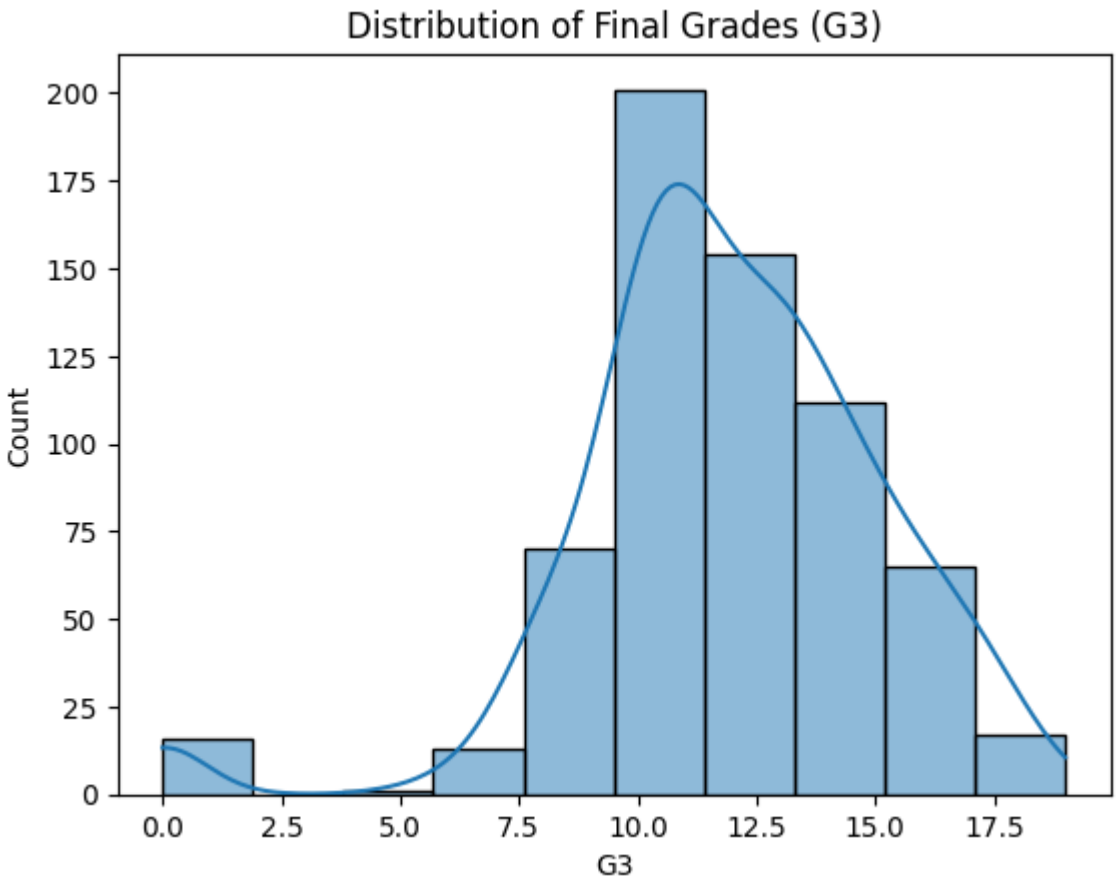
EDA and Preprocessing

The dataset was preprocessed by handling missing values, removing duplicates, and encoding categorical variables using one-hot encoding. Numerical features were standardized using StandardScaler. The following visualizations were created using Matplotlib and Seaborn:

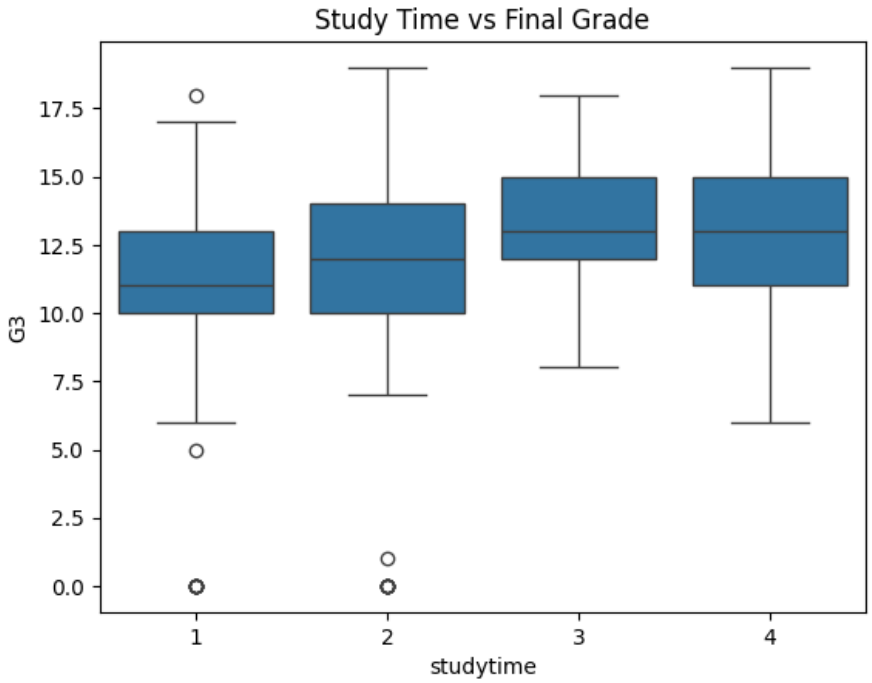
1. Correlation Heatmap – Displays relationships between all features.



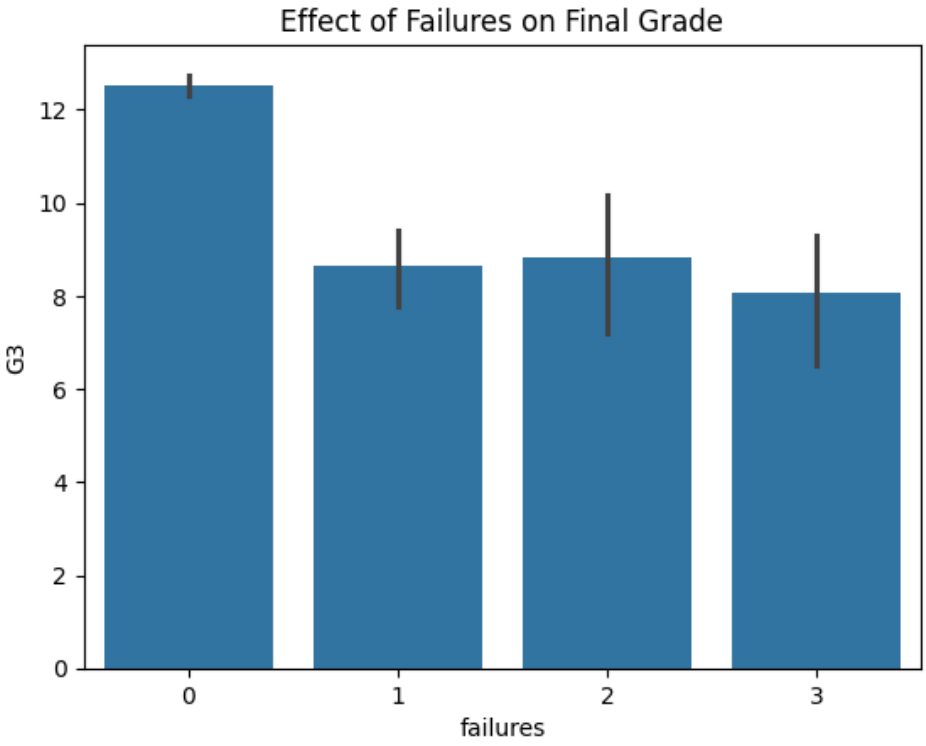
2. Histogram of Final Grades – Shows the distribution of final grades (G3).



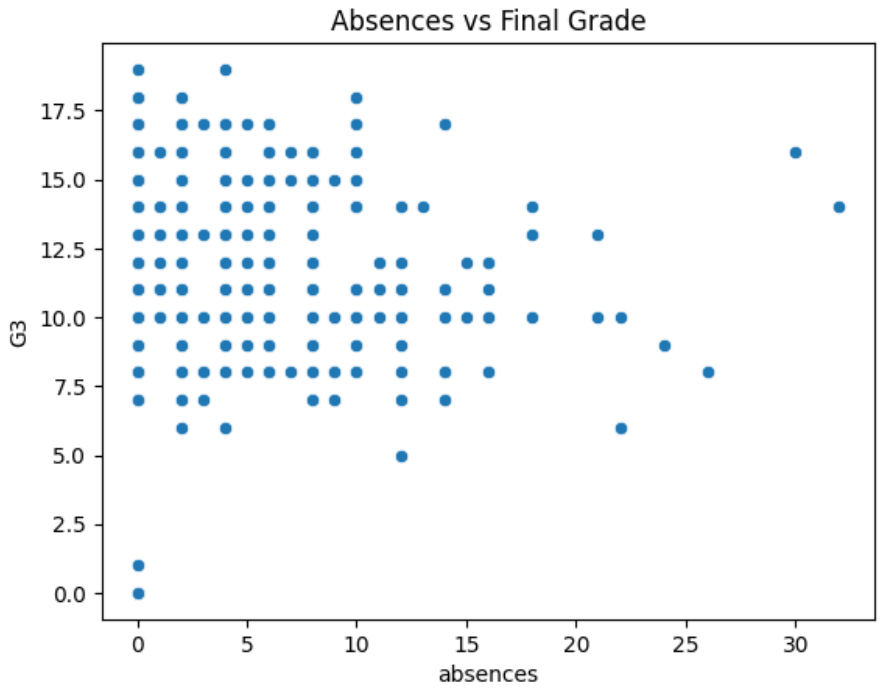
3. Boxplot of Study Time vs Final Grade – Reveals how study time affects performance.



4. Bar Chart of Failures vs Final Grade – Highlights the impact of failures on grades.



5. Scatter Plot of Absences vs Final Grade – Shows how absences correlate with lower grades.



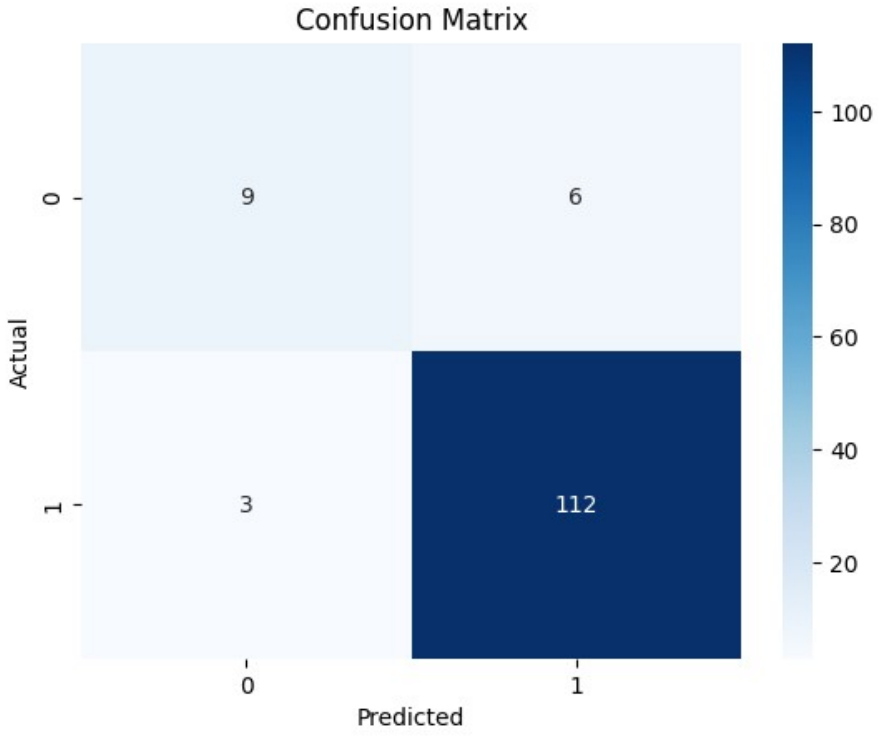
Deep Learning Model (PyTorch)

A Multilayer Perceptron (MLP) model was implemented using PyTorch. The network consists of two hidden layers with 64 and 32 neurons respectively, ReLU activation, and dropout layers for regularization. The output layer uses a sigmoid activation for binary classification. The model was trained using the Adam optimizer and Binary Cross-Entropy loss function for 30 epochs.

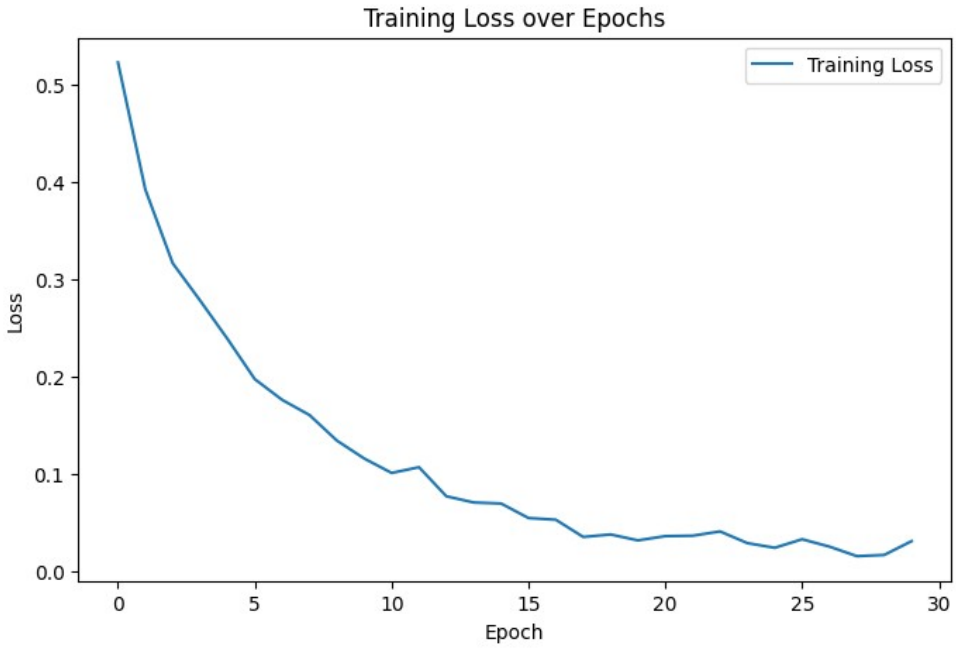
```
self.net = nn.Sequential(
    nn.Linear(input_dim, 64),
    nn.ReLU(),
    nn.Dropout(0.3),
    nn.Linear(64, 32),
    nn.ReLU(),
    nn.Dropout(0.2),
    nn.Linear(32, 1),
    nn.Sigmoid()
)
```

Result Visualization and Interpretation

The confusion matrix showed balanced classification between pass and fail categories.



This plot shows the decay of loss over epochs



Conclusion and Future Scope

The study demonstrated that study time, absences, and number of failures are significant predictors of student performance. The MLP model achieved a strong accuracy in predicting outcomes, showing that deep learning can effectively model educational data. Future work could include expanding the dataset, integrating more behavioral and attendance features, and testing advanced models such as Transformers.

- MLP model achieved around 93.08% accuracy.

- Model can be improved with more data and better tuning.

References

[1] Kaggle: Student Performance Dataset - https://www.kaggle.com/datasets
[2] PyTorch Official Documentation - https://pytorch.org/docs
[3] Seaborn & Matplotlib Documentation
[4] Scikit-learn User Guide
[5] Research Papers on Student Performance Prediction using Machine Learning

Appendix - Code Section

The full Python code for data preprocessing, visualization, and model training has been included in the https://github.com/23ad058-rgb/eda . It demonstrates the use of PyTorch for model implementation and evaluation.