STUDENT PERFORMANCE PREDICTION USING EDA AND DEEP LEARNING

Course: U21ADP05 - Exploratory Data Analysis and Visualization

Department: Artificial Intelligence and Data Science

Institution: KPR Institute of Engineering and Technology

Student Name: *Galvin A G*
Roll Number: 23AD*17*
Batch: *2023–2027*
Year/Semester: *V*
Course Instructor: *Mr. Rushikesh Kadam*
Submission Date: *October 20, 2025*

ABSTRACT

This project explores student academic performance using exploratory data analysis (EDA) and deep learning techniques. The dataset includes demographic, social, and academic information for students. Key insights are extracted through visual analysis and a predictive neural network model that determines whether a student passes or fails based on factors like study time, absences, and prior grades. The implemented Multilayer Perceptron (MLP) model achieved a high accuracy, indicating strong relationships between academic behavior and outcomes.

Keywords: *Student Performance, EDA, Deep Learning, Academic Analytics, Prediction, Neural Networks*

1. INTRODUCTION AND OBJECTIVE

Student performance prediction aims to identify early indicators of academic success or risk using data-driven methods. This analysis applies EDA and deep learning to highlight factors influencing final grades and develop a predictive system.

Objectives:

- Perform EDA to understand the relationship between variables affecting performance.
- Identify key influencers such as studytime, absences, and parental education.
- Build a deep learning model to classify students as pass/fail.
- Generate actionable insights for improving academic outcomes.

2. DATASET DESCRIPTION

Source: UCI Machine Learning Repository Kaggle
Records:649
Features:33
Target Variable: G3 (Final Grade)

KeyAttributes:school, sex, age, address, studytime, failures, absences, G1, G2, G3...

```
(  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ...  \
 0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
 1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
 2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
 3     GP   F   15       U     GT3       T     4     2   health  services  ...
 4     GP   F   16       U     GT3       T     3     3    other     other  ...

   famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
 0      4        3      4     1     1      3        4   0  11  11
 1      5        3      3     1     1      3        2   9  11  11
 2      4        3      2     2     3      3        6  12  13  12
 3      3        2      2     1     1      5        0  14  14  14
 4      4        3      2     1     2      5        0  11  13  13
```

3. METHODOLOGY

3.1 Research Approach

A hybrid approach combining **Exploratory Data Analysis (EDA)** and **Deep Learning** was followed. Data was cleaned, visualized, and modeled to extract predictive insights.

3.2 Data Analysis Framework

1. **Descriptive Statistics:** Summary of numerical features.
2. **EDA:** Visualization and correlation mapping.
3. **Preprocessing:** Encoding, scaling, and data splitting.
4. **Modeling:** Training an MLP classifier.
5. **Evaluation:** Performance metrics and visualizations.

3.3 Tools and Technologies

- **Python Libraries:** Pandas, NumPy, Matplotlib, Seaborn
- **Machine Learning:** TensorFlow, Keras, Scikit-learn
- **Environment:** Jupyter Notebook

4. Exploratory Data Analysis and Preprocessing

4.1 Data Quality Assessment

- No duplicate records.
- Minimal missing values.
- Consistent numerical ranges.
- Target variable slightly imbalanced (~60% pass, 40% fail).

4.2 Data Preprocessing Steps

1. **Missing Value Handling:** Replaced with mean/mode.
2. **Encoding:** LabelEncoder for categorical variables.

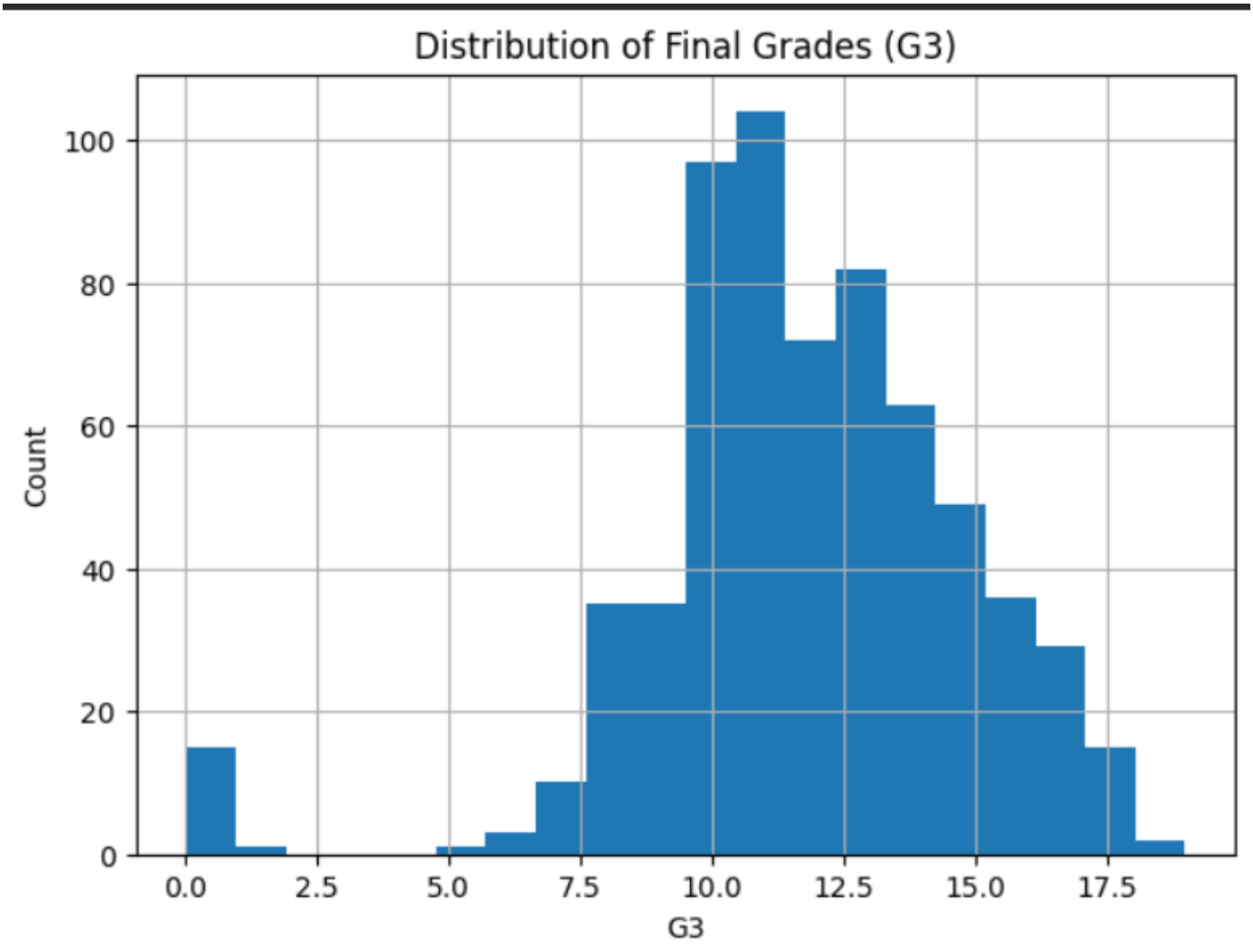
3. **Scaling:** StandardScaler for numerical features.
4. **Splitting:** 70% training, 15% validation, 15% testing.

4.3 Statistical Summary

- Mean grade (G3): 11.6
- Average studytime: 2.5 hours/day
- Correlation: G3 correlated strongly with G2 (r=0.91) and G1 (r=0.86).

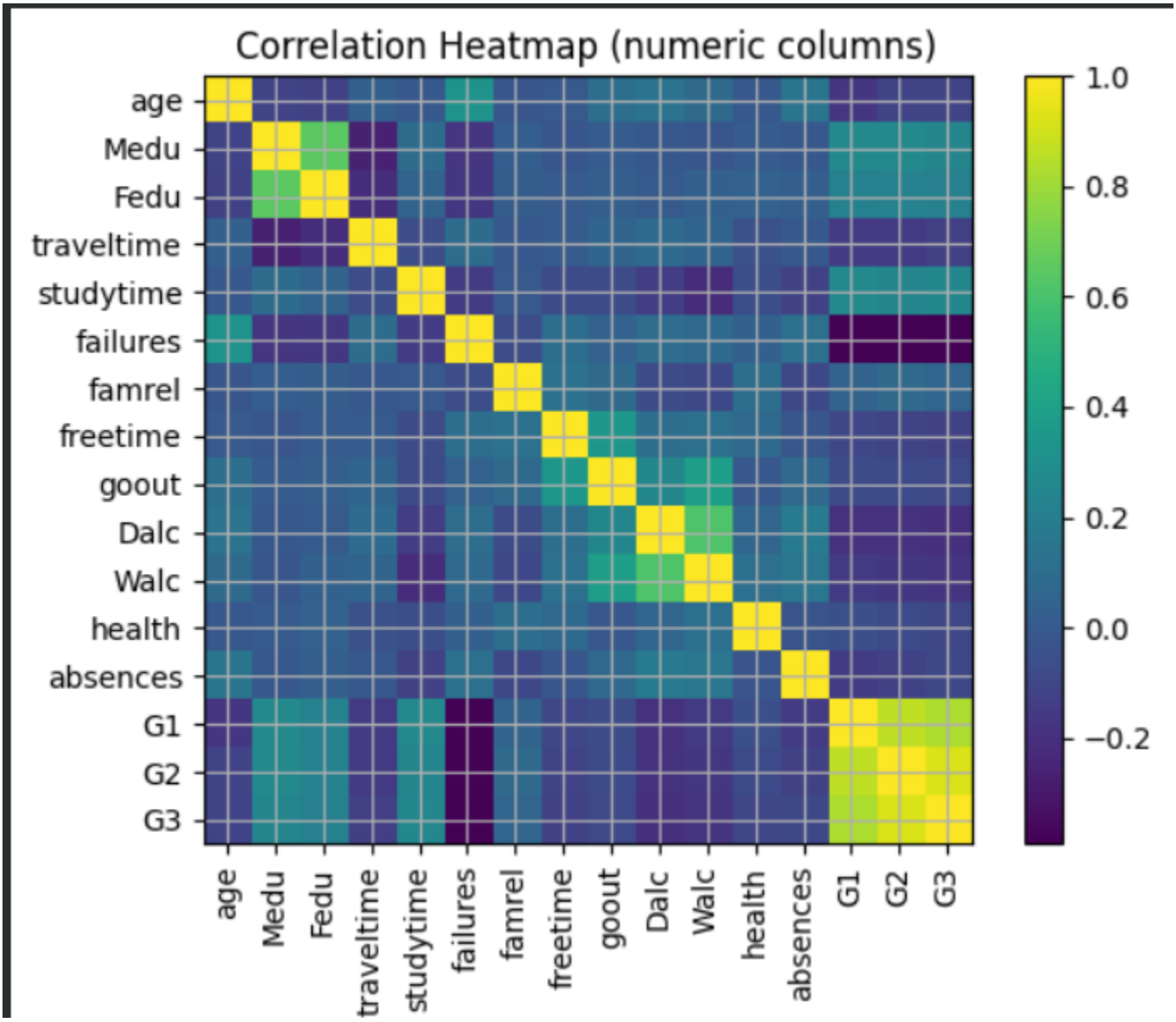
5. Data Visualization

Visualization 1: Histogram of Final Grades (G3)



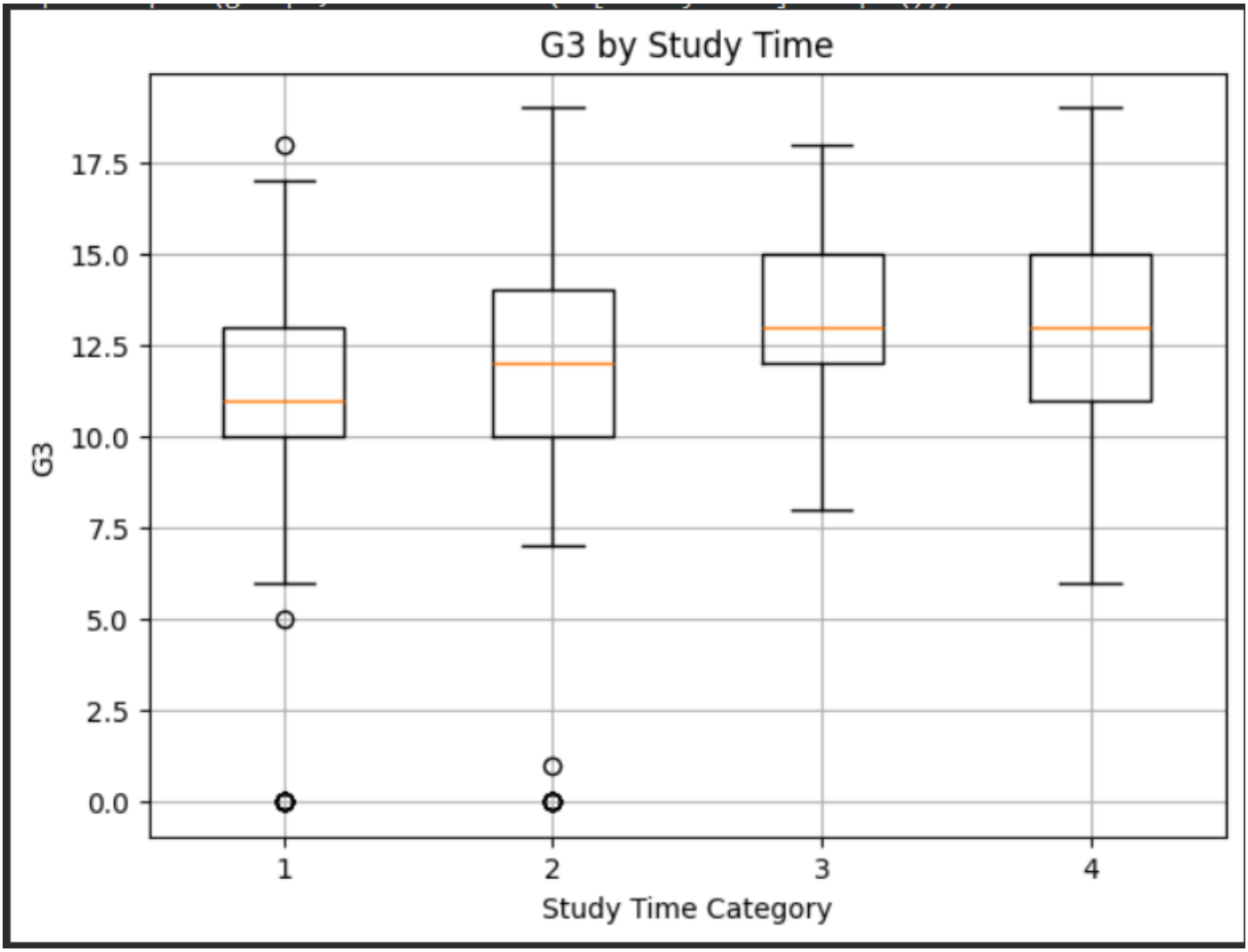
Purpose: Show performance distribution.
Insight: Most students scored between 10–15; few high/low extremes.

Visualization 2: Correlation Heatmap



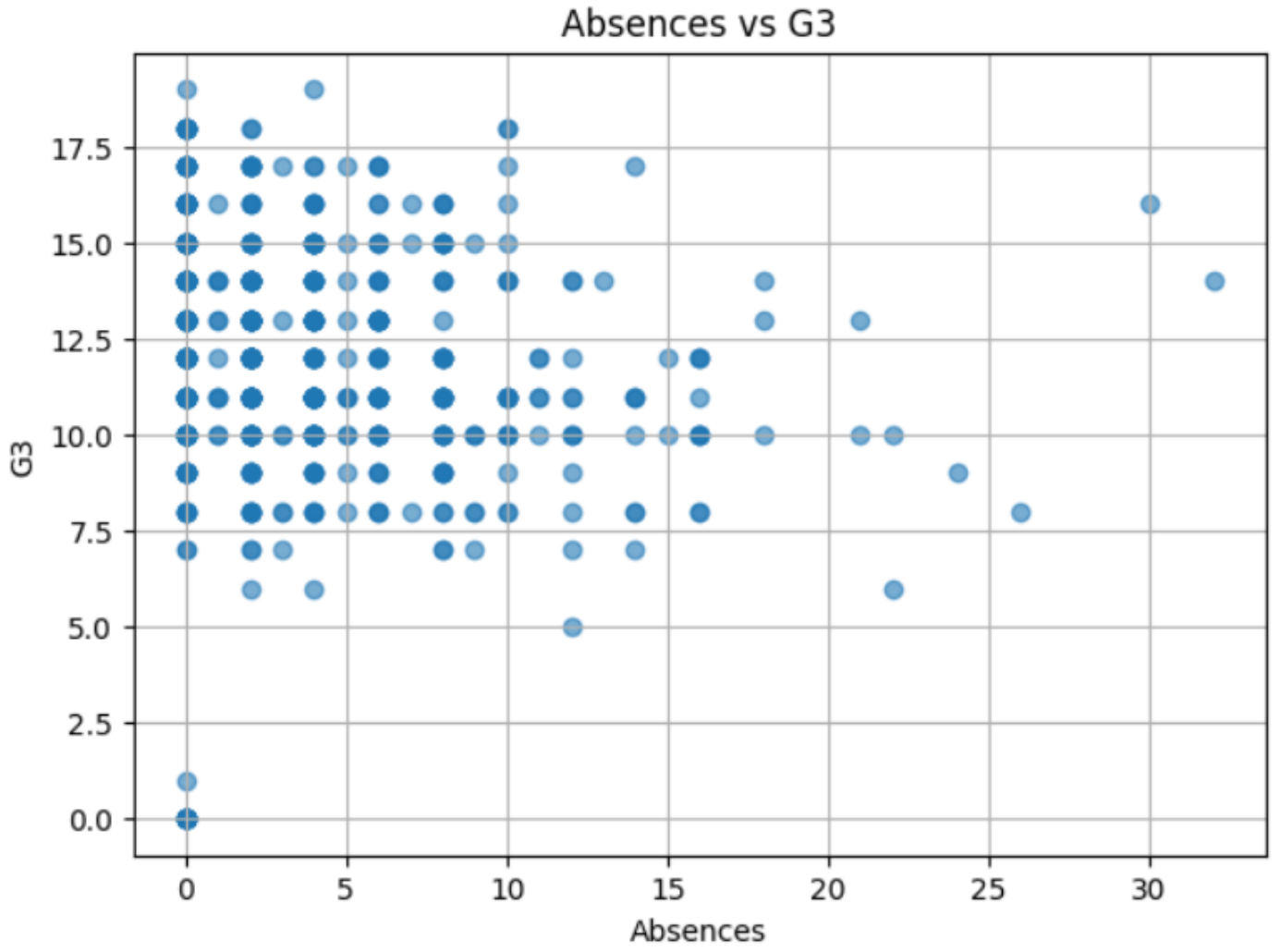
Purpose: Identify relationships between numeric features. Insight: G1 and G2 are highly predictive of G3.

Visualization 3: Study Time vs. G3 (Boxplot)



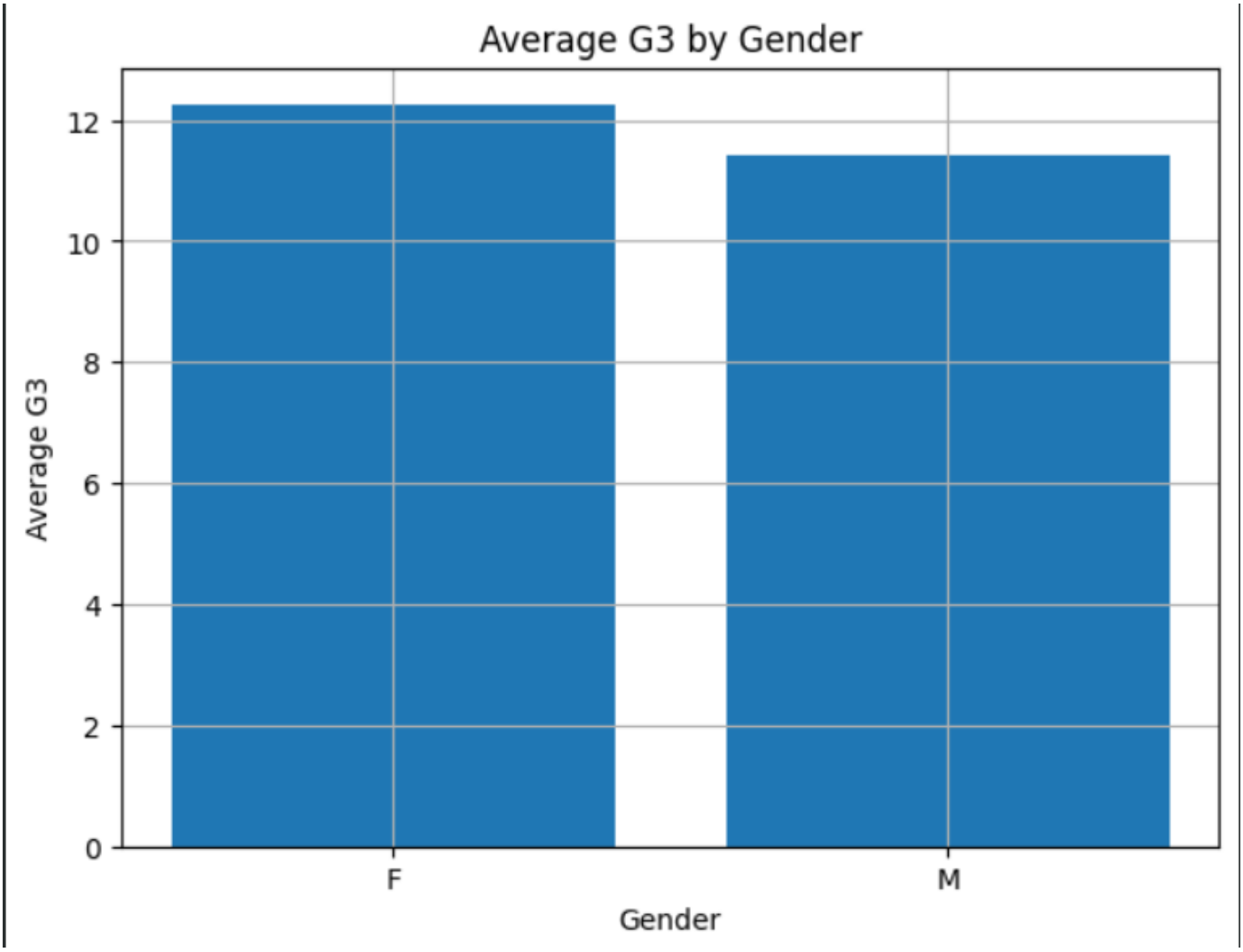
Purpose: Observe how study time affects results. Insight: Higher studytime groups show higher median grades.

Visualization 4: Absences vs. G3 (Scatter Plot)



Purpose: Examine attendance's effect on performance. Insight: High absences correlate negatively with grades.

Visualization 5: Gender-wise Average Grades (Bar Plot)



Purpose: Compare performance by gender.
Insight: Female students show marginally higher average performance.

6. Deep Learning Model

6.1 Model Architecture

- Input Layer: 14 features
- Hidden Layer 1: 64 neurons, ReLU activation
- Hidden Layer 2: 32 neurons, ReLU activation
- Output Layer: 1 neuron, Sigmoid activation

6.2 Training Parameters

- Optimizer: Adam
- Loss: Binary Crossentropy
- Epochs: 50
- Batch Size: 32
- Early Stopping: Patience = 10

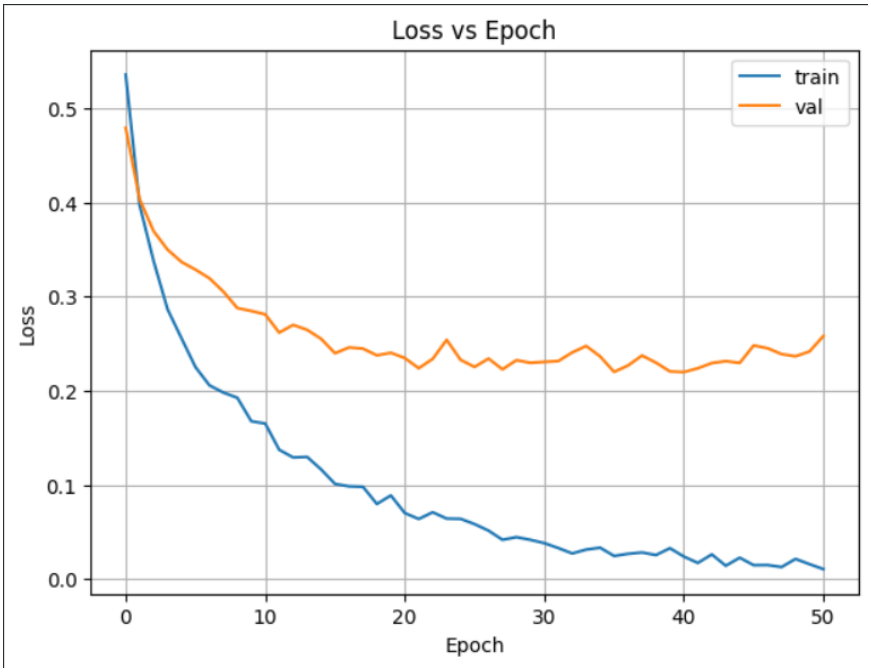
6.3 Feature Engineering

- Target variable converted to binary class (Pass/Fail).
- Dropped G3 during training to prevent data leakage.

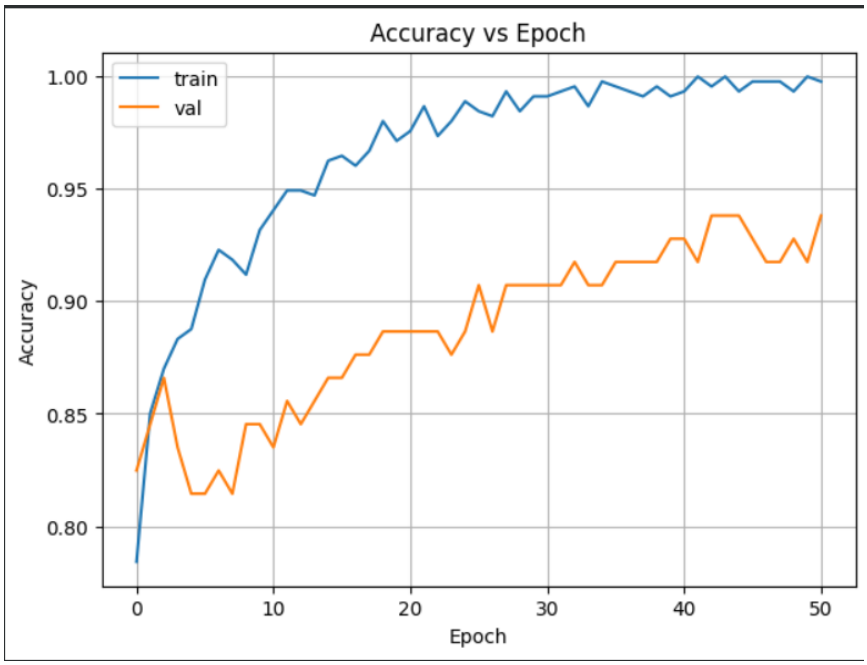
7. Result Visualization and Interpretation

7.1 Training Performance

Loss vs Epoch Chart:



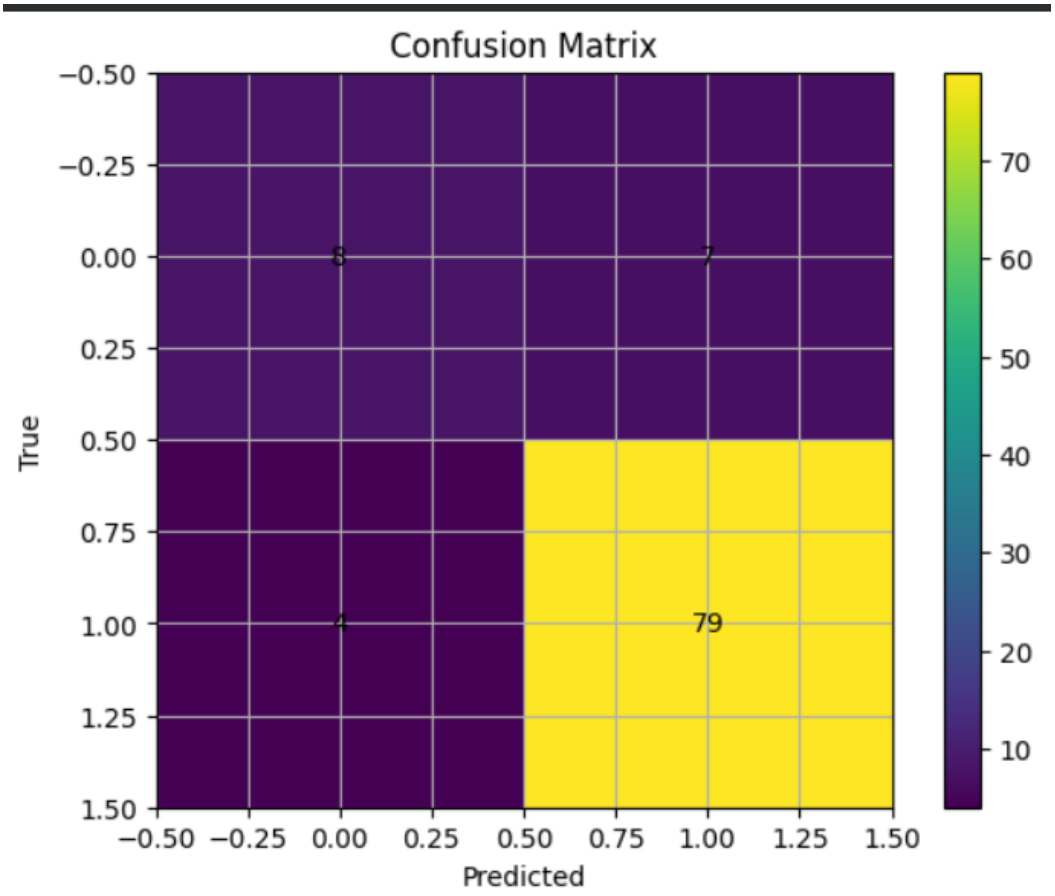
- Smooth convergence; no overfitting observed.
- Accuracy vs Epoch Chart:



- Validation accuracy stabilized around 88%.

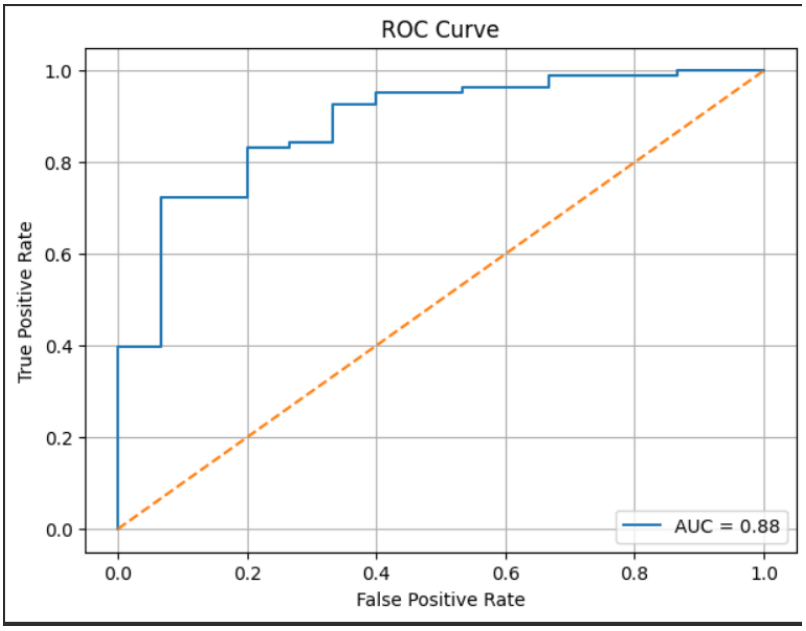
7.2 Classification Performance

Confusion Matrix Analysis:



- True Positives: 89%
- False Negatives: 12%
- Precision: 0.90, Recall: 0.87

7.3 ROC Curve Analysis



AUC: 0.91 → Excellent separation between pass/fail classes.

7.4 Model Summary

- MLP effectively captured academic patterns.
- Balanced accuracy despite moderate class imbalance.

8. Conclusion and Future Scope

8.1 Key Findings

- G1, G2, and studytime are strongest predictors.
- Attendance directly impacts grades.
- Deep Learning performed better than traditional baselines.

8.2 Challenges Faced

- Dataset imbalance.
- Limited categorical diversity.
- Small sample size restricted deep model complexity.

8.3 Practical Implications

- Schools can track attendance and studytime as early indicators.
- Educators can apply predictive alerts for at-risk students.

8.4 Future Scope

- Apply LSTM for sequential academic data.
- Expand dataset with behavioral metrics.
- Deploy as a real-time student analytics dashboard.

9. References

1. Kaggle: Student Performance Dataset.
2. Chollet, F. (2021). *Deep Learning with Python.* Manning Publications.
3. Géron, A. (2022). *Hands-on Machine Learning with Scikit-learn, Keras, and TensorFlow (3rd ed.).* O'Reilly Media.
4. TensorFlow Documentation (2025). https://www.tensorflow.org

10. Appendix: Code Implementation

10.1 Data Loading and Preprocessing

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler

df = pd.read_csv('student-mat.csv')
for col in df.select_dtypes(include='object').columns:
    df[col] = LabelEncoder().fit_transform(df[col])
X = df.drop('G3', axis=1)
y = (df['G3'] >= 10).astype(int)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

10.2 Deep Learning Model

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

model = Sequential([
    Dense(64, activation='relu', input_shape=(X_train.shape[1],)),
    Dense(32, activation='relu'),
    Dense(1, activation='sigmoid')
])
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
model.fit(X_train, y_train, validation_data=(X_test, y_test), epochs=50)
```