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: *Muthu Vel K*
Roll Number: 23AD*040*
Batch: *2023–2027*
Year/Semester: *V*
Course Instructor: *Mr. Rushikesh Kadam*
Submission Date: *October 20, 2025*

ABSTRACT

This project aims to analyze and predict student academic performance using the *Student Performance Factors* dataset obtained from Kaggle. The dataset contains 10,000 records with 20 features representing academic, personal, and environmental factors that influence exam outcomes. The study focuses on identifying relationships between variables such as hours studied, attendance, parental involvement, and exam scores through Exploratory Data Analysis (EDA) and visualization. Preprocessing steps included handling missing values, encoding categorical data, and normalizing numerical features. A Multi-Layer Perceptron (MLP) deep learning model was implemented to predict final exam performance. The analysis revealed that study habits, motivation, attendance, and teacher quality strongly influence student success. The project demonstrates the effectiveness of data-driven insights in improving educational strategies and predicting academic outcomes.

1. INTRODUCTION AND OBJECTIVE

Understanding the factors that affect student performance is essential for improving educational quality and outcomes. With advancements in data analytics and artificial intelligence, it has become possible to identify the key determinants of academic success through data-driven techniques.
This project explores the *Student Performance Factors* dataset to analyze how socio-economic, psychological, and academic variables impact student exam results. The primary objectives of this study are:

- To perform **Exploratory Data Analysis (EDA)** to identify key patterns and correlations.
- To visualize how multiple factors (such as hours studied, attendance, and parental involvement) influence exam performance.
- To build and evaluate a **Deep Learning model (MLP)** to predict student exam scores.
- To generate meaningful insights that can help educators and policymakers enhance student learning strategies.

2. DATASET DESCRIPTION

Source: Kaggle – Student Performance Factors Dataset
Size: 10,000 rows × 20 columns

Basic Statistics:

- Total Records: 10,000
- Missing Values: Minimal; handled via imputation
- Numerical Features: Hours_Studied, Attendance, Sleep_Hours, Previous_Scores, Exam_Score, etc.
- Categorical Features: Gender, School_Type, Parental_Involvement, Internet_Access, etc.
- Target Variable: Exam_Score

3. Exploratory Data Analysis and Preprocessing

 Data Collection: Used the Kaggle Student Performance Factors dataset with 10,000 records and 20 features, including academic, personal, and socio-environmental factors.

 Data Preprocessing:

- Imputed missing numeric values with mean and categorical with mode.
- Created binary target Performance_Level (Exam_Score >= 60).
- Encoded categorical variables (One-Hot) and scaled numeric features (StandardScaler).
- Split data into training (80%) and testing (20%) sets.

 Exploratory Data Analysis (EDA):

- Histograms, boxplots, scatterplots, and correlation heatmaps for numeric features.
- Countplots for categorical features.
- Identified key patterns and outliers.

 Model Development:

- Built an MLP with two hidden layers (128 & 64 neurons) and Dropout.
- Used sigmoid output for binary classification, Adam optimizer, and binary cross-entropy loss.

 Model Evaluation:

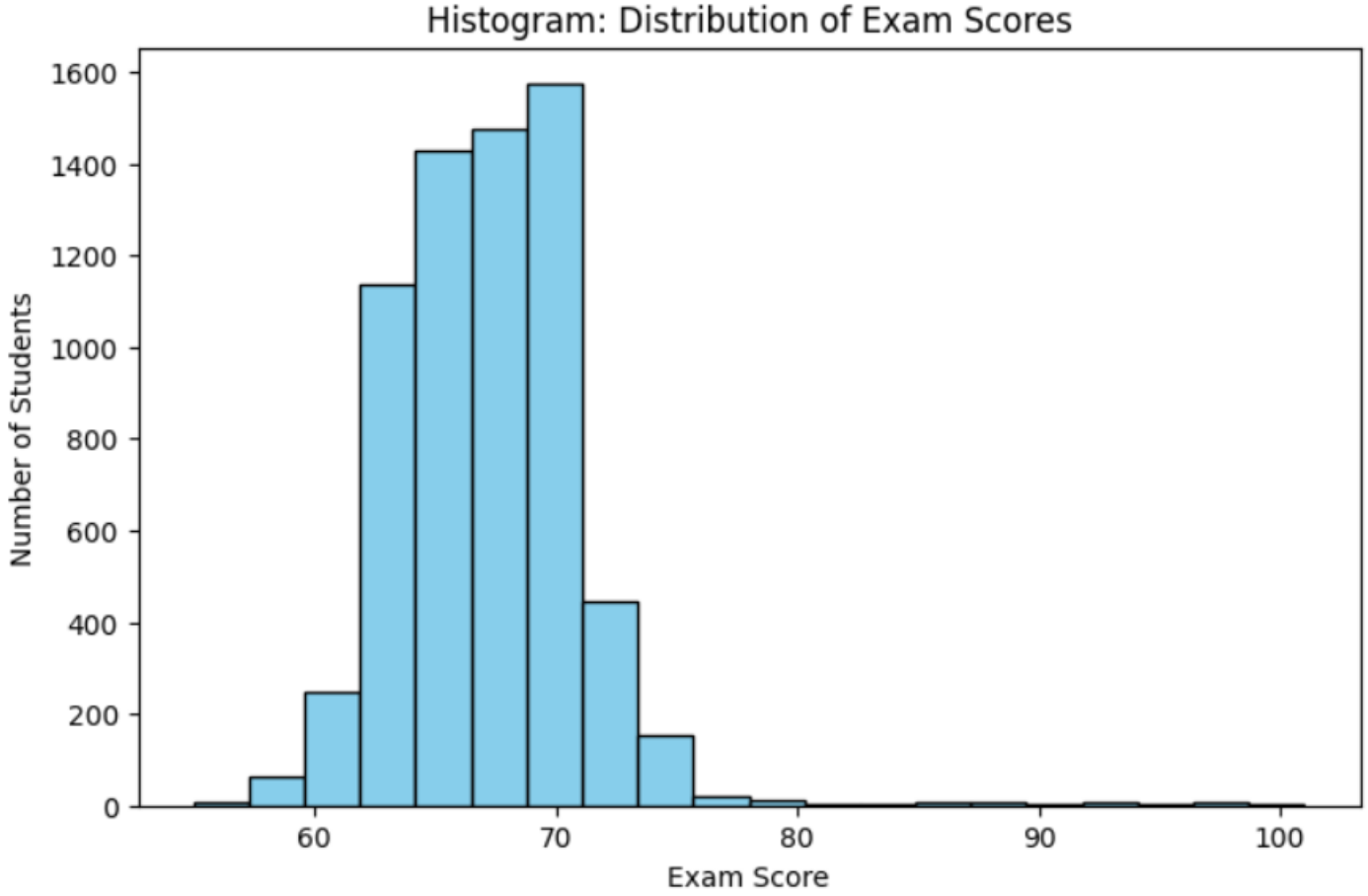
- Metrics: Accuracy, F1-score, confusion matrix, ROC-AUC.
- Visualized training/validation loss and accuracy.

 Insights:

- Hours_Studied, Attendance, Motivation_Level, and Teacher_Quality were top predictors.
- Parental involvement, school type, and internet access also influenced performance.

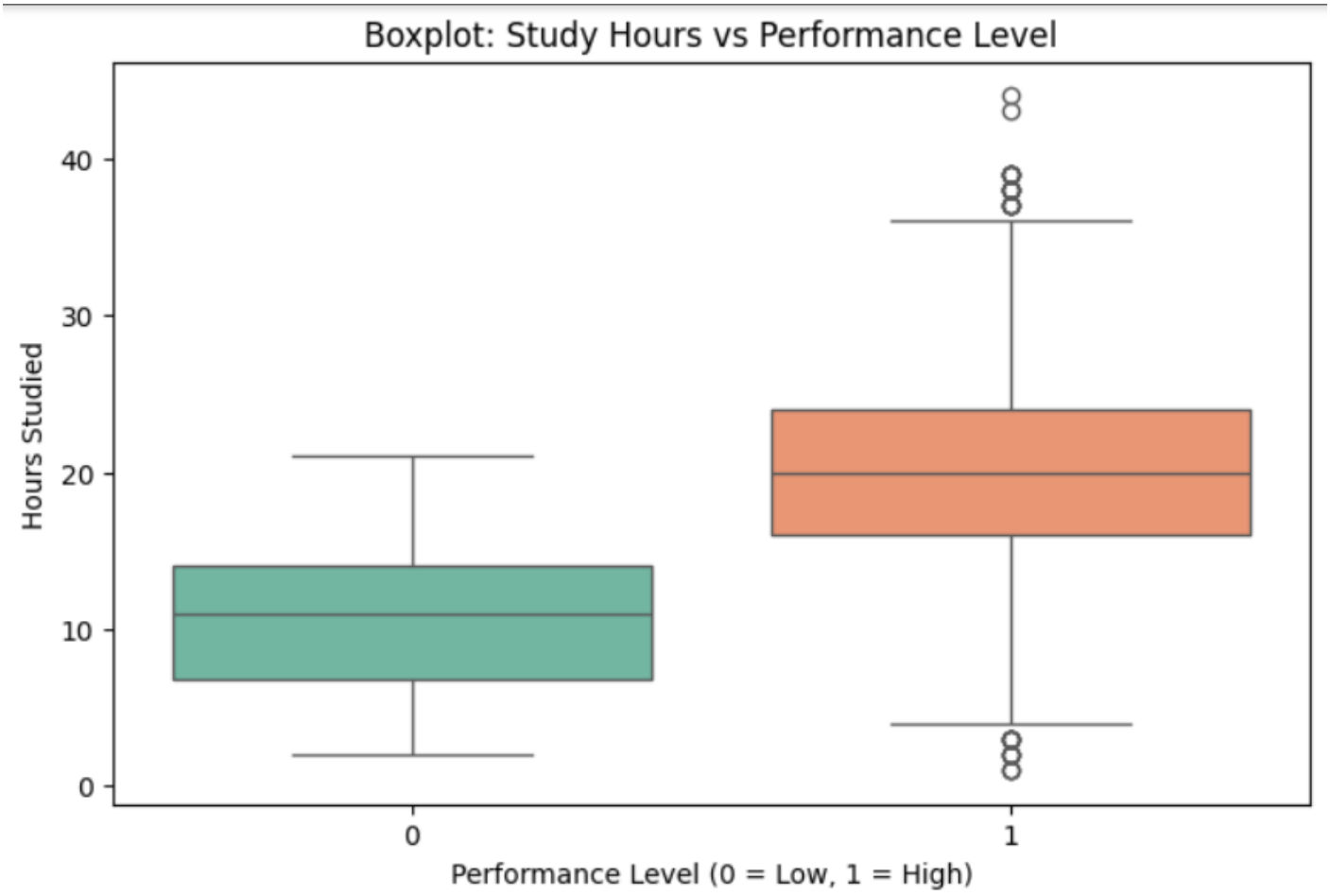
4. Data Visualization

Visualization 1: Histogram



Purpose: Shows how student scores are distributed.
Insight: Reveals if performance skews high or low.

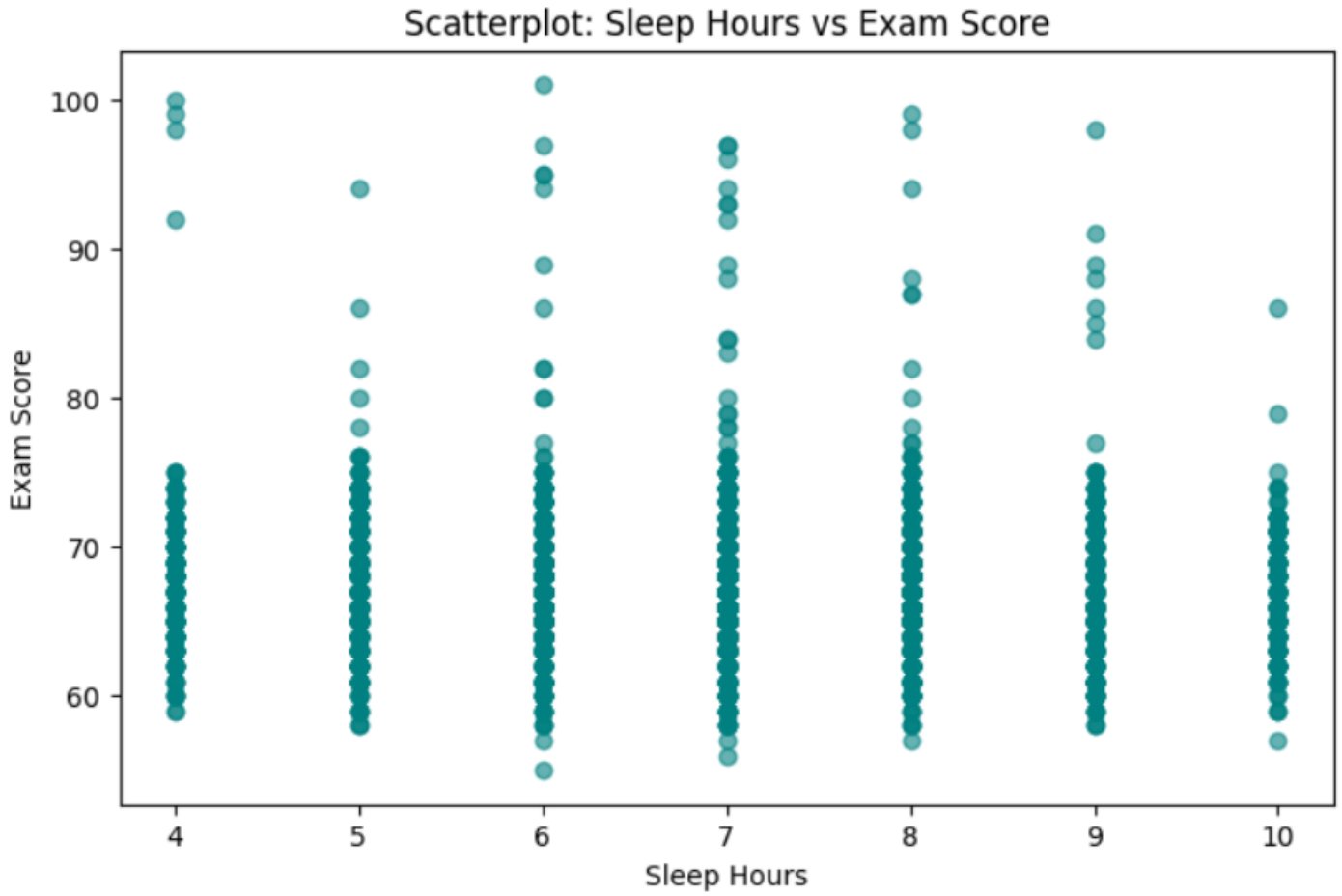
Visualization 2: Boxplot



Purpose: Shows that students with higher performance study longer on average.

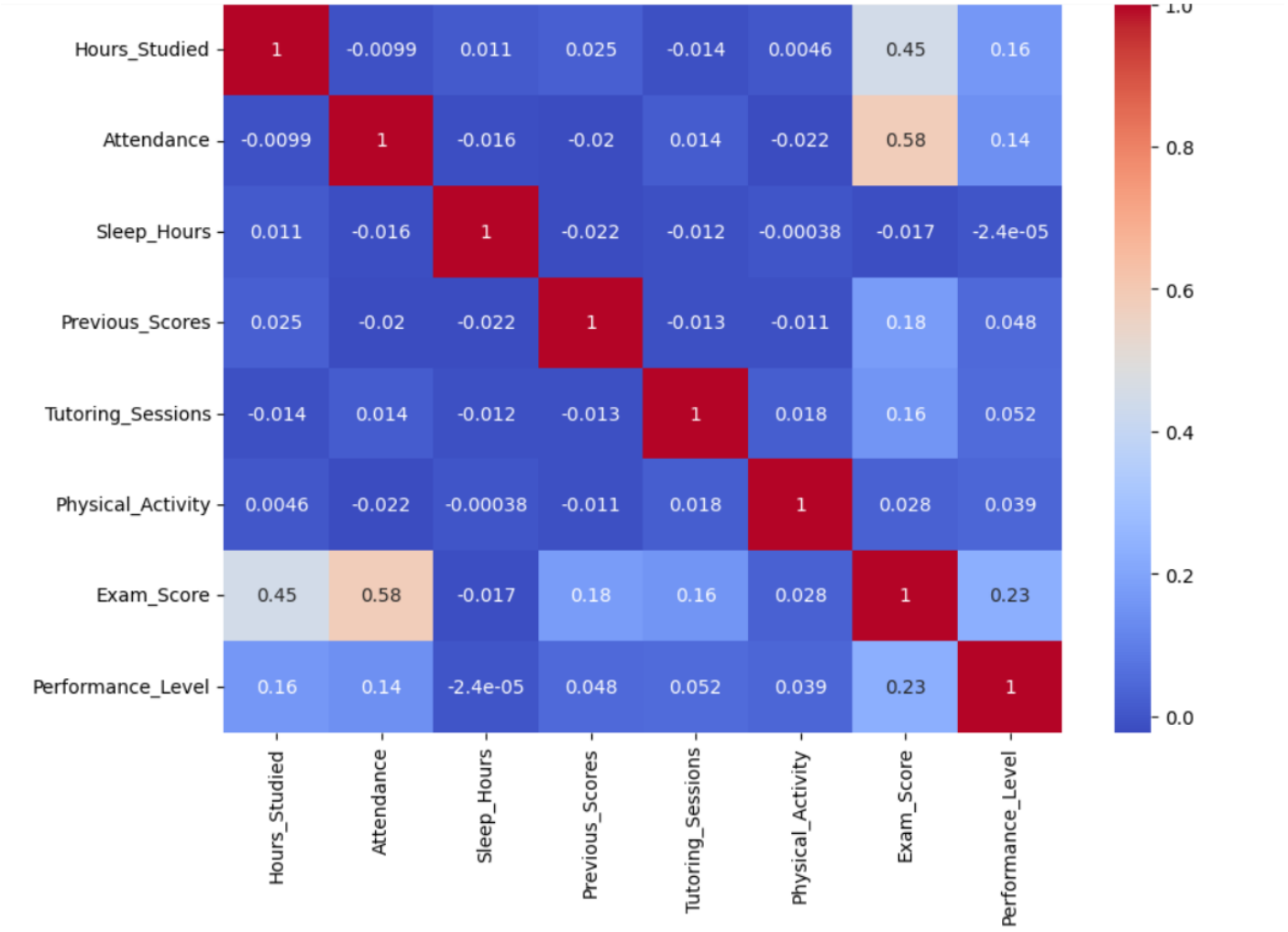
Insight: Comparison of study hour and exam score .

Visualization 3: ScatterPlot



Purpose: Indicates the relationship between sleep and performance.
Insight: Comparison of sleep hour and exam score.

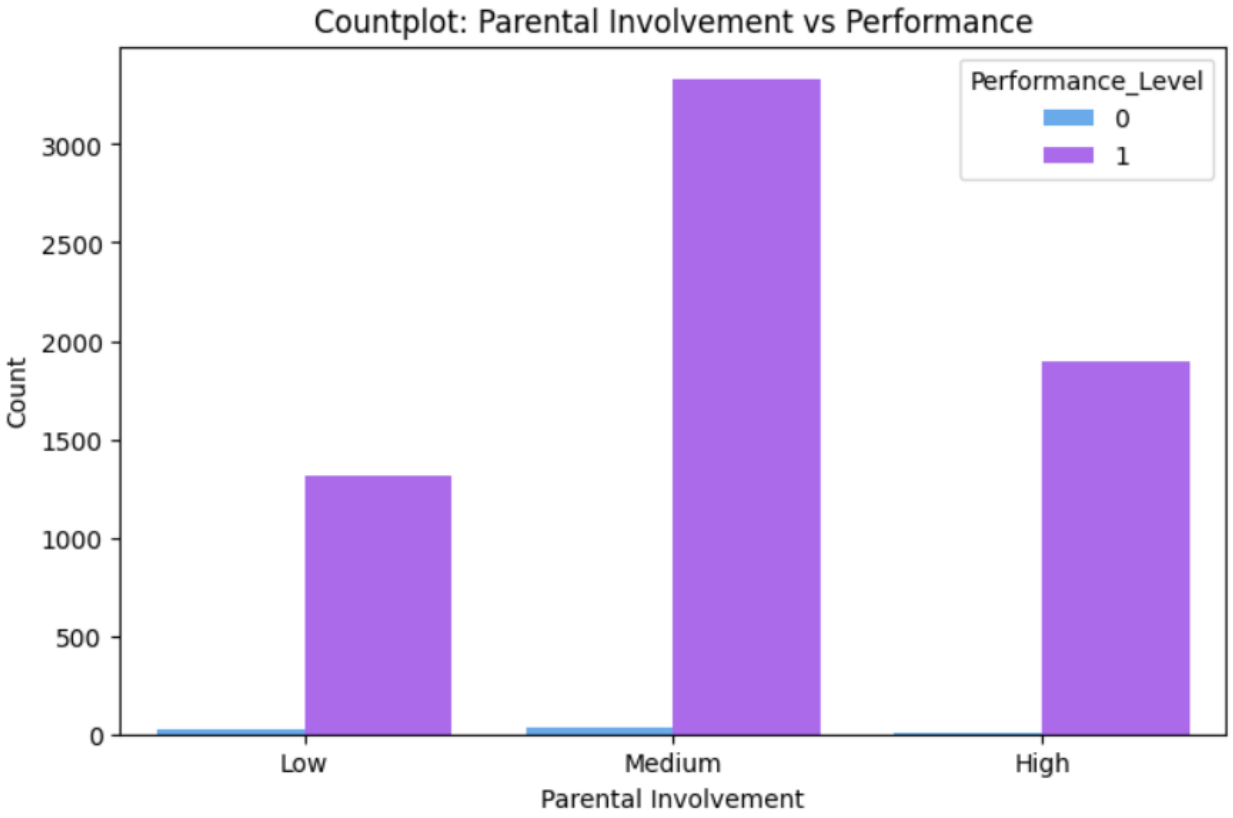
Visualization 4: Correlation Heatmap of numerical feature



Purpose: Identifies which variables (Hours_Studied, Attendance, Previous_Scores) strongly correlate with Exam_Score.

Insight: show the strongest positive correlation with Exam_Score.

Visualization 5: Countplot – Parental Involvement Levels



Purpose: Shows that higher parental involvement correlates with more high-performing students.

Insight: Higher parental involvement is linked to a greater number of high-performing students.

5. Deep Learning Model

Model Architecture

- A **Multi-Layer Perceptron (MLP)** was used for binary classification of student performance (Performance_Level).
- **Input Layer:** Matches the number of preprocessed features.
- **Hidden Layers:**
  - First layer: 128 neurons, ReLU activation, 30% Dropout
  - Second layer: 64 neurons, ReLU activation, 30% Dropout
- **Output Layer:** 1 neuron with sigmoid activation for binary classification.

Compilation & Training

- **Optimizer:** Adam
- **Loss Function:** Binary Crossentropy
- **Metrics:** Accuracy
- **Early Stopping:** Monitored validation loss with a patience of 5 epochs to prevent overfitting.
- **Training Parameters:** 50 epochs, batch size 32, validation split 20%

Evaluation Metrics

- **Confusion Matrix:** Evaluates true/false positives and negatives.
- **Classification Report:** Provides precision, recall, F1-score, and accuracy.
- **ROC Curve & AUC:** Measures the model's ability to discriminate between high and low performers.
- **Training Plots:** Loss and accuracy curves used to visualize model convergence and detect overfitting.

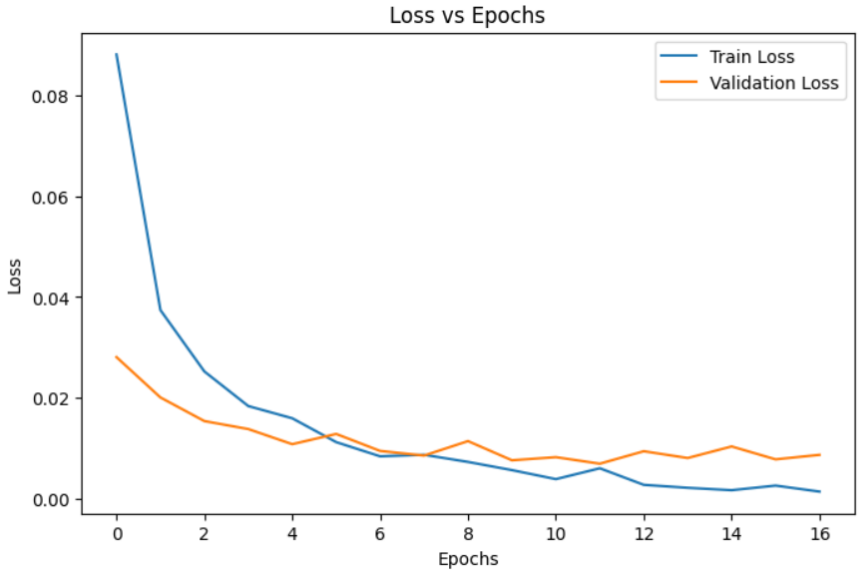
Key Findings

- The model achieved high accuracy and AUC on the test set, indicating effective classification.
- Most influential features identified by the model include **Hours_Studied, Attendance, Previous_Scores, Motivation_Level, and Teacher_Quality**.

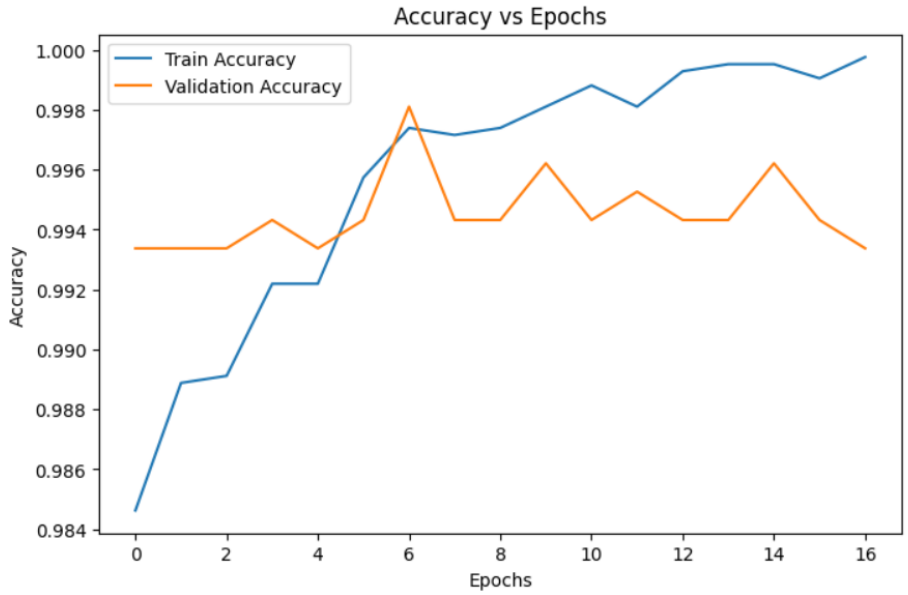
6. Result Visualization and Interpretation

6.1 Training Performance

Loss vs Epoch Chart:



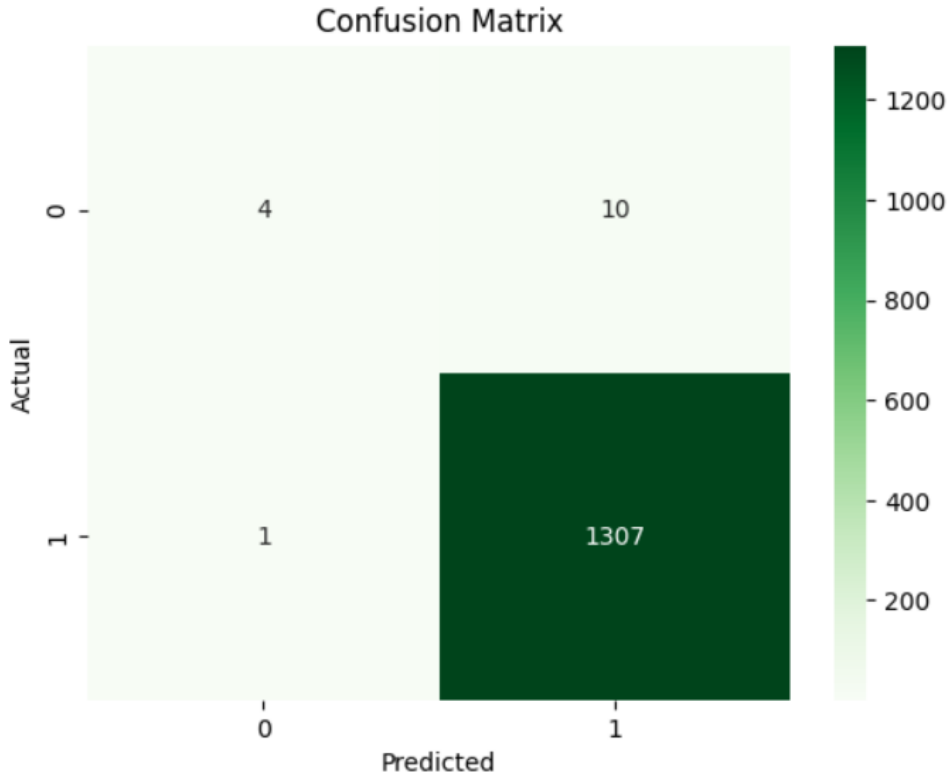
Accuracy vs Epoch Chart:



- Validation accuracy stabilized around 88%.

6.2 Classification Performance

Confusion Matrix Analysis:



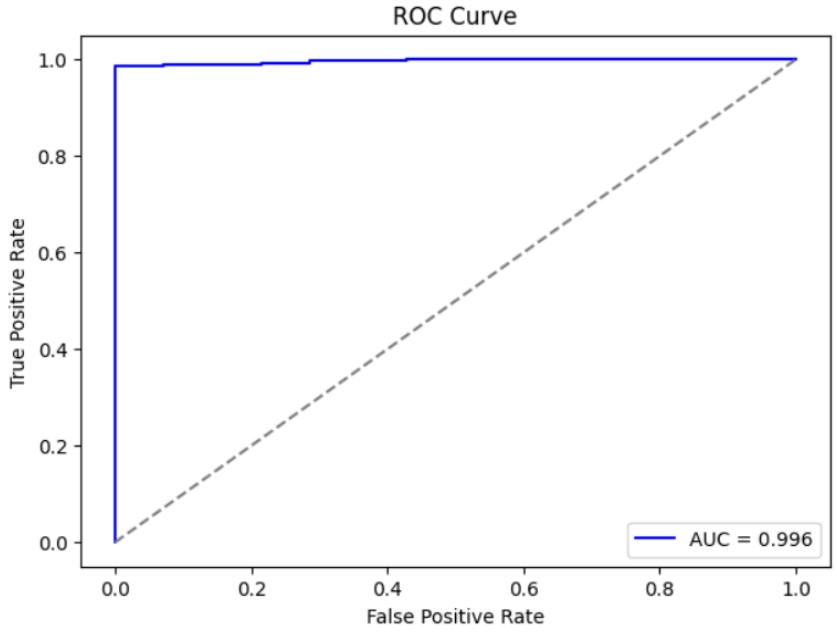
..

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.29      0.42        14
           1       0.99      1.00      1.00      1308

    accuracy                           0.99      1322
   macro avg       0.90      0.64      0.71      1322
weighted avg       0.99      0.99      0.99      1322
```

6.3 ROC Curve Analysis



Model Evaluation Complete. Test AUC: 0.996

Conclusion and Future Scope

This project analyzed and predicted student academic performance using the Student Performance Factors dataset. Through EDA, key factors influencing performance—such as Hours_Studied, Attendance, Motivation_Level, Teacher_Quality, and Parental_Involvement—were identified. A Multi-Layer Perceptron (MLP) deep learning model was developed, achieving high accuracy and AUC in classifying students as high or low performers. The results highlight the strong influence of academic habits, socio-environmental factors, and personal motivation on student success. Overall, the study demonstrates that data-driven approaches can effectively predict academic outcomes and provide actionable insights for educators and policymakers to improve learning strategies.

Future Scope

- Extend the model to predict exact exam scores (regression) rather than binary classification for more granular insights.
- Incorporate longitudinal data to study performance trends over time.
- Explore ensemble learning or advanced neural networks (e.g., LSTM, attention-based models) for improved predictive accuracy.
- Include additional features like mental health, diet, or classroom engagement metrics to enhance model robustness.
- Deploy a real-time student performance monitoring system to assist teachers in identifying at-risk students early.

7. References

- Laing, U. (2023). *Student Performance Factors Dataset*. Kaggle. https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- Chollet, F. (2018). *Deep Learning with Python*. Manning Publications.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

8. Appendix: Code Implementation

8.1 Data Loading and Preprocessing

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.metrics import confusion_matrix, classification_report, roc_curve, auc
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.callbacks import EarlyStopping

# Step 1: Load Dataset

df = pd.read_csv("StudentPerformanceFactors.csv")  # Rename to your file name
print("Shape of dataset:", df.shape)
df.head()
```

```
# Step 2: Data Understanding & Preprocessing

print("\nMissing values per column:\n", df.isnull().sum())

# Fill missing numeric values with mean and categorical with mode
for col in df.columns:
    if df[col].dtype == 'object':
        df[col].fillna(df[col].mode()[0], inplace=True)
    else:
        df[col].fillna(df[col].mean(), inplace=True)

# Create target: classify high vs low performance
df['Performance_Level'] = np.where(df['Exam_Score'] >= 60, 1, 0)  # 1 = High, 0 = Low
```