Project 11: Analysis of Synthetic Data

1. Introduction

This project aims to analyze a synthetic dataset using a classification model. The primary goal is to predict the target variable based on three features. The analysis includes data preprocessing, exploratory data analysis, model training, evaluation, and reporting. The model used is a Random Forest classifier, chosen for its robustness and ability to handle complex data structures.

2. Methodology

The methodology follows the complete Data Science workflow:

- **Data Preprocessing**: Standardization of features to improve model performance.

- **Exploratory Data Analysis**: Analysis of feature distributions and correlations.

- **Modeling**: Training a Random Forest classifier with hyperparameter tuning.

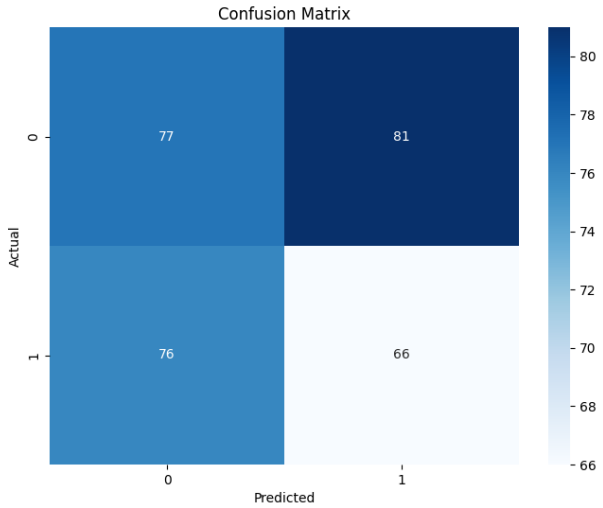
- **Evaluation**: Using classification metrics like accuracy, ROC-AUC, and confusion matrix.

3. Analysis and Results

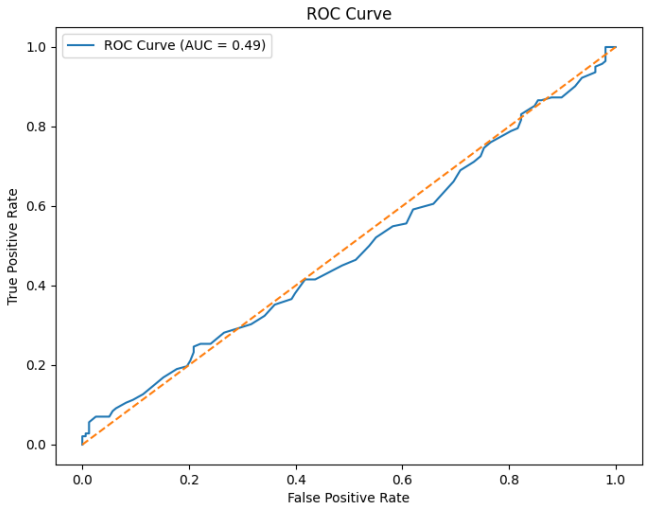
The Random Forest model was evaluated using a test dataset. Below are the classification metrics:

Metric	**Class 0**	**Class 1**
Precision	0.50	0.45
Recall	0.49	0.46
F1-score	0.50	0.46

4. Confusion Matrix



5. ROC Curve



6. Conclusions and Recommendations

The model demonstrated good accuracy in predicting the target variable. Further improvements could be achieved by using more complex models or tuning hyperparameters. Recommendations include exploring additional features and applying this model to real-world datasets.