Airline Passenger Satisfaction Classification Report

1.Introduction

The objective of this project was to classify airline passenger satisfaction using various features from a dataset obtained from Kaggle. The dataset contains multiple variables such as flight distance, inflight service ratings, and demographic information. By building and evaluating different machine learning models, we aimed to identify the most effective model for predicting passenger satisfaction.

2.Data Exploration and Preprocessing

Initial Exploration

1. **Data Loading and Overview**:

   The dataset was loaded using pandas, and an initial overview of the data was obtained. This included checking the shape of the dataset, identifying duplicate entries, and inspecting missing values.

2. **Handling Missing Values**:

   The dataset had missing values in the arrival_delay_in_minutes column. These were filled with the median value of the column to ensure data integrity.

3. **Data Visualization**:

   Histograms were plotted to understand the distribution of numerical variables.

   Count plots were created for categorical variables to visualize their frequency distribution.

   Box plots were used to detect outliers in numerical columns.

Data Encoding and Feature Engineering

1. **Categorical Encoding**:

   Categorical variables were encoded using One-Hot Encoding to convert them into numerical format suitable for machine learning algorithms.

2. **Correlation Analysis**:

   A correlation matrix was plotted to understand the relationships between different features and to detect multicollinearity.

3.Model Building

Three different models were built and evaluated:

1. **Logistic Regression**:

   A logistic regression model was trained using the processed dataset.

   The model achieved an accuracy score of 0.87.

2. **Decision Tree**:

A decision tree classifier was used as an alternative model.

The decision tree model also achieved an accuracy score of 0.94.

3. **Random Forest**:

A random forest classifier was trained, which is an ensemble method combining multiple decision trees.

The random forest model achieved the highest accuracy score among the three models, 0.96.

4.Model Evaluation

1. **Confusion Matrix**:

Confusion matrices were plotted for each model to visualize the true positives, true negatives, false positives, and false negatives.

These matrices helped in understanding the performance of each model in detail.

2. **ROC Curve and AUC**:

The ROC curves for all three models were plotted.

The Area Under the Curve (AUC) was calculated for each model:

-Logistic Regression: AUC = 0.92

-Decision Tree: AUC = 0.94

-Random Forest: AUC = 0.99

5.Conclusion

The project successfully built and evaluated multiple machine learning models to classify airline passenger satisfaction. The Random Forest classifier emerged as the best model with the highest accuracy and AUC score. The models' performances indicate that they can effectively predict passenger satisfaction based on the provided features.

Future work could involve further hyperparameter tuning, exploring additional advanced models, and potentially integrating more features to enhance prediction accuracy. Additionally, real-world deployment would require considerations for model interpretability and continuous monitoring to ensure sustained performance.