Report

Data Preprocessing

Imputation:
The 'Arrival Delay in Minutes' column with missing values is imputed using the median value. Imputation is a common technique to handle missing data, ensuring a complete dataset for further analysis.

One-Hot Encoding:
Categorical features ('Gender', 'Traveler Type', 'Type of Travel', 'Class') are transformed using one-hot encoding. This process converts categorical variables into binary vectors, allowing machine learning models to interpret and utilize these features effectively.

Outlier Removal:
Outliers in numeric features are identified and capped using the Interquartile Range (IQR) method. This method helps maintain data integrity by addressing extreme values that might skew the model's performance.

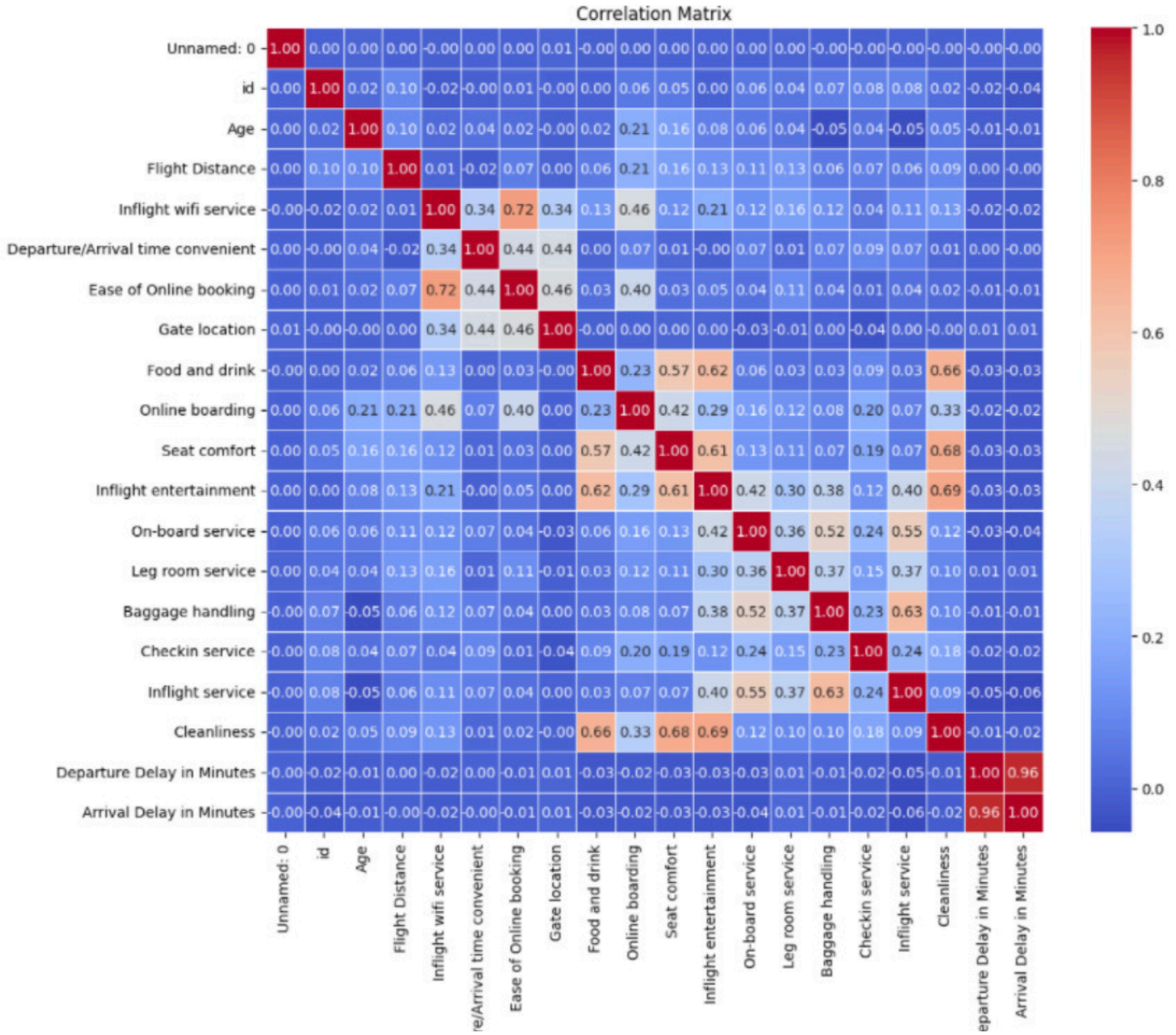
Exploratory Data Analysis (EDA)
Data Loading and Initial Exploration:

A comprehensive overview of the dataset is presented, including the first few rows, summary statistics, and data type information. This step is crucial for understanding the dataset's structure and characteristics.

Missing Values Analysis:
A thorough examination reveals no missing values in the dataset, ensuring a complete and reliable dataset for subsequent analysis.

Correlation Analysis:

A heat-map visualising the correlation matrix of numeric features provides insights into feature relationships. This aids in understanding how different features interact with each other, assisting in feature selection and interpretation.

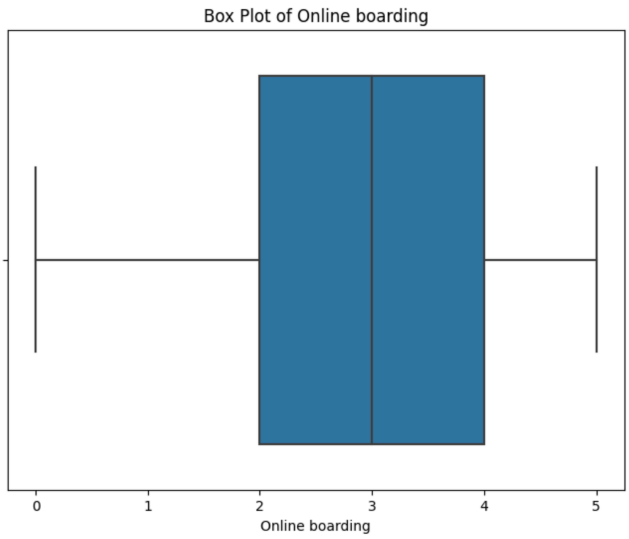


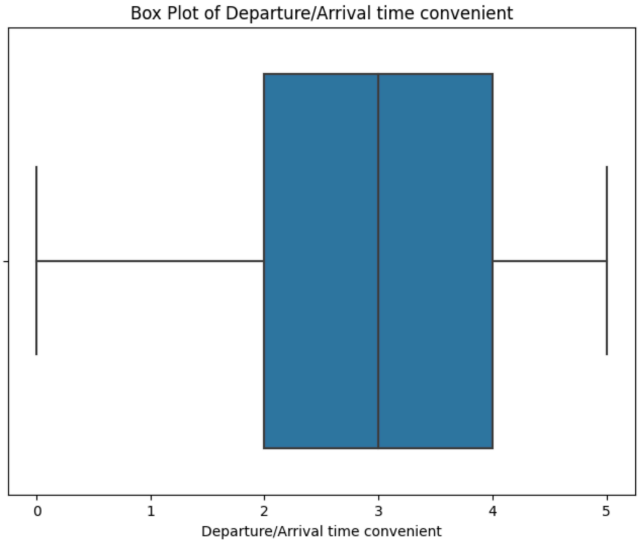
Target Variable Distribution:
Visualization of the distribution of the target variable ('contentment') provides insights into the balance or imbalance of classes. This is essential for understanding the dataset's inherent characteristics.

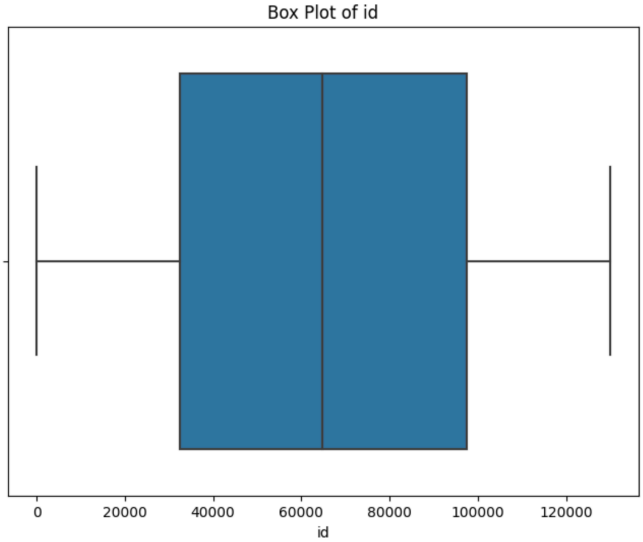
Feature Distribution:

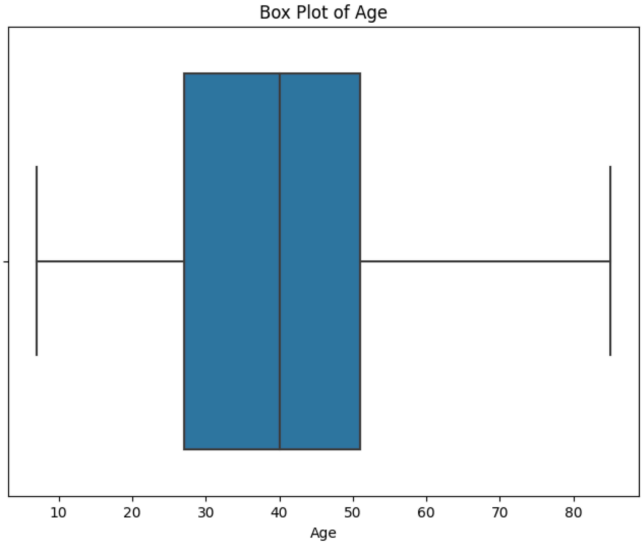
A histogram displays the distribution of the 'Arrival Delay in Minutes' feature, offering insights into its central tendency and spread. Understanding feature distributions is crucial for choosing appropriate preprocessing techniques.

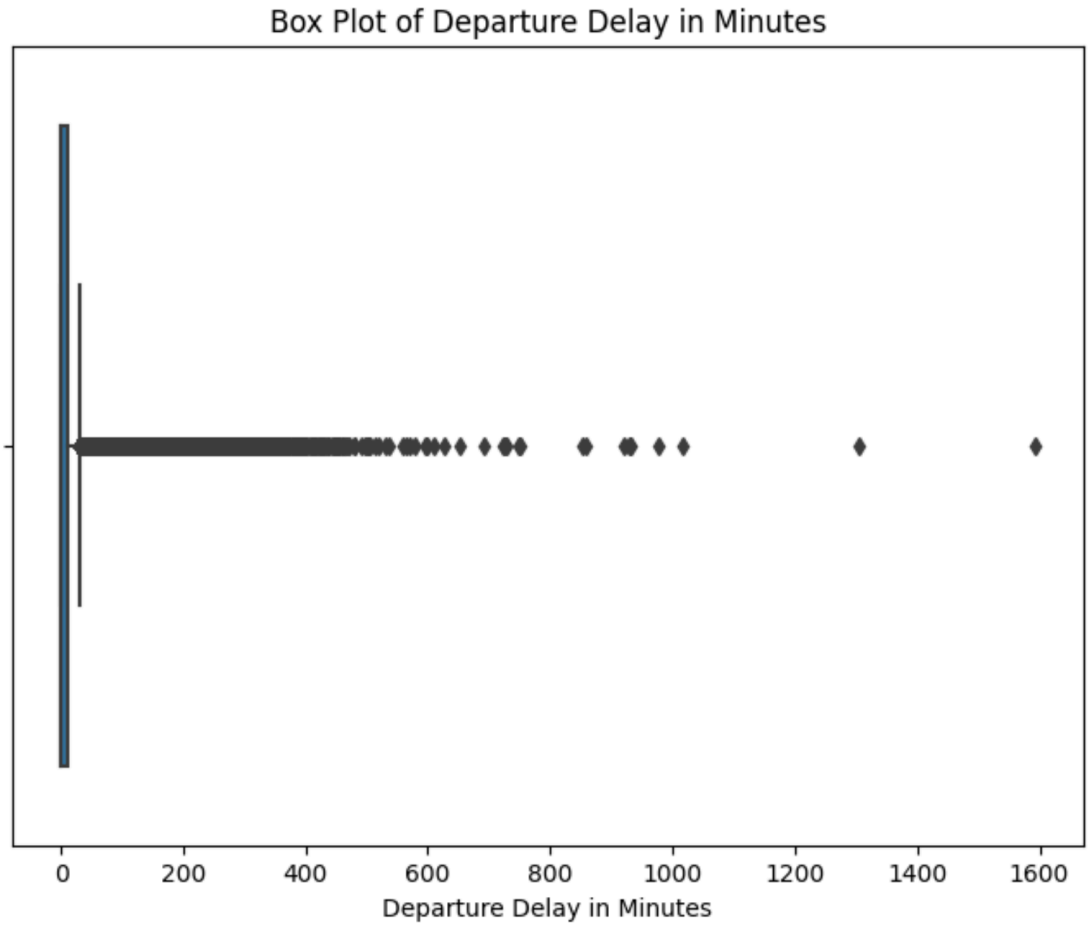
Box Plots for Numeric Features:
Box plots are employed to illustrate the distribution of numeric features, facilitating the identification of potential outliers. Outliers can significantly impact model performance and thus need careful consideration.

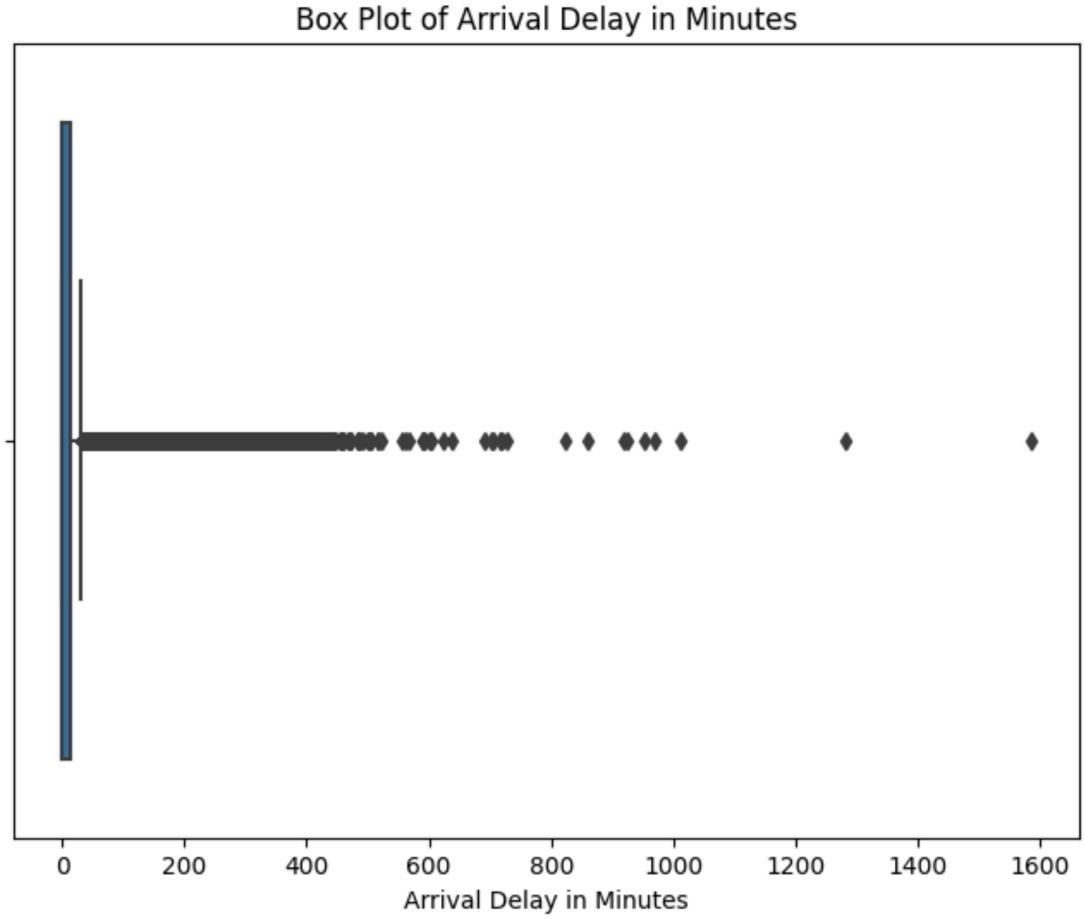






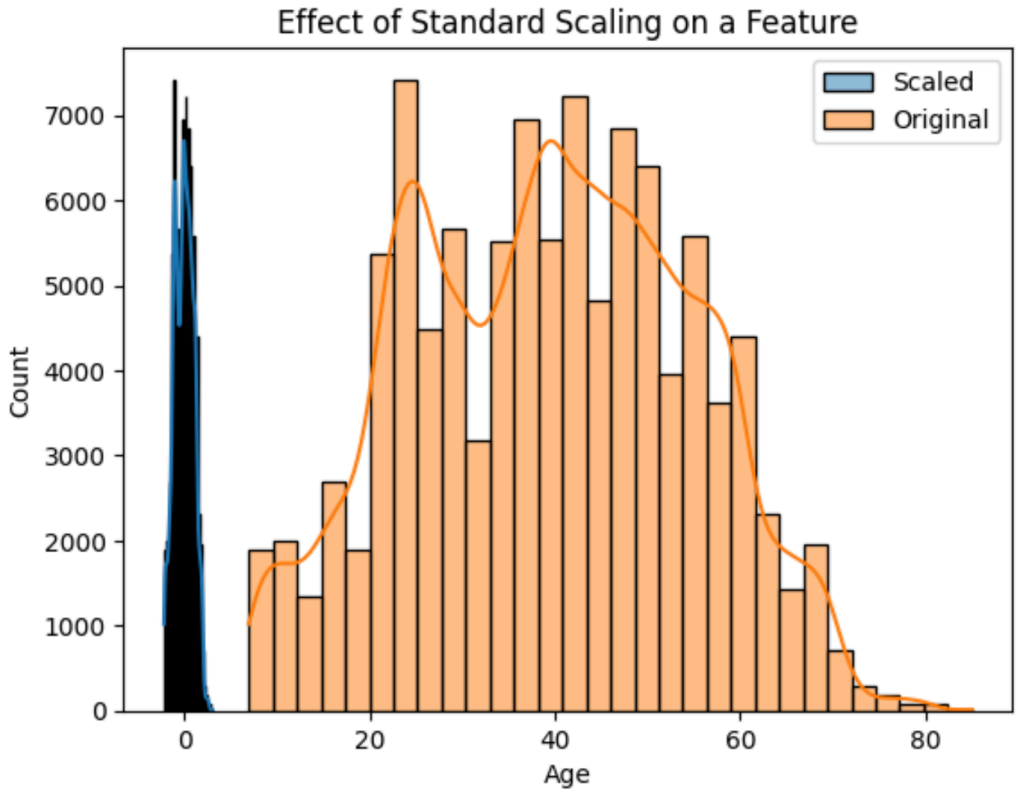



Box Plot of Departure Delay in Minutes
0
200
400
600
800
1000
1200
1400
1600
Departure Delay in Minutes


Box Plot of Arrival Delay in Minutes
0
200
400
600
800
1000
1200
1400
1600
Arrival Delay in Minutes

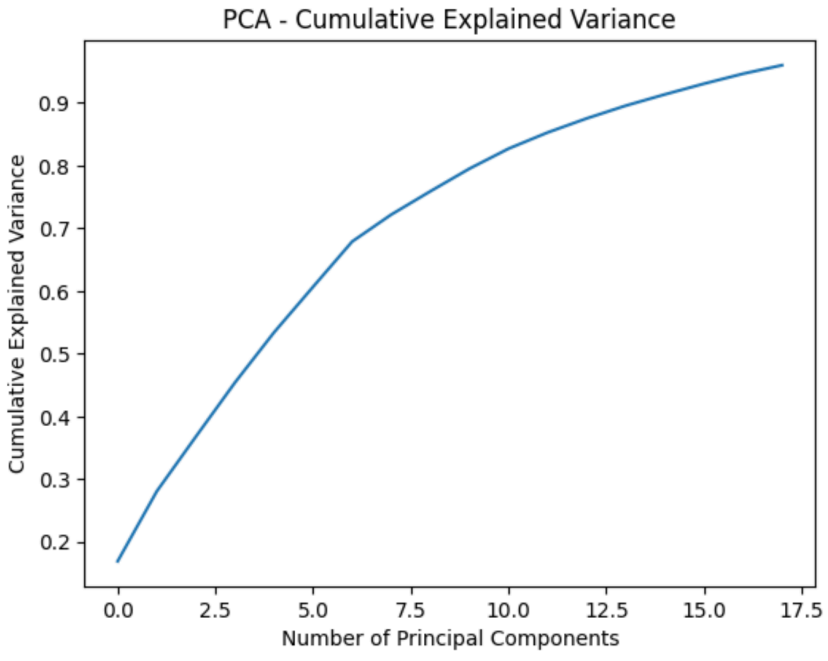
Data Splitting and Preprocessing

The dataset is split into training and validation sets to assess model performance effectively. Standard scaling is applied to the features, ensuring that numeric variables are on a similar scale, preventing certain features from dominating others.



Dimensionality Reduction (PCA):

Principal Component Analysis (PCA) is utilized for dimensionality reduction while retaining 95% of the variance. This reduces the complexity of the dataset and helps mitigate issues related to the curse of dimensionality.



Modeling

Support Vector Machine (SVM):
A Support Vector Machine with an RBF kernel and specified hyperparameters is employed. SVMs are known for their versatility in handling both linear and non-linear relationships in data.

Nu-Support Vector Machine (NuSVM):
A variation of the SVM algorithm, the Nu-SVM, is used with specific hyperparameters. Nu-SVM can be advantageous in scenarios where the exact proportion of support vectors is unknown.

Multi-Layer Perceptron (MLP):
A Multi-Layer Perceptron with a specific architecture, including two hidden layers and ReLU activation, is chosen. MLPs are a type of artificial neural network known for their ability to capture complex relationships in data.

Boosting using AdaBoost with Decision Trees:
An AdaBoost ensemble with decision trees as base models is implemented. AdaBoost is an ensemble learning technique that combines weak learners to create a robust predictive model.

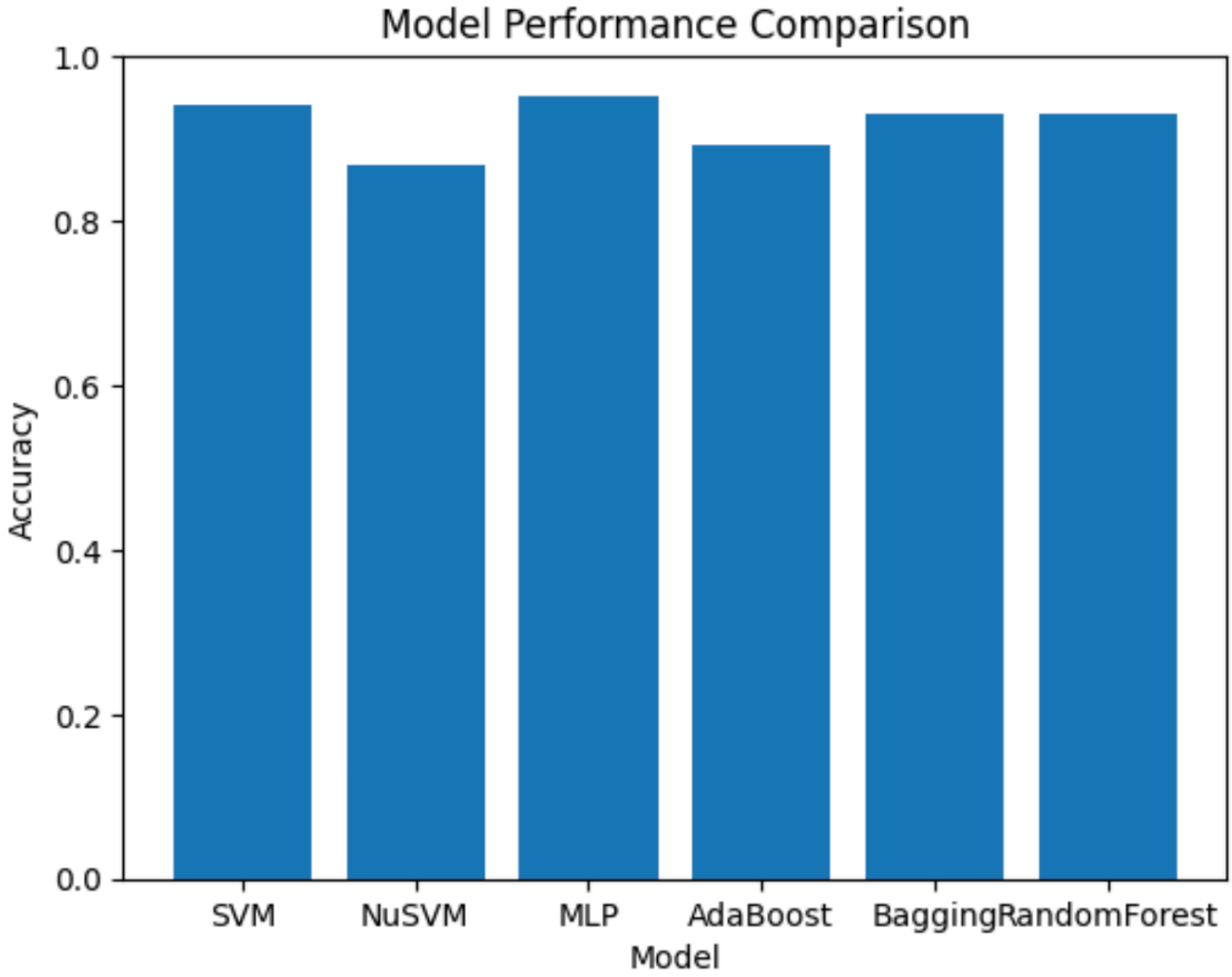
Bagging using Decision Trees:
A Bagging ensemble with decision trees as base models is employed. Bagging helps reduce overfitting and variance by aggregating predictions from multiple models.

Random Forest:
A Random Forest model is utilized, consisting of an ensemble of decision trees. Random Forests are powerful and robust due to their ability to reduce overfitting and handle complex relationships.

Neural Network Model:
A Neural Network model with a specific architecture, including two hidden layers and ReLU activation, is constructed using the TensorFlow Keras library. Neural networks are known for their capability to learn intricate patterns in data.

Model Evaluation

All models are evaluated on a validation set, and their accuracies are recorded. Model evaluation involves assessing how well each model generalizes to unseen data, providing insights into their performance.



Best Model Selection

The model with the highest accuracy on the validation set is selected as the best model. Accuracy serves as a primary metric for model selection in this context.

Final Predictions

The best model (MLP) is employed to make predictions on the test set. Predictions are converted to binary values, aligning with the binary nature of the target variable.

Report By:

IMT2021014 Karan Raj Pandey
IMT2021050 Anuj Arora
IMT2021092 Ayush Singh