**NAME: MAFFO YOKIE BEUTCHA REGINE**

**Report on Titanic Dataset Analysis and Classification Model**

**Introduction**

This report describes a comprehensive data processing and machine learning pipeline applied to the Titanic dataset to predict passenger survival. This pipeline includes data cleaning, outlier handling, normalization, feature engineering, encoding, model training, and evaluation using a Random Forest classifier. The goal of this analysis is to improve the prediction accuracy of survival on the Titanic based on several features.

1. **Importing Libraries**

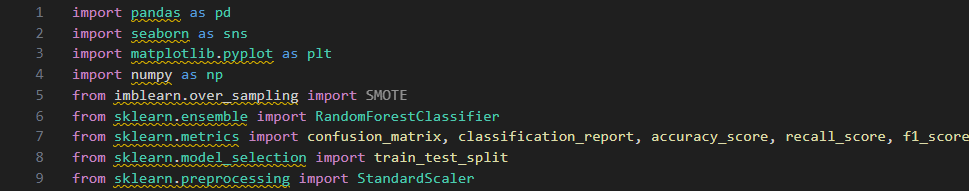


Figure 1importing libraries

* **pandas**: For data manipulation and analysis.
* **seaborn** and **matplotlib.pyplot**: For data visualization.
* **numpy**: For numerical operations.
* **imblearn.over\_sampling.SMOTE**: For addressing class imbalance (not utilized in this implementation).
* **sklearn.ensemble.RandomForestClassifier**: For creating the classification model.
* **sklearn.metrics**: For evaluating model performance.
* **sklearn.model\_selection.train\_test\_split**: For splitting the dataset into training and testing sets.
* **sklearn.preprocessing.StandardScaler**: For standardizing feature values.

1. **Data Cleaning Function**

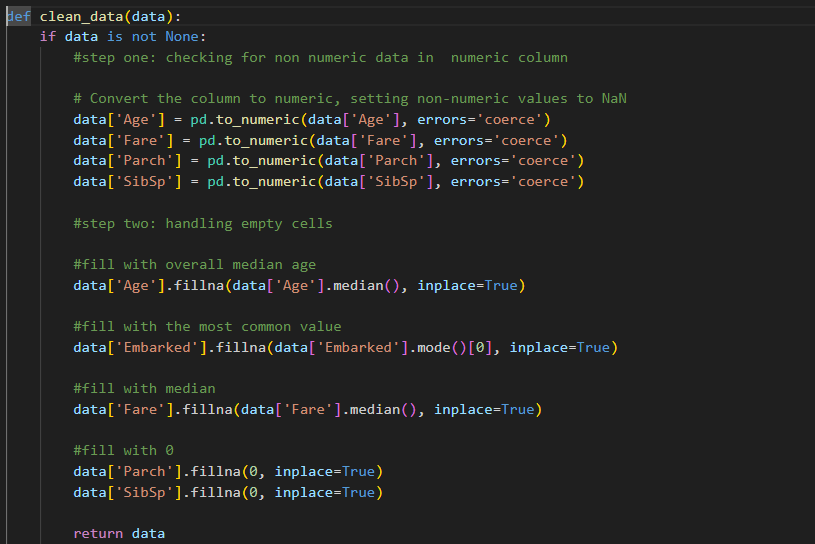


Figure 2 2. Data Cleaning Function

* **clean\_data(data)**: Cleans the dataset by:
  + Converting columns to numeric types, replacing non-numeric values with NaN.
  + Filling missing values using median for 'Age' and 'Fare', mode for 'Embarked', and zero for 'Parch' and 'SibSp'.

**3. Data Collection**



Figure 3Data Collection

* Loads the Titanic dataset from a CSV file into a DataFrame.

**4. Data Cleaning**



Figure 4Data Cleaning

* Invokes the clean\_data function to preprocess the dataset.

**5. Outlier Detection and Visualization**

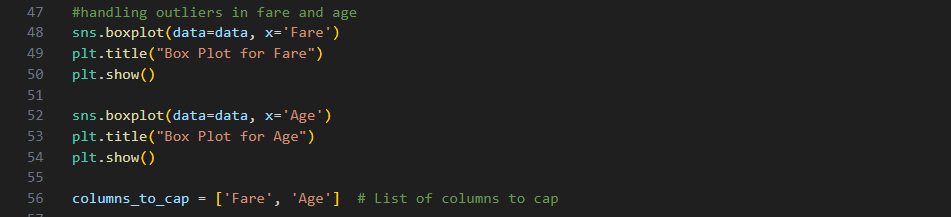


Figure 5 Outlier Detection and Visualization

* Creates box plots for 'Fare' and 'Age' to visually inspect for outliers.

**6. Removing Outliers**

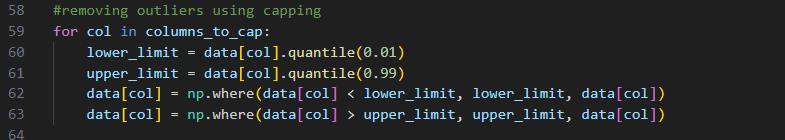


Figure 6. Removing Outliers

* Caps outliers in 'Fare' and 'Age' at the 1st and 99th percentiles to reduce their impact on the analysis.

**7. Data Normalization**

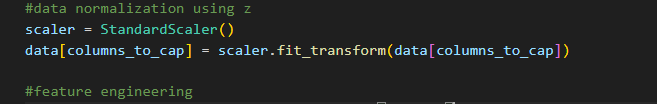


Figure 7 Data Normalization

* Standardizes the 'Fare' and 'Age' features to have a mean of 0 and a standard deviation of 1.

**8. Feature Engineering**

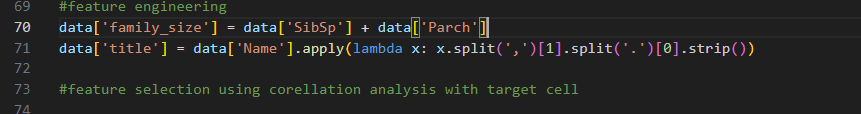


Figure 8Feature Engineering

* Creates new features:
  + family\_size: Sum of siblings/spouses and parents/children.
  + title: Extracted from passenger names to capture social status.

**9. Encoding Categorical Features**

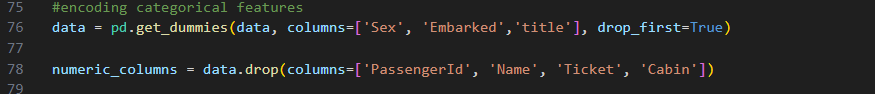


Figure 9Encoding Categorical Features

* Converts categorical variables into dummy variables for model training.

**10. Correlation Analysis**

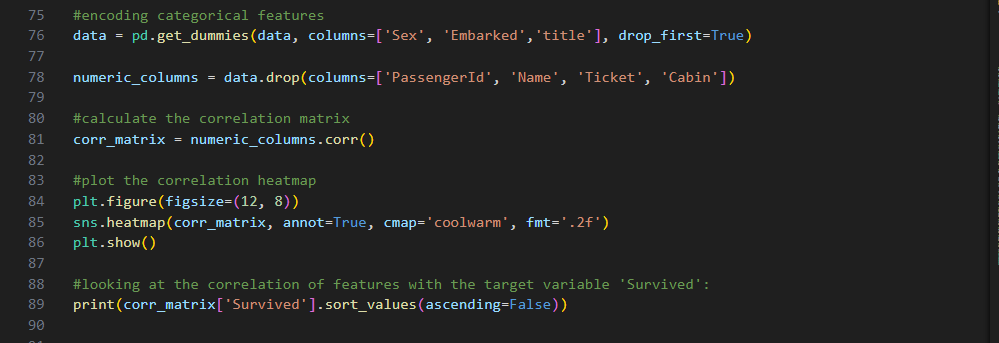


Figure 10Correlation Analysis

* Computes and visualizes the correlation matrix to identify relationships between features and the target variable ('Survived').

**11. Model Building**

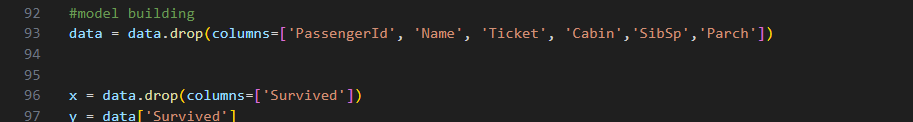


Figure 11Model Building

* Prepares the dataset for model training by dropping unnecessary columns and splitting the data into training and testing sets.

**12. Training the Random Forest Model and evaluating the model**

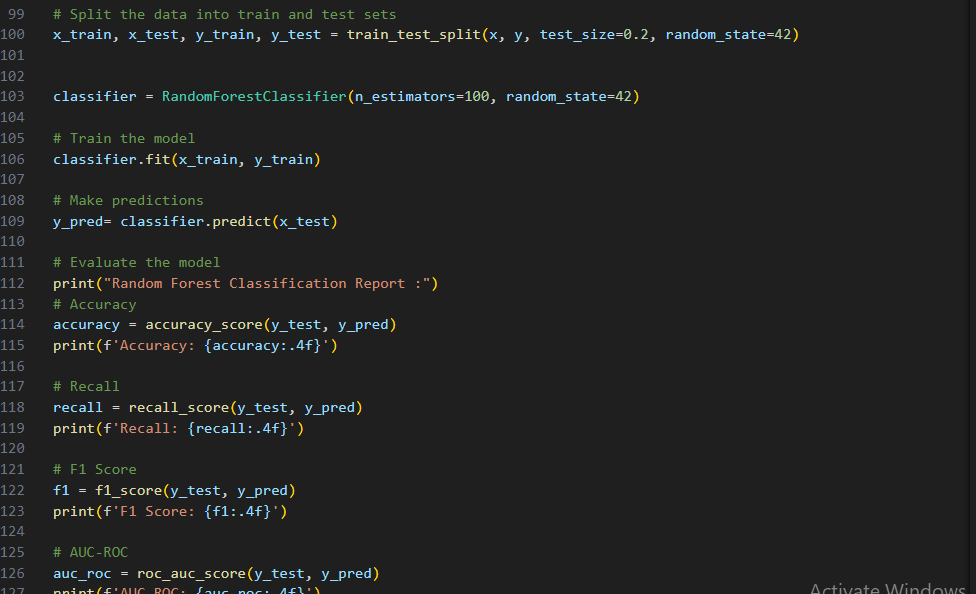
* 

Figure 12Training the Random Forest Model and evaluating the model

* Initializes and trains a Random Forest classifier using the training data.

**13. Making Predictions**



Figure 13Making Predictions

* Uses the trained model to predict outcomes on the test set.



* Evaluates the model's performance using metrics such as accuracy, recall, F1 score, and AUC-ROC, providing insights into its predictive power.

**15. Confusion Matrix Visualization**

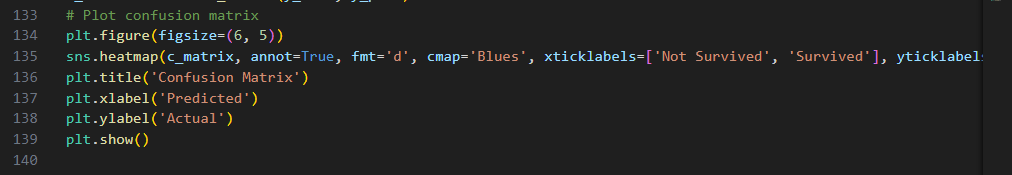


Figure 14Confusion Matrix Visualization

* Computes the confusion matrix and visualizes it using a heatmap to show true vs. predicted classifications, aiding in understanding model performance