Titanic Dataset Analysis, Visualization, and Modeling Report

Introduction:

The Titanic dataset is a classic dataset widely used for exploring machine learning techniques. This report covers data exploration, visualization, preprocessing, and model building to predict passenger survival on the Titanic.

Data Exploration and Cleaning:

Missing Values:

Checked for missing values in the dataset.

Filled missing values in the 'Age' column with the mean value.

Filled missing values in the 'Embarked' column with the most frequent value.

Dropped the 'Cabin' column due to a large number of missing values.

Outlier Detection:

Identified outliers using box plots for numerical columns.

Used the interquartile range (IQR) method to remove outliers from the 'Fare' column.

Data Visualization:

Passenger Analysis:

Explored the distribution of passenger ages using a histogram.

Analyzed passenger class distribution using a count plot.

Investigated common embarkation points using a count plot.

Explored the distribution of fares paid by passengers using a histogram.

Examined the gender distribution in the dataset using a count plot.

Survival Analysis:

Visualized the distribution of survived and non-survived passengers using a count plot.

Explored how the survival rate varies across different passenger classes using a bar plot.

Investigated the correlation between survival and gender using a bar plot.

Explored patterns in survival based on passenger age using a histogram.

Analyzed the impact of siblings/spouses (SibSp) and parents/children (Parch) on survival using bar plots.

Examined the correlation between survival and embarkation point using a bar plot.

Relationships and Correlations:

Explored the correlation between fare (Fare) and passenger class (Pclass) using a box plot.

Analyzed patterns in the distribution of fares based on embarkation point (Embarked) using a box plot.

Investigated how the number of siblings/spouses (SibSp) and parents/children (Parch) vary across different passenger classes (Pclass) using box plots.

Explored the correlation between passenger age (Age) and fare (Fare) using a scatter plot.

Additional Analysis:

Extracted titles from the 'Name' column and visualized their distribution using a count plot.

Visualized the first 20 columns of the dataset using bar plots.

Plotted a bar plot of age distribution based on gender.

Model Building:

Preprocessing:

Encoded categorical variables ('Sex' and 'Embarked') using Label Encoding.

Split the dataset into features (X\_train) and target (y\_train).

Model Training:

Trained a Logistic Regression model on the training data.

Evaluated the model on the test data, showing accuracy and classification report.

Model Tuning:

Utilized GridSearchCV to tune hyperparameters for Logistic Regression.

Identified the best parameters and evaluated the tuned model.

Random Forest Model:

Applied a Random Forest Classifier to the dataset.

Evaluated the Random Forest model and displayed accuracy and classification report.

Confusion Matrix:

Visualized the confusion matrix using a heatmap.

Conclusion:

The analysis, visualization, and modeling provide valuable insights into the Titanic dataset. The predictive models, including Logistic Regression and Random Forest, demonstrate reasonable accuracy. Further exploration and fine-tuning of models could enhance predictive performance.