Shrey Shah
(20009523)

AAI – 551 FINAL PROJECT REPORT

"Data Analysis and Modeling on the Titanic Dataset"

Code description :

The code begins by importing several libraries including pandas, seaborn, matplotlib, sklearn, and warnings. It then reads in the Titanic dataset using pandas, and explores the data using the head(), tail(), and describe() functions. The code checks for null values in the dataset using isnull().sum() and removes rows with null values in the Embarked column. The Age column has null values, which are filled with the mean value of the column using fillna().

The code then defines a function called custom_summary() that calculates descriptive statistics for several columns in the dataset: PassengerId, Survived, Pclass, Age, SibSp, Parch, and Fare. The function also includes custom comments for identifying skew and outliers in the data. To identify skew, the function categorizes the skew of each column into several categories based on the value of the skewness statistic. To identify outliers, the function uses the interquartile range (IQR) to define upper and lower bounds, and checks if any values in the column fall outside of these bounds.

After calculating the descriptive statistics, the code performs one-hot encoding on the Sex column, replacing the values with 0s and 1s. It then creates a new column called FamilySize that is the sum of the SibSp and Parch columns. The code then creates a new dataframe with the FamilySize and Sex columns as well as the Survived column, which is the target variable.

The code then splits the data into training and test sets using the train_test_split() function from sklearn. It fits a logistic regression model to the training data and makes predictions on the test data, evaluating the predictions using the accuracy_score() function. The code also fits a linear regression model to the training data and makes predictions on the test data, evaluating the predictions using the mean_squared_error() function.

In summary, this code performs data exploration, cleaning, and preparation on the Titanic dataset, and fits and evaluates two different machine learning models on the prepared data. The models used are logistic regression and linear regression. The logistic regression model is evaluated using accuracy, and the linear regression model is evaluated using mean squared error.

GITHUB Link :https://github.com/shreyshah6699/AAI-551