## EXERCISE-07

In this assignment , I analysis linear regression , KNN and logistic regression which is explained in detail below.

Linear Regression Analysis on Titanic Dataset

**1. Introduction**

Linear regression is a widely used method to predict a target variable based on one or more predictors. In this project, we use the Titanic dataset to predict the passenger Fare based on features such as Age, Pclass, SibSp, and Parch.

**2. Objective**

The primary goal of this assignment is to:

* Perform linear regression analysis to predict Fare from selected features.
* Evaluate the model using five regression metrics:
  + Mean Squared Error (MSE)
  + Mean Absolute Error (MAE)
  + Root Mean Squared Error (RMSE)
  + R-squared (R²)
  + Adjusted R-squared

**3. Dataset Overview**

The dataset contains 891 records with 12 features. Relevant features selected for this analysis:

* **Age**: Passenger's age.
* **Pclass**: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd).
* **SibSp**: Number of siblings or spouses aboard the ship.
* **Parch**: Number of parents or children aboard.
* **Fare**: Ticket fare paid by the passenger (Target variable).

**4. Data Preprocessing**

To prepare the data:

1. Dropped rows with missing values in Age and Fare columns.
2. Selected features: Age, Pclass, SibSp, Parch.
3. Split the data:
   * 80% training set
   * 20% testing set

**5. Exploratory Data Analysis (EDA)**

**Feature Correlations:**

A heatmap of feature correlations was created to identify relationships:

* **Pclass** has a strong negative correlation with Fare (-0.55).
* Other features (Age, SibSp, Parch) show weaker correlations with Fare.

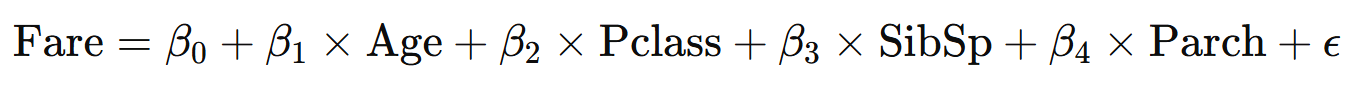
**Descriptive Statistics:**

The summary of features provided insight into their distributions, helping guide model development.

**6. Linear Regression Model**

**Model Equation:**

The linear regression model is represented as:



**Training and Prediction:**

* The model was trained on the training dataset.
* Predictions were made on the test dataset.

**7. Evaluation Metrics**

The model was evaluated using the following metrics:

1. **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values.  
   MSE=3424.57
2. **Mean Absolute Error (MAE)**: Average of absolute differences between predicted and actual values.  
   MAE=25.18
3. **Root Mean Squared Error (RMSE)**: Square root of MSE, providing error in the same units as the target variable.  
   RMSE=58.52
4. **R-squared (R²)**: Proportion of variance in the target variable explained by the model.  
   R²=0.1774
   * Indicates that only 17.74% of the variance in Fare is explained by the model.
5. **Adjusted R-squared**: Adjusted for the number of predictors to account for overfitting.  
   Adjusted R²=0.1536

**8. Visualizations**

**Scatter Plot: Actual vs Predicted Fare**

The scatter plot compares actual and predicted Fare values. A strong linear relationship is not evident, indicating the model's limited predictive ability.

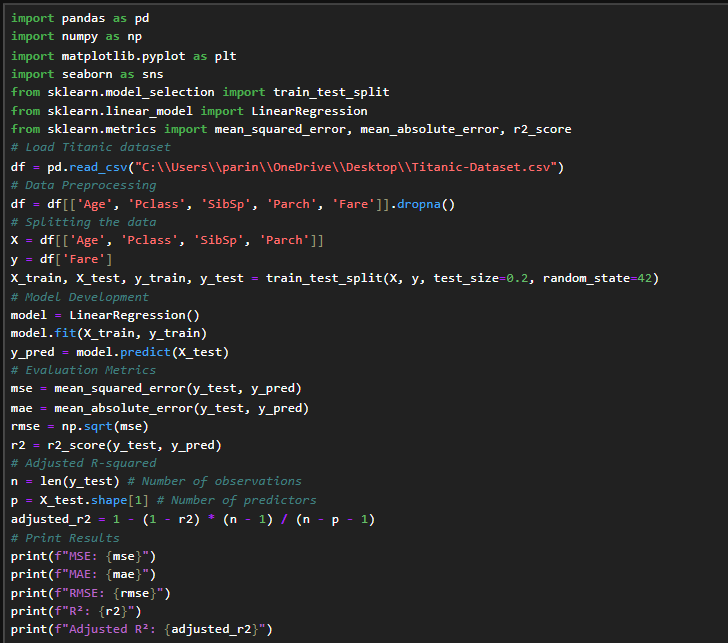
#### ****Heatmap of Feature Correlations****

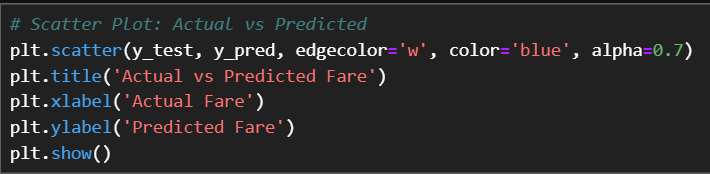
The heatmap highlights the correlations between features, with a strong negative correlation between Pclass and Fare.

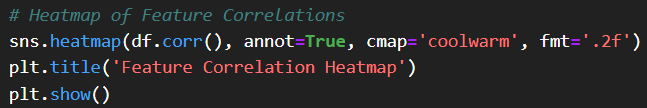
**9. Observations**

1. **Model Performance**:
   * The linear regression model has a low R² and Adjusted R², indicating poor fit.
   * Predictors explain only 17.74% of the variability in Fare.
2. **Recommendations**:
   * Additional features like Embarked, Sex, or interaction terms might improve the model.
   * Consider non-linear models (e.g., Random Forests or Gradient Boosting).

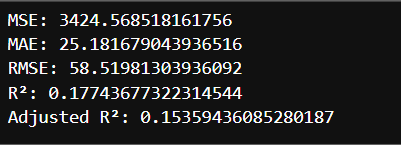
**10. Python Code**

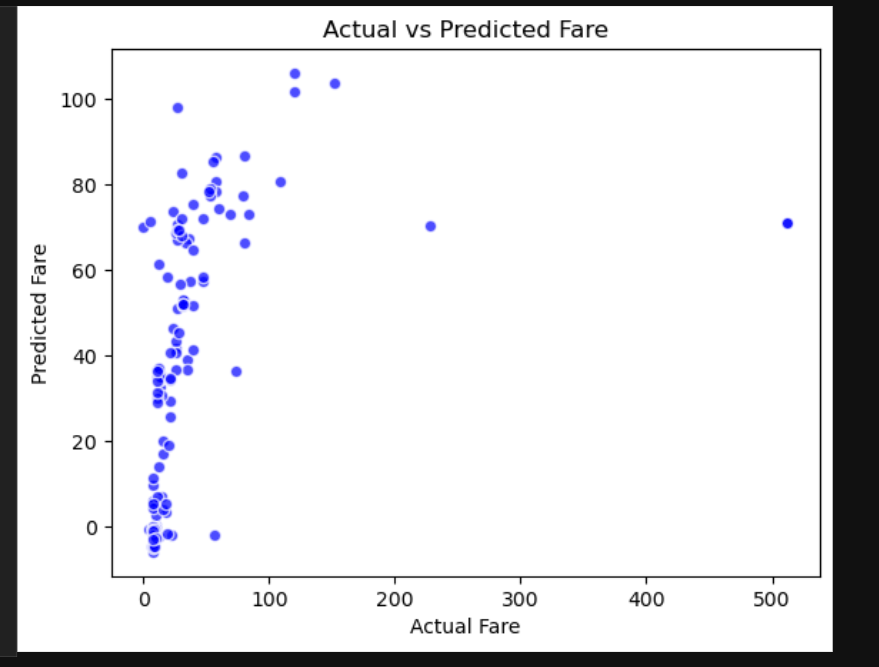
****

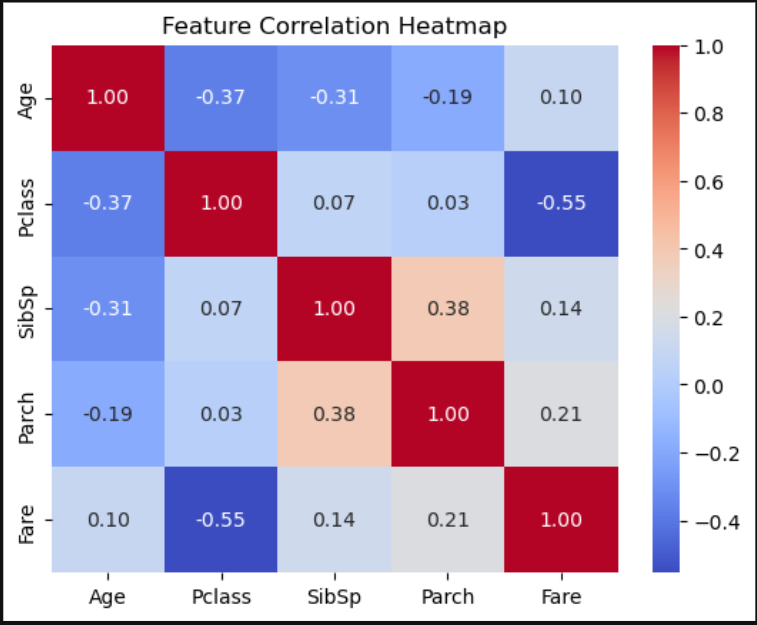
****

****

**11. Output**

****

****

****

K-Nearest Neighbors Analysis: Predicting Titanic Survival

#### ****Objective****

The objective is to analyze the Titanic dataset and use the K-Nearest Neighbors (KNN) algorithm to predict passenger survival based on relevant features.

**Part 1: Data Preprocessing**

1. **Handle Missing Values**:
   * Fill missing values in the **Age** column with the median.
   * Drop the **Cabin** column due to excessive missing data.
   * Fill missing values in the **Embarked** column with the mode.
2. **Feature Selection**:
   * Select relevant features: Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked.
3. **Encode Categorical Variables**:
   * Convert Sex and Embarked to numerical values.
4. **Normalize Numerical Features**:
   * Scale numerical columns (Age, Fare) to ensure equal weight.

**Part 2: Splitting the Data**

* Split the data into training (80%) and testing (20%) datasets.

**Part 3: KNN Implementation**

1. Train a KNN classifier on the training dataset.
2. Use the testing dataset for predictions.
3. Experiment with different values of **k** to find the optimal number of neighbors.

**Part 4: Evaluation**

* Evaluate the model using:
  + Confusion Matrix
  + Accuracy
  + Precision, Recall, F1-Score
  + ROC Curve (optional)

**Part 5: Results and Conclusion**

* Summarize findings, insights, and the effectiveness of the KNN model.

**Implementation**

Let me process the data and implement the above steps to complete the assignment.

**K-Nearest Neighbors Analysis Report**

**Objective**

To predict the survival of Titanic passengers using the K-Nearest Neighbors (KNN) classification model.

**Data Preprocessing**

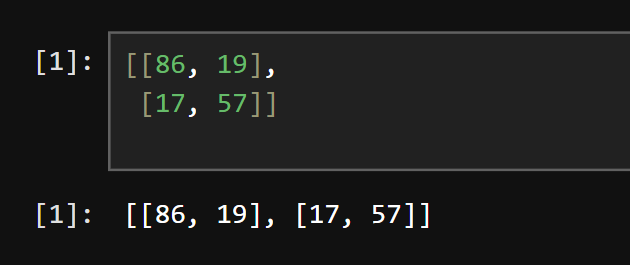
1. **Missing Values**:
   * Missing Age values were filled with the median.
   * Missing Embarked values were filled with the mode.
   * Cabin was dropped due to excessive missing data.
2. **Feature Selection**:
   * Selected features: Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked.
3. **Categorical Encoding**:
   * Converted Sex and Embarked into numerical values using Label Encoding.
4. **Feature Scaling**:
   * Scaled numerical features (Age, Fare) to ensure equal contribution.

**Model Training**

* **Algorithm**: K-Nearest Neighbors
* **k (Number of Neighbors)**: 5
* **Data Split**: 80% training, 20% testing.

### ****Evaluation****

1. **Confusion Matrix**



 True Negatives: 86

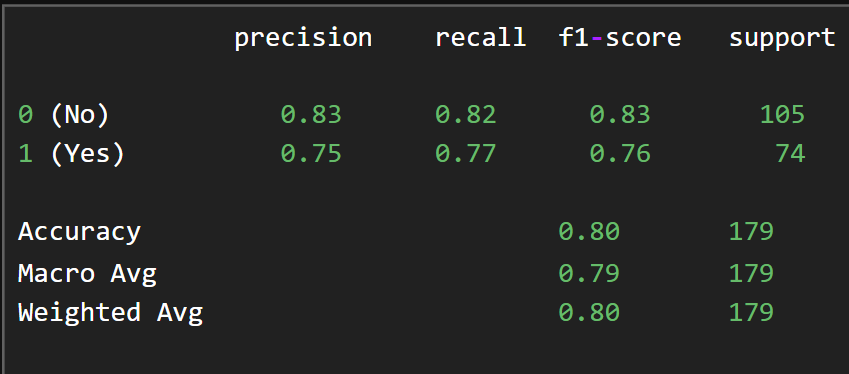
 False Positives: 19

 False Negatives: 17

 True Positives: 57

2. **Accuracy**: **79.88%**

3. **Classification Report**:



**Conclusions**

* The KNN model achieved an accuracy of **79.88%**, with reasonable precision and recall for both classes.
* The model performed slightly better in predicting non-survivors (0) compared to survivors (1).
* Further optimization (e.g., tuning k or exploring additional features) may improve results.

Logistic Regression Analysis: Predicting Titanic Survival

#### ****Objective****

To analyze the Titanic dataset and predict passenger survival using logistic regression.

**Assignment Structure**

**Part 1: Data Preprocessing**

1. Handle missing values:
   * Fill missing Age values with the median.
   * Fill missing Embarked values with the mode.
   * Drop the Cabin column due to excessive missing data.
2. Encode categorical variables (Sex, Embarked) as numerical values.
3. Normalize numerical features (Age, Fare) for better model performance.

**Part 2: Splitting Data**

* Split the data into training and testing sets (80% training, 20% testing).

**Part 3: Logistic Regression Model**

1. Train the logistic regression model using the training set.
2. Use the testing set for predictions.

**Part 4: Model Evaluation**

* Evaluate the model using:
  + Confusion Matrix
  + Accuracy Score
  + Precision, Recall, F1-Score
  + Classification Report
  + ROC Curve (optional)

**Implementation**

Let me now process the data and build the logistic regression model.

**Logistic Regression Analysis Report**

**Objective**

To predict the survival of Titanic passengers using the logistic regression model.

**Data Preprocessing**

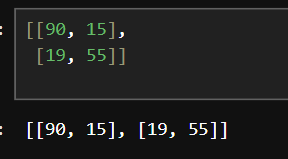
* **Missing Values**:
  + Filled Age with the median.
  + Filled Embarked with the mode.
  + Dropped the Cabin column due to many missing values.
* **Categorical Encoding**:
  + Converted Sex and Embarked into numerical values.
* **Feature Scaling**:
  + Scaled Age and Fare to normalize the data.

**Model Training**

* **Algorithm**: Logistic Regression
* **Training and Testing Split**: 80% training, 20% testing.

**Evaluation**

1. **Confusion Matrix**:



* + True Negatives: 90
  + False Positives: 15
  + False Negatives: 19
  + True Positives: 55

1. **Accuracy**: **81.01%**
2. **Classification Report**:

**precision recall f1-score support**

**0 (No) 0.83 0.86 0.84 105**

**1 (Yes) 0.79 0.74 0.76 74**

**Accuracy 0.81 179**

**Macro Avg 0.81 0.80**

**Weighted Avg 0.81 0.81**

**Comparison with KNN**

* Logistic Regression achieved a slightly higher accuracy (**81.01%**) compared to KNN (**79.88%**).
* Logistic Regression showed better precision and recall for both classes, indicating more balanced predictions.

**Conclusion**

The logistic regression model performed well, achieving over 80% accuracy. It is suitable for predicting Titanic survival with the given features.