**Kairiz CYber Technologies**

***Submitted By: Abdullah Amjad***

***Cell No: 03309772003***

***Gmail****:* ***abdullahamjad17301@gmail.com***

***Instructor: Sir Afaq and Mam Noor***

***CEO: Kainat Rizwan***

***Date: 7 July 2024.***

**Task: 1**

1. **Install and Import Libraries:**

* pip: Used to install Python packages.
* pandas: For data manipulation and analysis.
* numpy: For numerical operations.
* matplotlib and seaborn: For data visualization.
* scikit-learn: For machine learning and preprocessing.

1. **Load Data:**

First, I will read all the data with **pd.read csv** and then display the data and count the missing values by head**()** and **isnull().sum().**

1. **Handle Missing Values:**

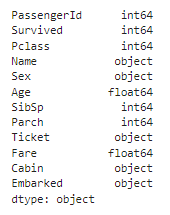
Then I will handle missing values like fille missing age and fare values with the median using **fillna().**

1. **Remove Duplicates:**

After Data load and fill missing values then I will remove the duplicates records from test and train data set using **drop\_duplicates().**

1. **Check and Convert Data Types:**

First, I will check the datatypes by **dtypes** then I will change the datatypes according to my need because everyone has its own way to do something that’s why I will change datatypes according to my need.



1. **Summary Statistics.**

Then I will work on some statistics like mean, median, minimum and maximum values to take some summary according to data and I use the **describe()** then after that work on the features.

1. **Visualize Data:**

Then I will visualize the data and create histograms for numerical features using **hist().** Then Create boxplots for 'Age' and 'Fare' with **sns.boxplot()** and then for comparison I will create scatter plots for 'Age' vs. 'Fare' with **sns.scatterplot().**

**Histograms:**

A group of blue and white graphs

Description automatically generated

A graph with blue lines

Description automatically generated

**Box plots (Age and Fare).**

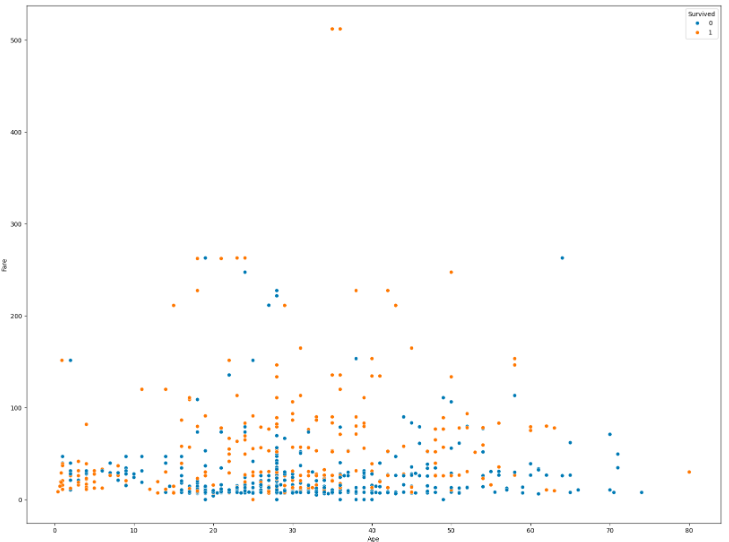
A blue line with black lines

Description automatically generated

A graph with a blue line

Description automatically generated

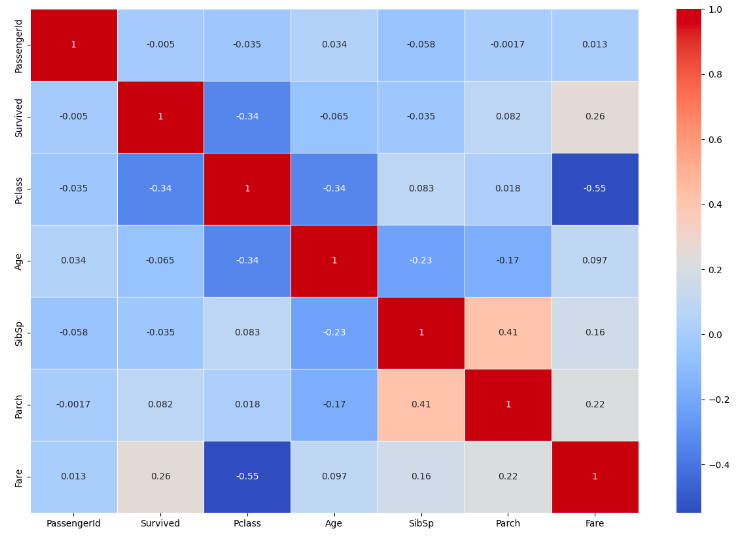
**Scatter plots:**



1. **Correlation Matrix:**

Basically, histogram is used for specific interval and boxplots visualize the spread and skewness of numerical data and highlighting the median and other things. Scatter plots is used to depict the relationship between two numerical variables and shows the difference between the variables.

On the other hand, **Correlation matrices** present the correlation coefficients between multiple numerical features, indicating the strength and direction of linear relationships.



1. **Create New Features.**

After these visualization then I will add some more features like ‘**FamilySize’** by summing '**SibSp'** and **'Parch'** and **adding 1**. And then second feature is '**IsAlone'** to indicate if a passenger is alone or not.

1. **One-Hot Encoding:**

In the hot encoding first, I will create the some dummy variable in Sex and Embarked columns in both the train and test datasets then I will use **align()** function to ensure that the both train and test datasets have the same columns.

1. **Standardize Numerical Features:**

After that I use the **StandardScaler.** Basically, **StandardScaler** is a convenient and efficient tool from scikit-learn to standardize features. Here I use this in Age, Fare and FamilySize.

1. **Now Resample the Data for Imbalance.**

In this step I have to balance the dataset if it is not balance and I will use the **value\_counts()** to check the imbalance in the Survived column.

In the last final step, I will print the rows of the train and test datasets to see the changes in the output using **head().**

**Helping Material:**

<https://youtu.be/dRBYkDNkrBI?si=PIQPKgqY8uMOOzq2>

<https://youtu.be/EKk_mcobsF8?si=V-iImJmV8JqIOd8t>

<https://youtu.be/r-Dihn9-uTU?si=LZkWIMSq45x7LxMI>

<https://youtu.be/5VD4V3uIkJc?si=Zz6Mqg95N9QlQ_Wh>

<https://youtu.be/tse_8LLWtfY?si=3H7A29-2D0OjuRNE>

<https://youtu.be/8Vq9MkTklCM?si=ZnWckXURTLkxLVHl>

<https://youtu.be/1fFVt4tQjRE?si=ba_0GIY6DUrqa0IS>

<https://youtu.be/-xqHsLE0AfM?si=TeaybhhqxY3WOns7>

<https://community.alteryx.com/t5/Data-Science/Life-or-Death-Prediction-with-the-Titanic-Dataset/ba-p/178966> (Forum)

<https://forums.fast.ai/t/kaggle-titanic-and-decisiontreeregressor/79655> (Forum)

**Thanks.**