### 

### 

### 

### 

### 

### 

### 

### **Title:** *Titanic Dataset Analysis*

#### **Name:** Antoine Gaton

#### **Date:** October 13, 2024

#### **Course:** CS491

#### **Assignment:** Titanic Dataset Analysis with Unsupervised and Supervised Learning

### 

### 

### 

### 

### 

### 

### 

### 

### 

### **Introduction:**

This assignment involves analyzing the Titanic disaster dataset using both unsupervised and supervised machine-learning techniques. The goal is to:

1. **Unsupervised Learning**: Perform clustering using the KMeans algorithm to describe patterns in the data without using labels.
2. **Supervised Learning**: Build two models (Logistic Regression and Random Forest) to predict which passengers survived the disaster based on available features.

### **1. Data Preprocessing:**

To prepare the dataset for analysis, several steps were taken to clean and transform the data:

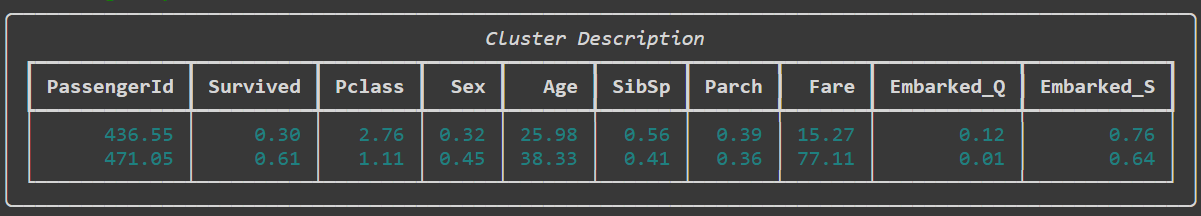
1. **Dropping Irrelevant Columns**:
   1. The columns *Name*, *Ticket*, and *Cabin* were dropped because they do not provide meaningful information for predicting survival. These columns are either unique to each passenger or have too many missing values.
2. **Handling Missing Values**:
   1. The *Age* and *Embarked* columns had missing values, which could affect the performance of machine learning models. To address this:
      1. For the Age column, missing values were filled with the *median age* of the passengers. The median is a robust measure that reduces the impact of outliers compared to the mean.
      2. For the Embarked column, missing values were filled with the *mode*, which represents the most frequent port of embarkation. This ensures that the missing values are replaced with a plausible category.
3. **Encoding Categorical Features**:
   1. The *Sex* column, which contains categorical values, male and female, was encoded as numerical values, 0 for male and 1 for female This transformation allows the machine learning models to process the data, as most models require numerical inputs.
   2. For the Embarked column, which contains multiple categorical values of the ports where passengers boarded the ship, *one-hot encoding* was applied. This converts each category into its binary column (e.g., Embarked\_Q, Embarked\_S). One-hot encoding prevents the model from assuming any ordinal relationship between the categories, which could distort the analysis.

These preprocessing steps are essential for preparing the data in a way that machine learning models can understand, ensuring better model performance and more accurate predictions.

### **2. Unsupervised Learning (KMeans Clustering):**

I applied KMeans clustering with 2 clusters to categorize passengers. The goal was to find natural groupings in the data. The clusters were described based on the mean values of the features.

#### **Cluster Description:**

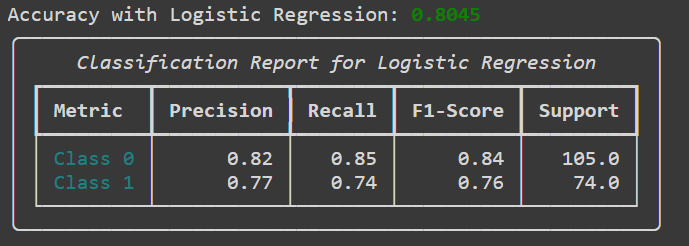


### **3. Supervised Learning:**

#### **Logistic Regression:**

Logistic Regression is used as a baseline classification model to predict survival. The accuracy and classification report of this model are included below:

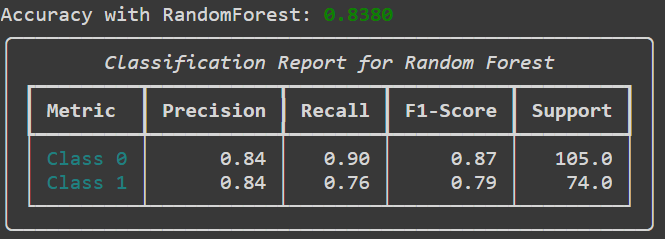
* **Accuracy**: 80.45%
* **Classification Report**:
  + Precision, Recall, F1-Score for Class 0 and Class 1



#### **Random Forest Classifier:**

Random Forest is an ensemble learning model that typically performs better than basic classifiers like Logistic Regression. It builds multiple decision trees and averages the results to improve performance.

* **Accuracy**: 83.80%
* **Classification Report**:
  + Precision, Recall, F1-Score for Class 0 and Class 1



### **4. Conclusion:**

The Random Forest model outperformed the Logistic Regression model, achieving an accuracy of 83.80%. This demonstrates the strength of ensemble methods in handling complex datasets like the Titanic data. The KMeans clustering analysis revealed interesting insights, showing that passengers in *Cluster 1* were, on average, older and had a higher survival rate. These natural groupings provide a better understanding of the data's structure, which can aid in refining the supervised learning models.

Further improvements could be made by tuning the hyperparameters of the Random Forest model to enhance its performance. Additionally, exploring more advanced models such as *XGBoost* or *Gradient Boosting* could help push the accuracy closer to or above 90%.

Moreover, experimenting with adding more clusters during the KMeans analysis may uncover more granular groupings, providing a deeper understanding of passenger characteristics. By identifying more nuanced subgroups, the clustering could inform feature engineering and potentially improve the performance of the supervised models.