***Report of the Project Done in Partial Fulfilment of the Requirements for the Award of Master of Science in Data Analytics***

**Big Data Analytics** - **IMAT5322**

TITANIC SURVIVAL REPORT



*Done by*

Juny Maria John – P2743266

*Under the guidance of*

**Ms Bangli Liu**

**Mr Bamba M**

*Professor at the Department of Computing, Engineering & Media*

Master In Science

Data Analytics

Faculty of Computing, Engineering & Media (CEM)

De Montfort University

Leicester – LE1 9BH

**MARCH- 2023**

**ABSTRACT**

The Titanic dataset is a .csv file that contains information about the passengers who were onboard the Titanic when it sank after hitting an iceberg on April 15, 1912. It holds crucial information about the people on board. The main objective of this project is to choose the best classification model to evaluate data for accurate survival prediction. The CatBoost classification model serves as an excellent example of how to approach a real-world data science problem, from data pre-processing to experimental result analysis.

**Keywords-** classification, CatBoost, pandas, NumPy, matplotlib, seaborn, clustering.

**INTRODUCTION**

**PROBLEM:**

Inaccurate survival predictions from classification models may occur due to unclean data in both the train and test sets of the Titanic dataset.

**TASK BREAKDOWN AND PROPOSED IMPLEMENTATION:**

1. Prepare the data by cleaning it and converting it into organized data.

2. Extract the characteristics that depict the data.

3. Execute the categorization of the extracted features.

4. Comparison of results using various tools and models

5. Derive the best approach for predicting survival on the Titanic.

**LITERATURE REVIEW:**

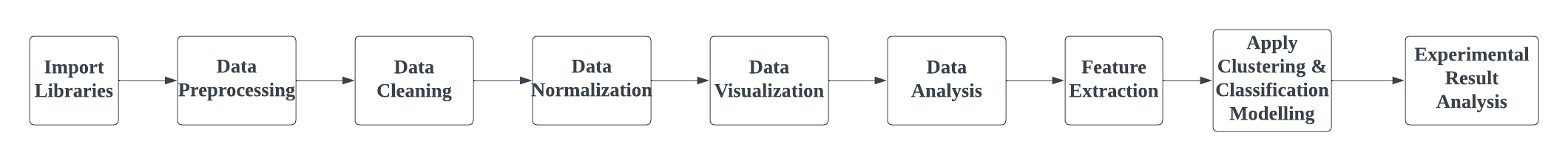
The Titanic dataset is extensively explored for analysis with the objective of survival prediction based on a set of input features. Several ML algorithms and analyses have been used on the dataset namely: In [1], logistic regression, decision trees, and random forests were used to predict survival using features such as age, gender, ticket class, and fare. In [2], feature selection was used to identify the most important features for survival prediction, while in [3], data normalization was used to improve the accuracy of the SVM. In [4], imputation techniques were used to fill in missing values in the dataset. Overall, the Titanic dataset remains a valuable resource for studying machine learning and predictive modelling.

**INNOVATIVE THINKING**:

Extract features from family dynamics and conduct modelling analysis.

**DETAILS OF APPROACH**

**FLOW CHART:**



**METHODOLOGY:**

* Import the dataset and comprehend the main libraries.
* Conduct an Exploratory Data Analysis (EDA) by utilizing Pandas.
* Identify missing values (NaN) using the data cleaning process – Raw data contains noise.
* Handling NaN either by removing null/replacing it with a non-null value(mean/median/mode).
* Standardize the data for modelling [transforming categorical into numerical features].
* Gain insights for modelling by visualizing data characteristics and by reviewing them.
* Enhance the model's performance and address underfitting by extracting features from the existing data.(Feature Engineering)
* Utilize various ML classifiers and clustering from the Scikit-learn library and use parameter metrics to assess performance.
* Conduct experimental analysis by comparing multiple models to determine the most effective approach.

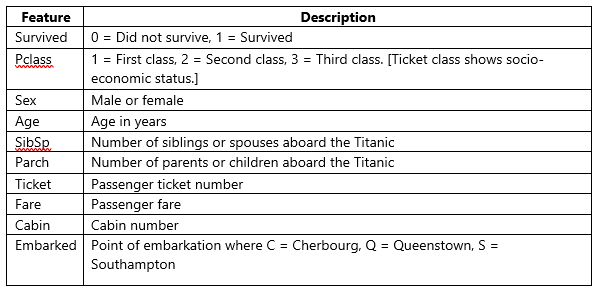
Note: All steps are done on both train and test data. Based on the information we currently have i.e. we want to teach MLs model to understand the relationship between passenger features and survival outcomes, and then use that model to forecast survival on new datasets.

Through exploratory data analysis (EDA), we identified the correlation between variables such as Gender, Age, Ticket class (Pclass), Fare, Sibling/Spouse (Sipsib), Parent/Child (Parch), and Port of Embarkation on the target (Survival). Utilizing these variables and engineered features, we constructed a dataset to employ both classification and clustering models for prediction purposes. Ultimately, after assessing the models using training accuracy, K-fold cross-validation, and modifying the feature data, we determined that CatBoost is the optimal model and we are confident that it will provide accurate predictions on new, unseen data.

**TOOLS USED:** Kaggle Notebook

**PROGRAMMING LANGUAGE:** Python

**TITANIC DATASET FEATURES:**

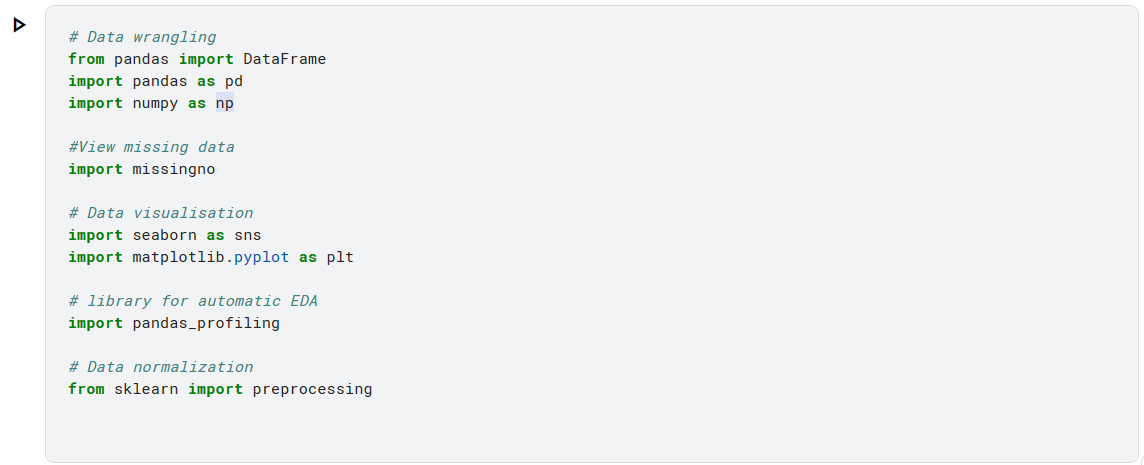


**RESULT ANALYSIS**

**1. IMPORTING LIBRARIES AND TITANIC DATASET:**

The first step is to import .csv datasets that are properly formatted with valid data and the necessary libraries.



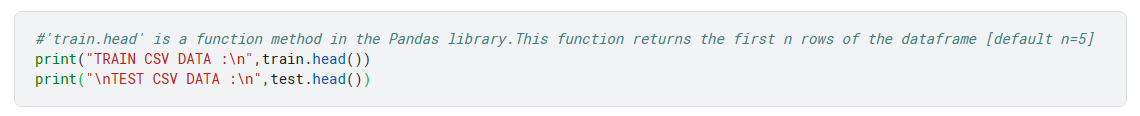


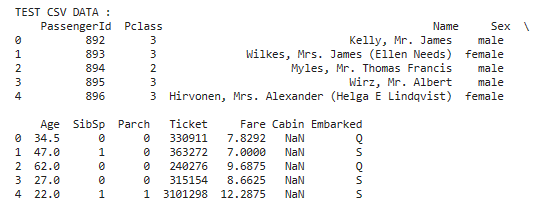
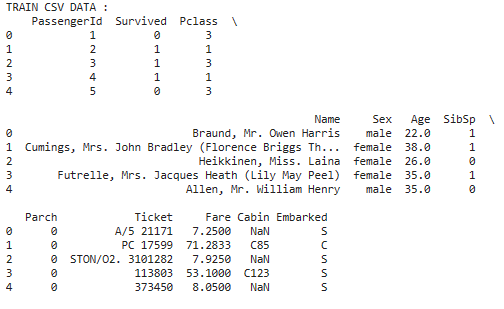


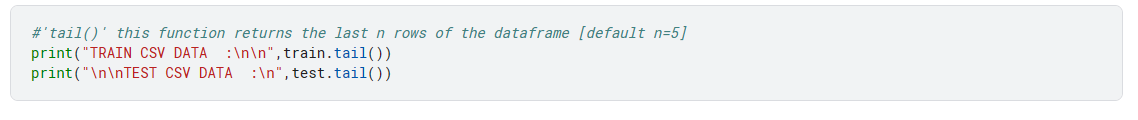
**2.** **DATA PRE-PROCESSING**

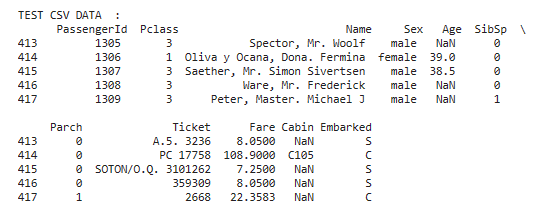
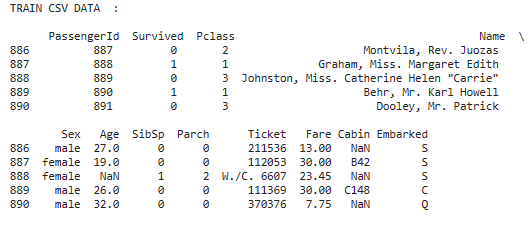


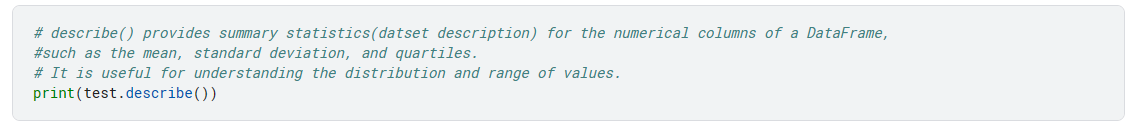
Analyzing data using functions.

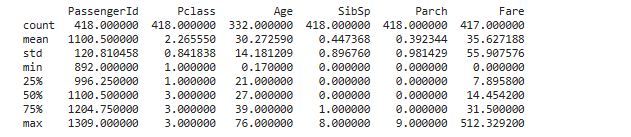


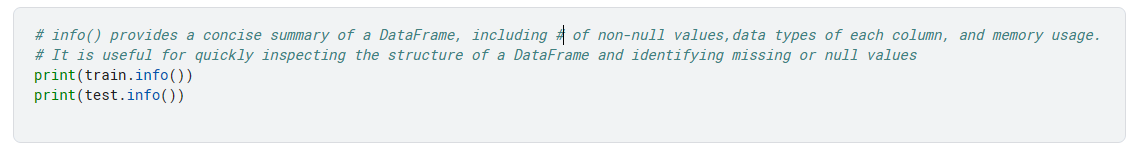


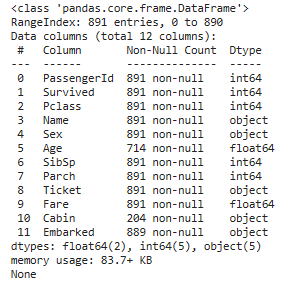
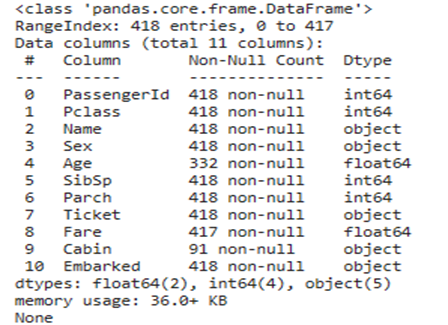




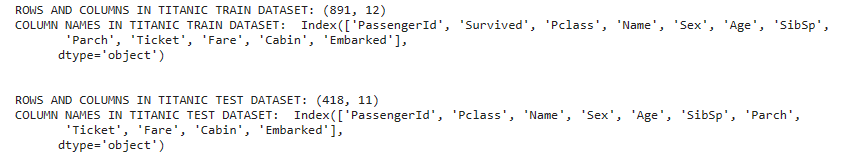






****



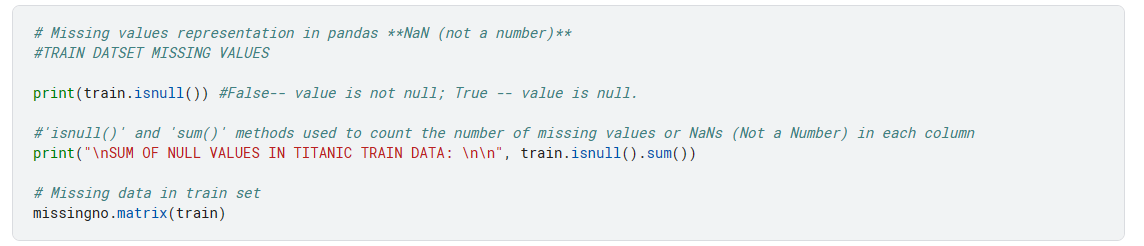


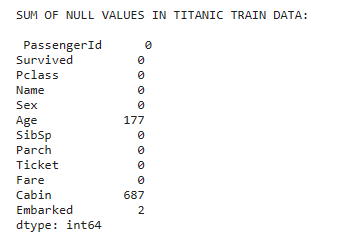
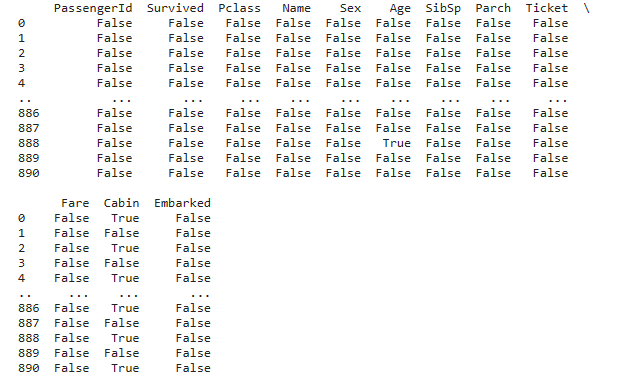
A **significant distinction** between the ‘train’ and ‘test’ datasets is that the ‘test’ dataset lacks one column compared to the ‘train’ dataset. Specifically, the ‘Survived’ column, which represents the **target variable**, is absent in the ‘test’ dataset. This is done to utilize our analysis of the ‘train’ dataset to forecast the survival of passengers in the ‘test’ dataset.

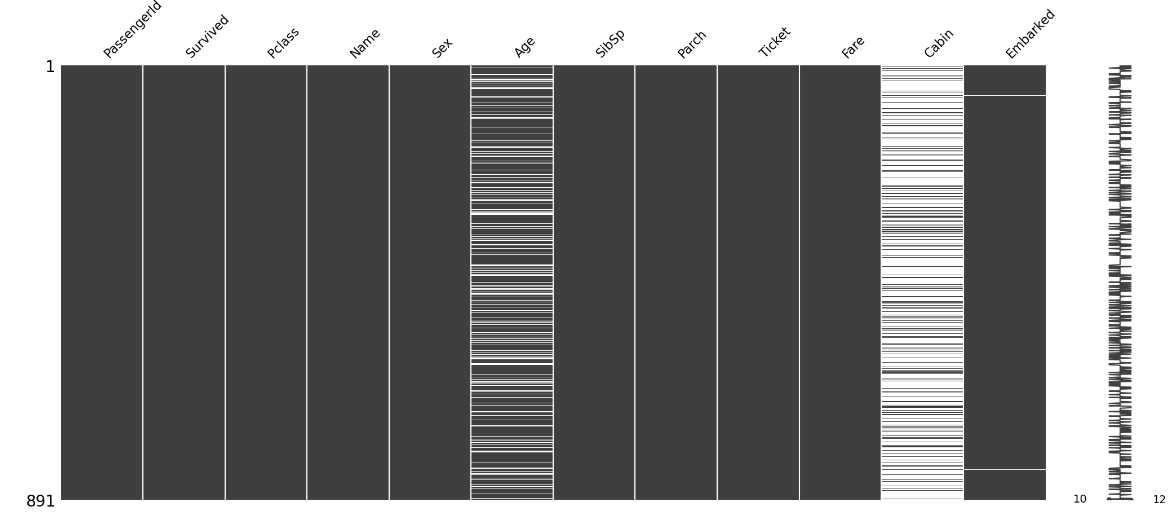
**3. DATA CLEANING**

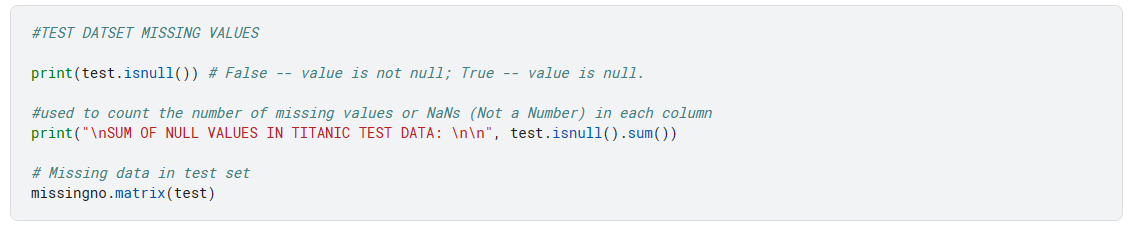
The process of detecting, correcting, replacing/removing inaccuracies, incongruities, and discrepancies in a data set.

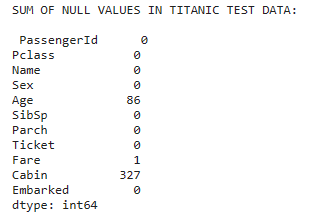
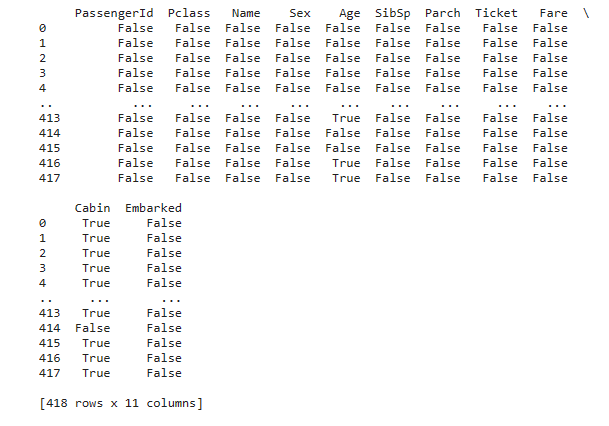
**3.1 Detecting missing values**











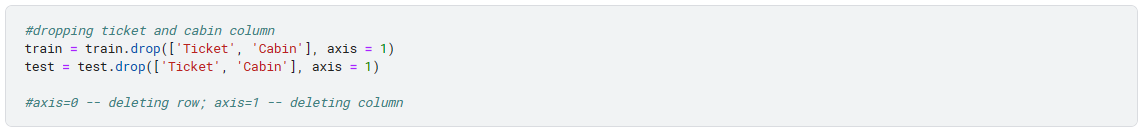


* Blank space in matrix shows NaN
* Age, Cabin, and Embarked columns in the Training Dataset have NaN
* Age, Fare, and Cabin in the Test Dataset have NaN.

**3.2 Handling Missing Data**

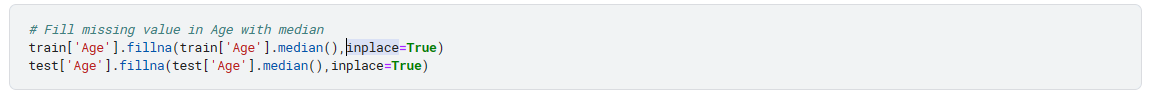
While dealing with missing data in data analysis care must be taken as it can affect the accuracy and validity of the results.

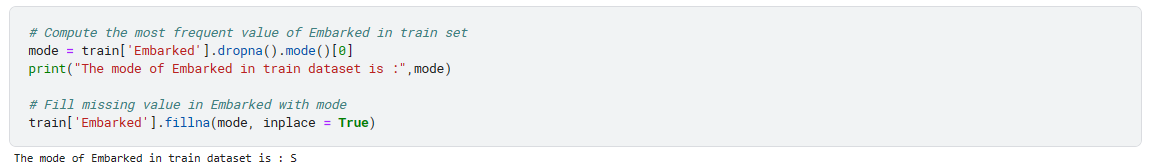
**3.2.1 Delete Column in Test and Train Data**

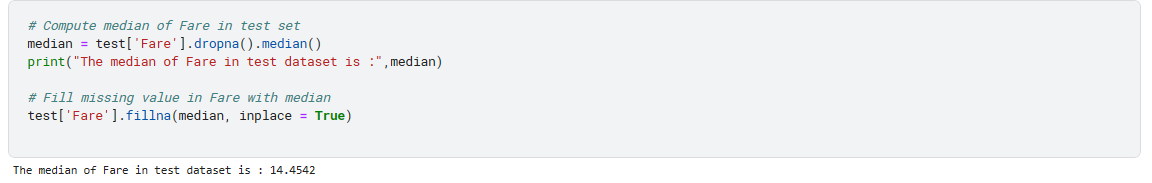


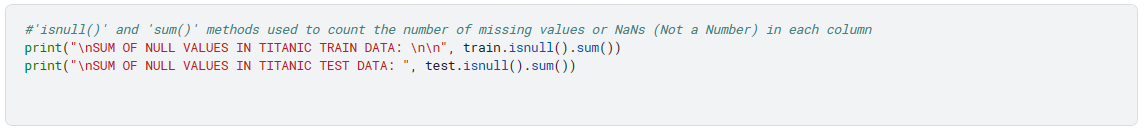
* As **Cabin** has more than 70% missing data, it is better to remove columns with a high percentage of missing data, as it can affect the accuracy and reliability of the model.
* Removing the **Ticket** to remove features to help reduce the complexity of the model and prevent overfitting.

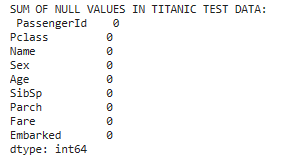
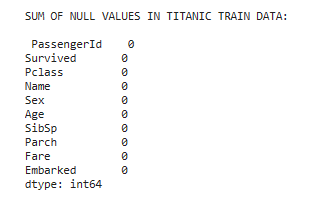
**3.2.2 Replace with Mean/Median/Mode**





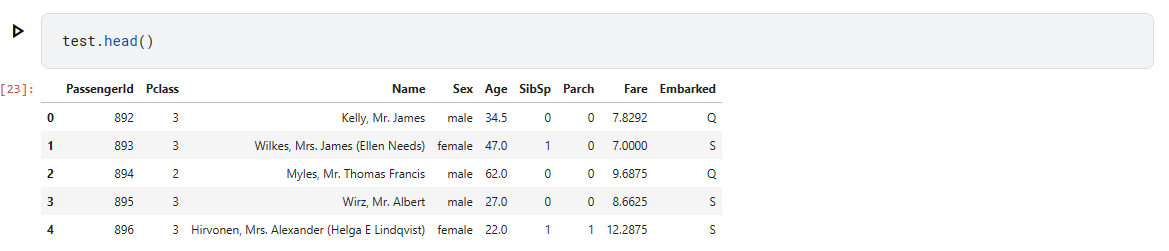


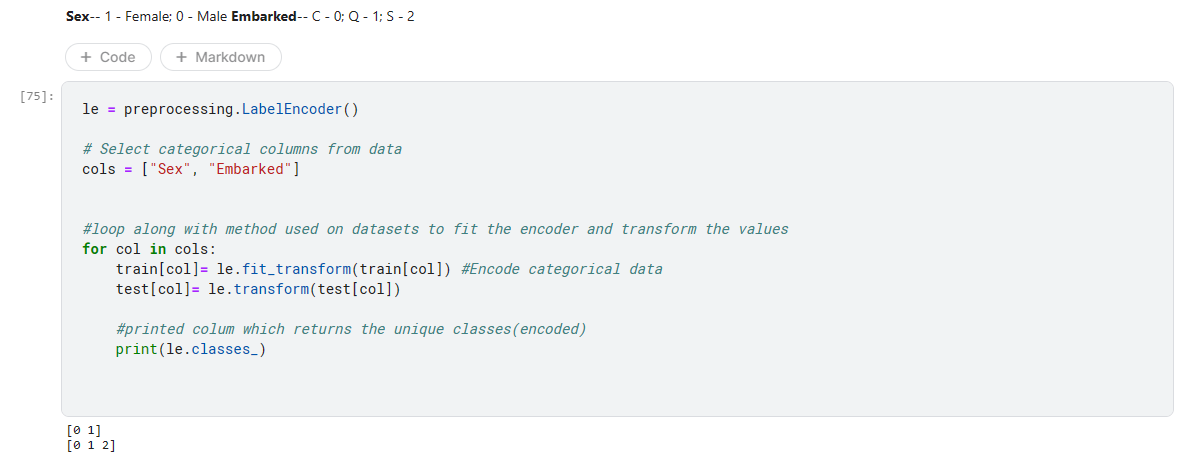




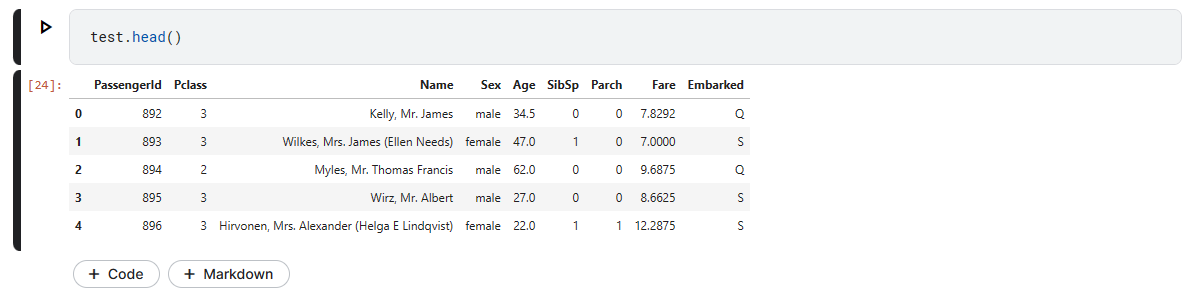
**4. DATA NORMALISATION**

The purpose of normalization is to make sure that the parameters of attributes are similar so that there is no conflict with the ML algorithms.





Data normalization done on **'Sex' and 'Embarked'** as they are categorical values.

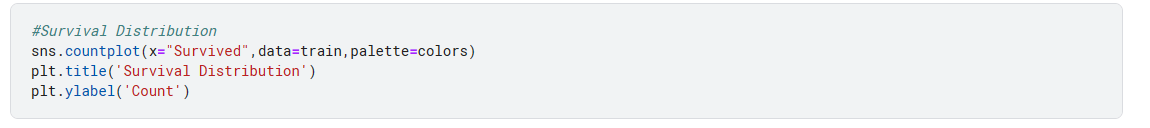


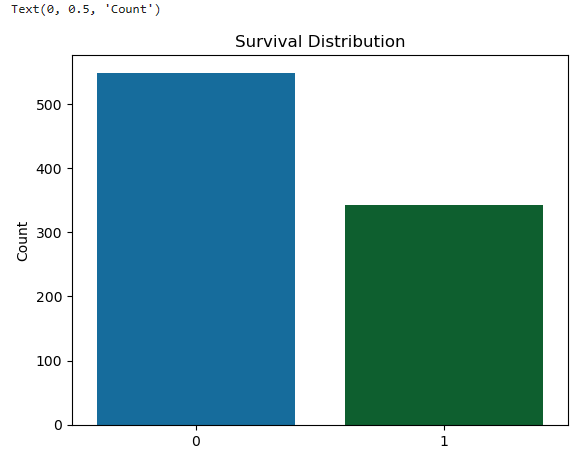
**5. DATA VISUALISATION**

Data is graphically represented in data visualization.

**5.1 Using Matplotlib and Seaborn Libraries**

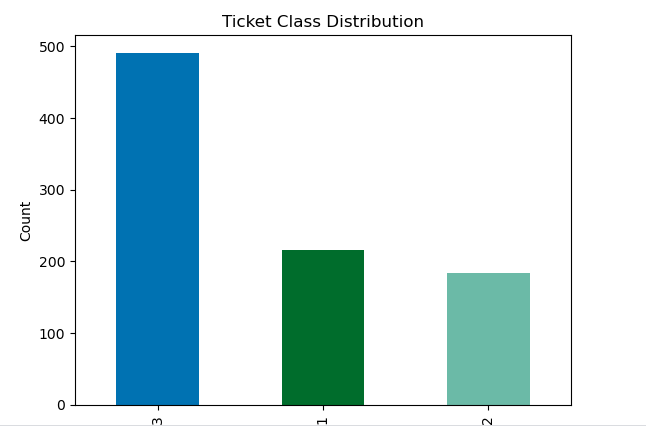
Each characteristic is plotted against the overall number of people aboard. To acquire a total tally, perform a passenger count without normalization. This allows us to acquire information about the individuals onboard.





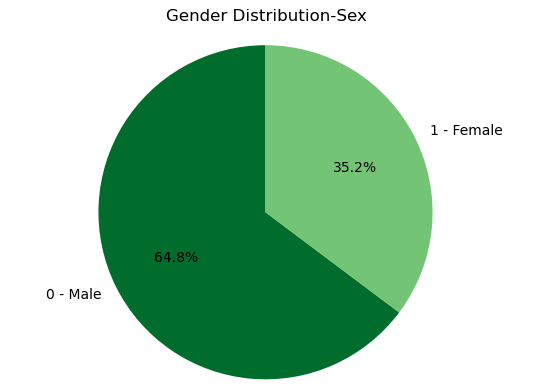
* Survival rate is significant low.





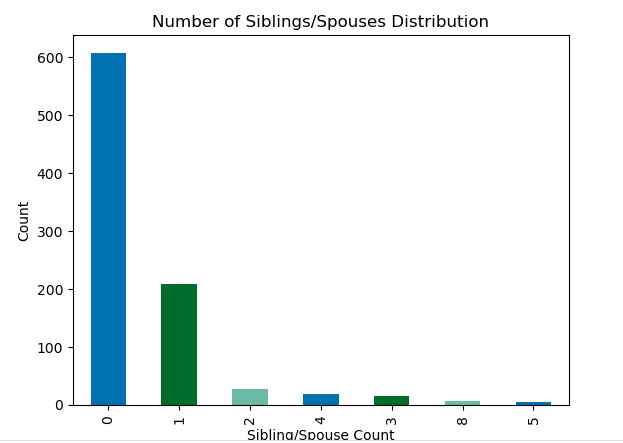
* Number of individuals onboarded is higher in Pclass3



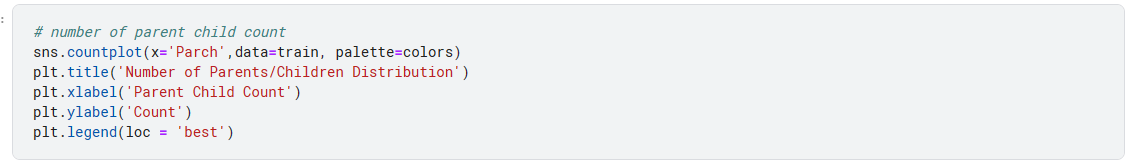


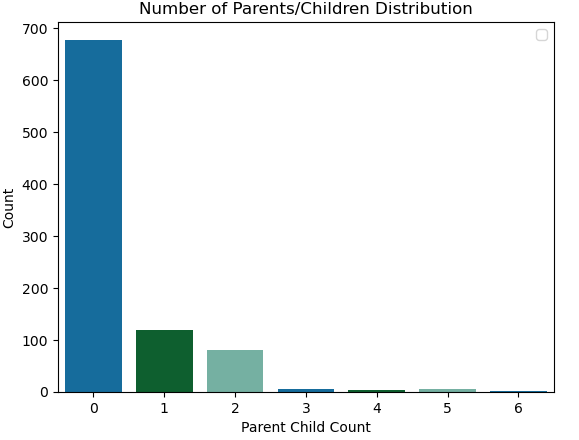
* The ratio of males onboarded is higher than females





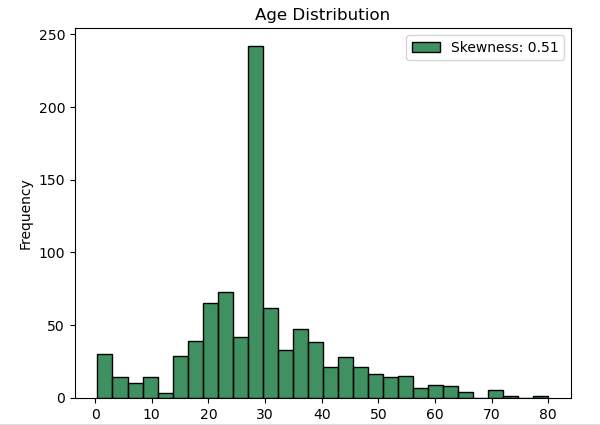
* The number of passengers travelling alone is higher.

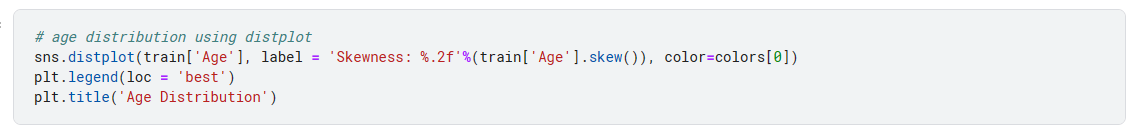


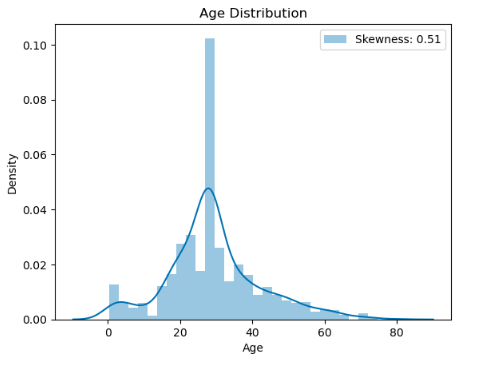


* The number of parents/children travelling alone is high.

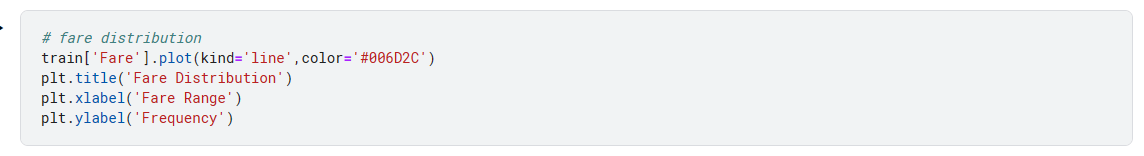


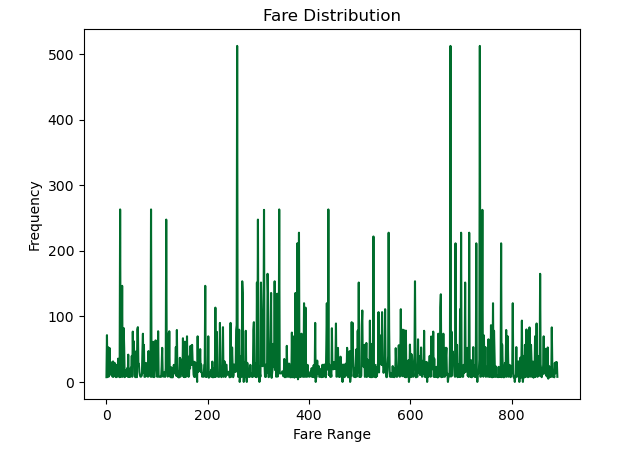


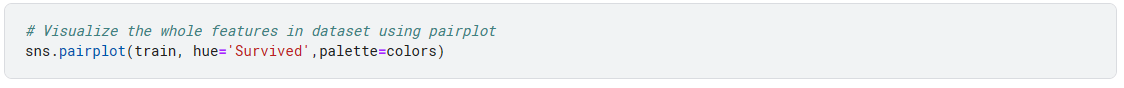


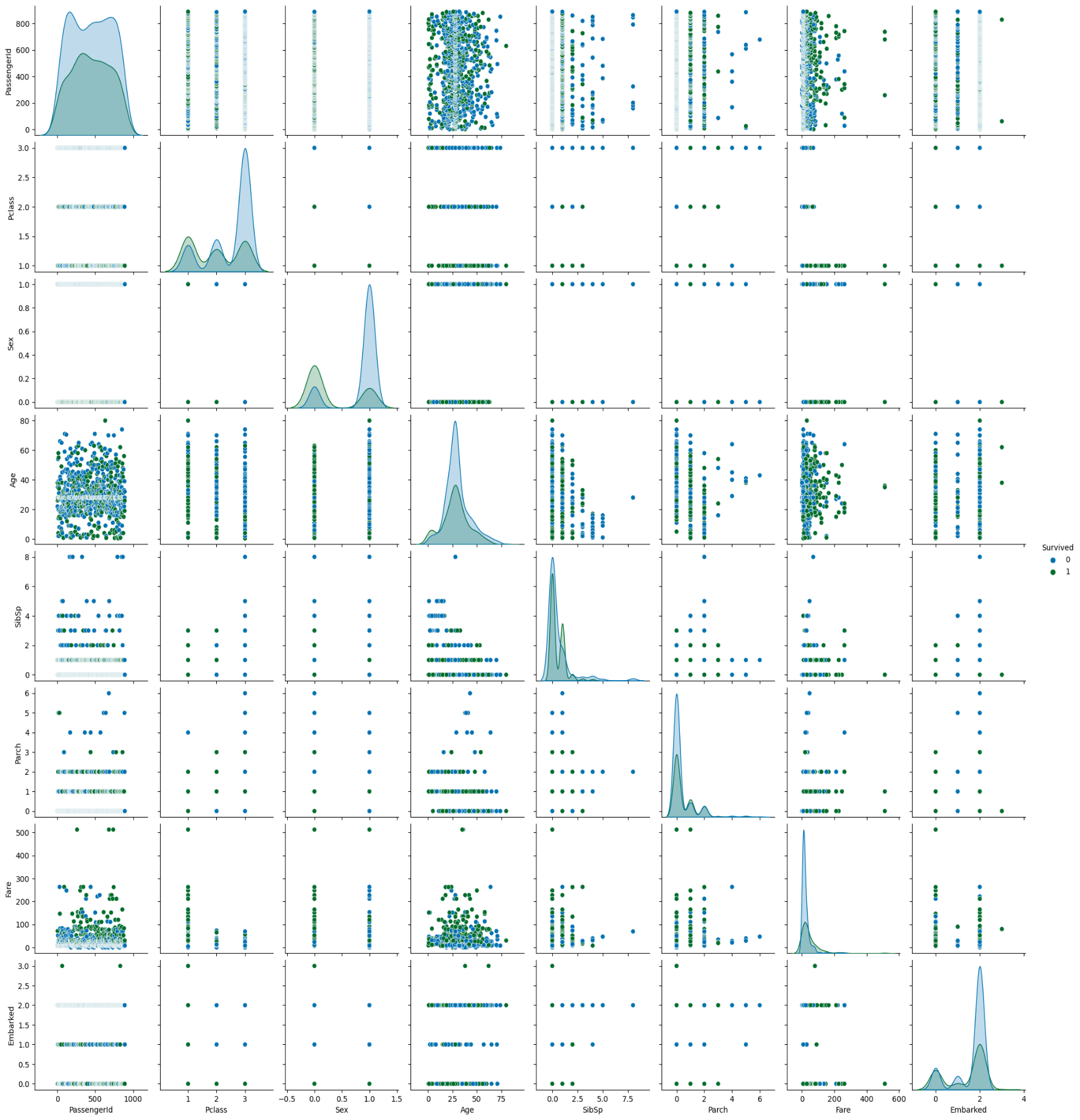


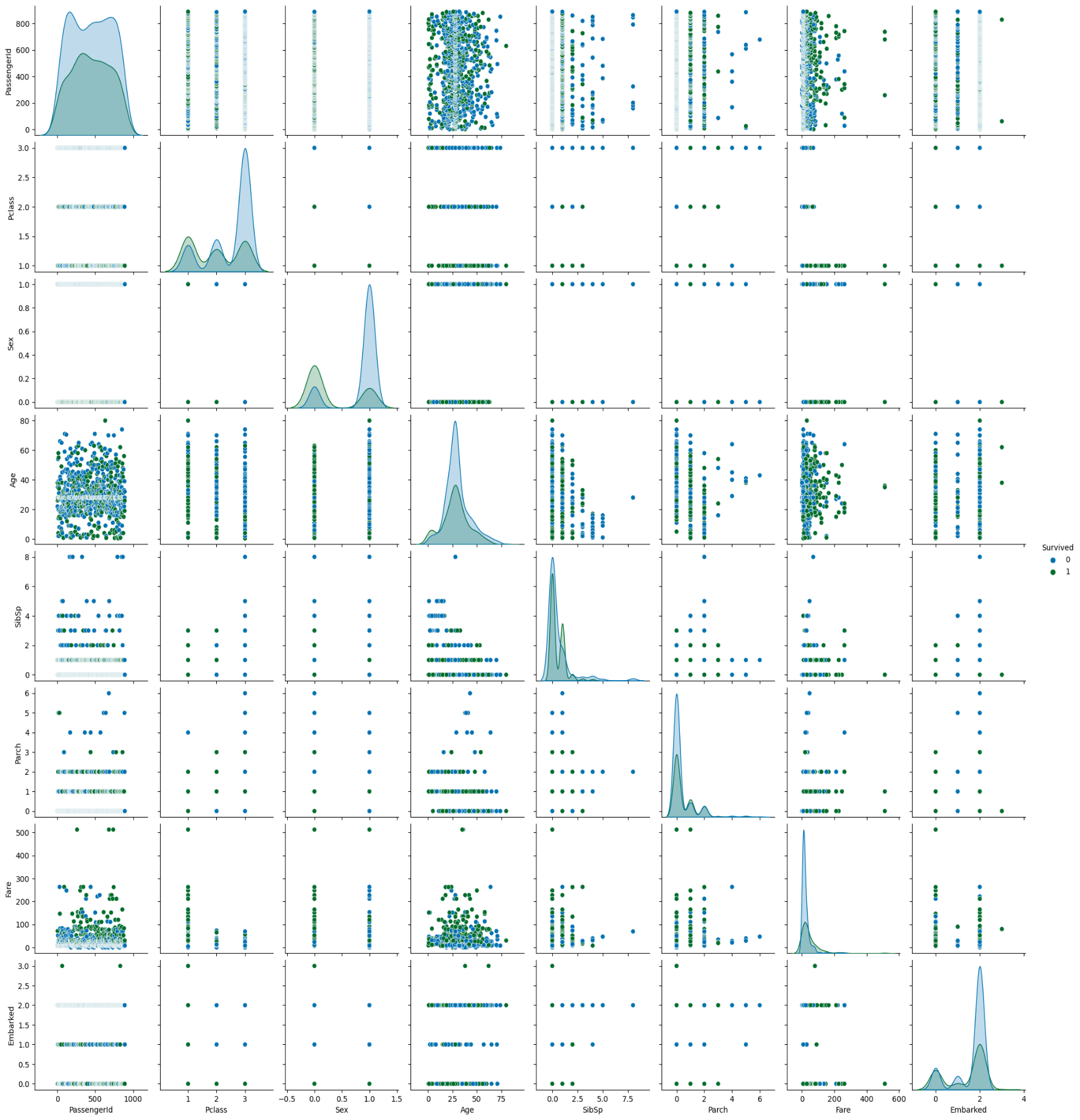
* The passengers aged between 25-30 years are higher than other age groups.







****

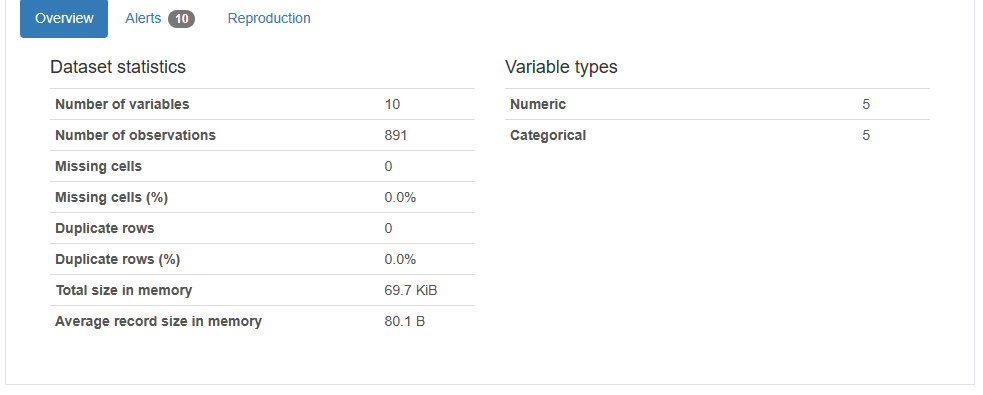
****

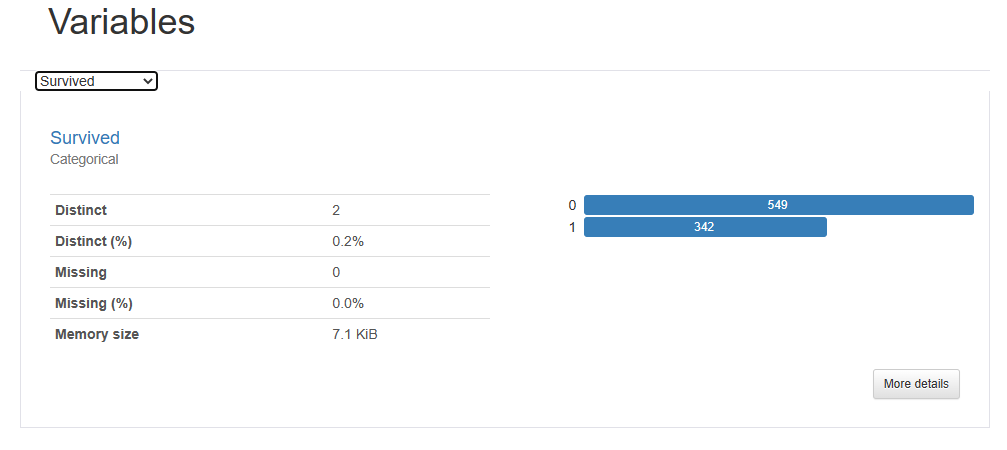
**5.2 Using Pandas Report:**

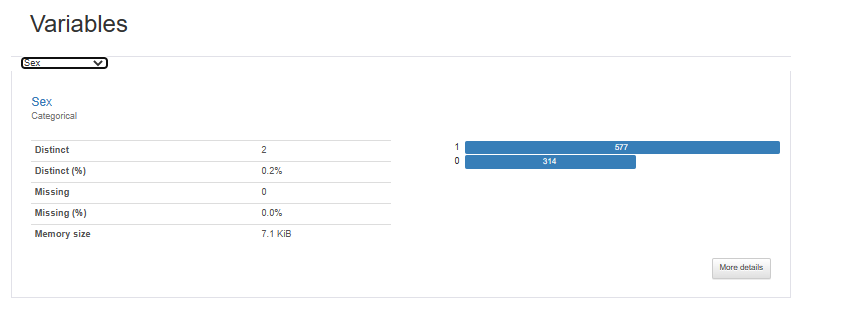


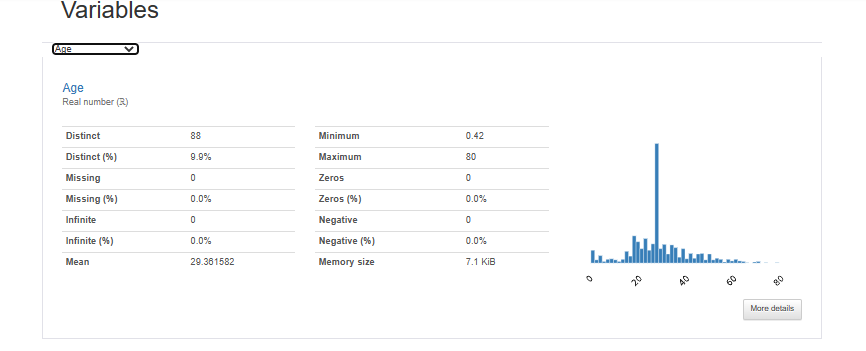


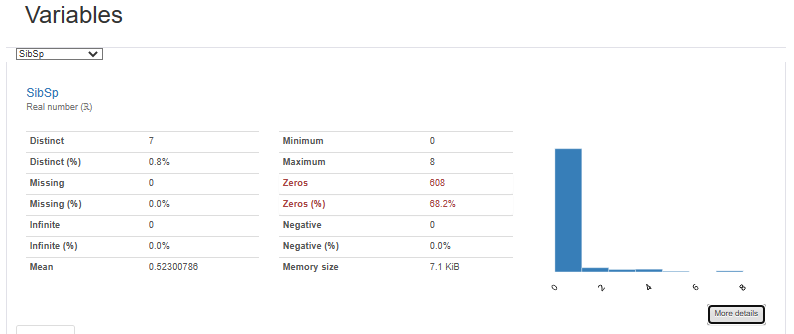
* Pandas\_profiling summarizes both train/test data and generates a report

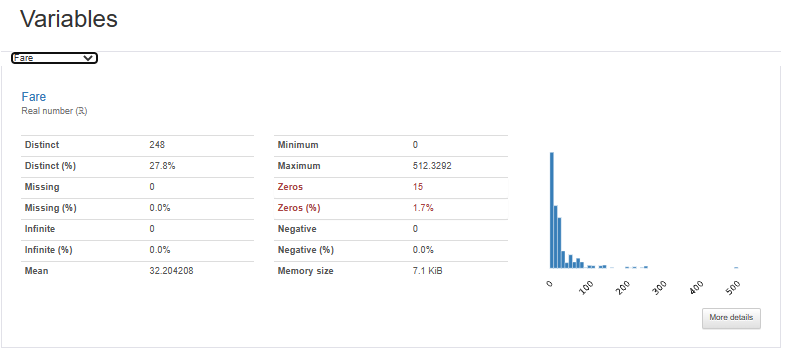




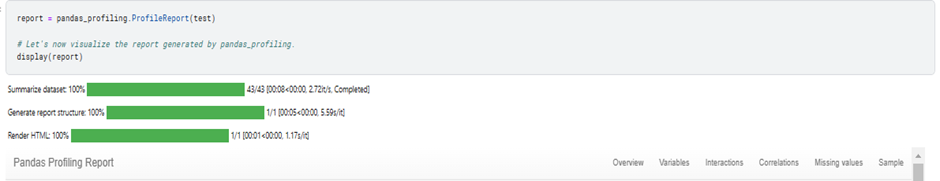


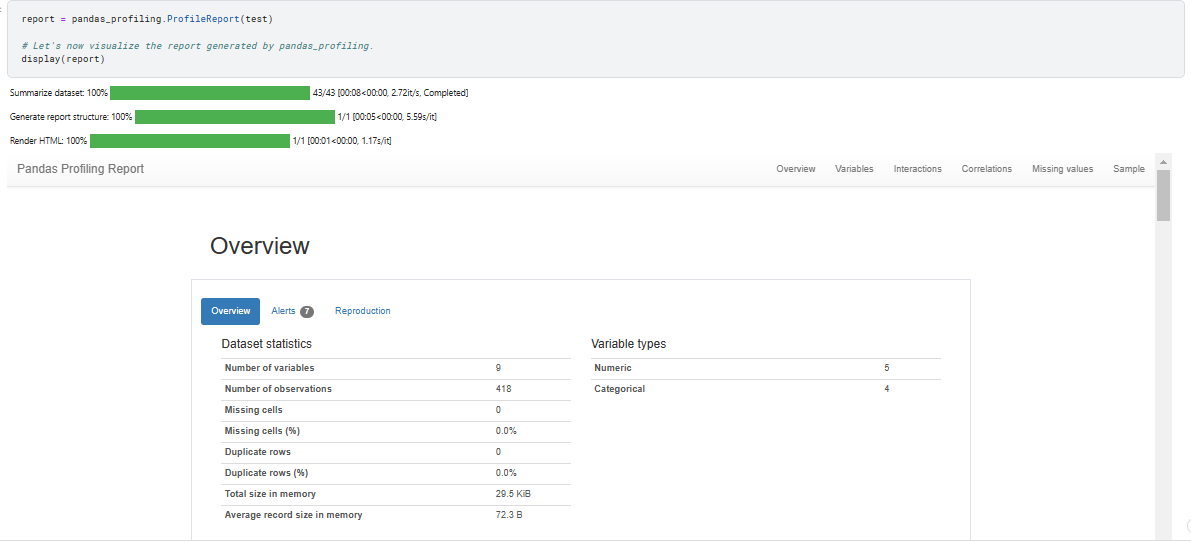


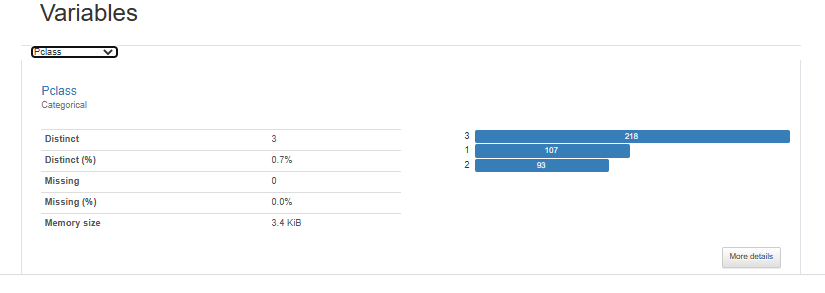




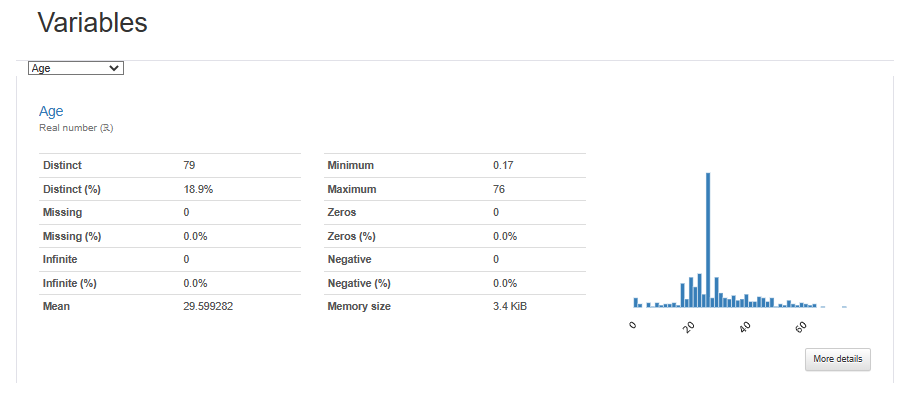
**Report for the test dataset.**

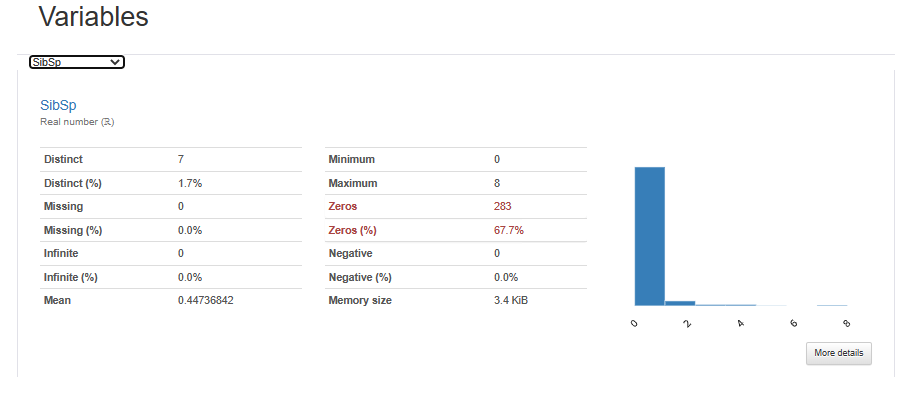


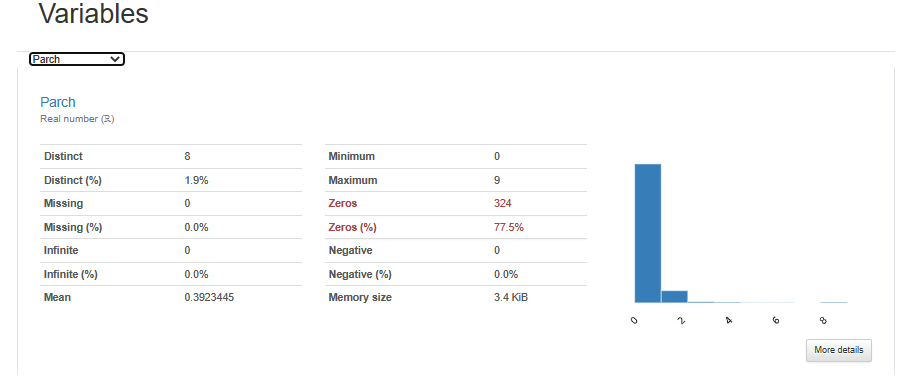












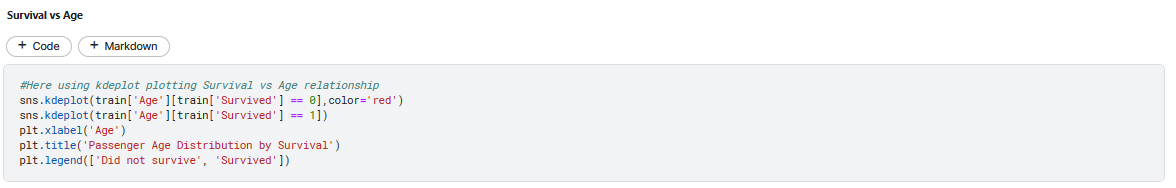
**6.**  **DATA ANALYSIS**

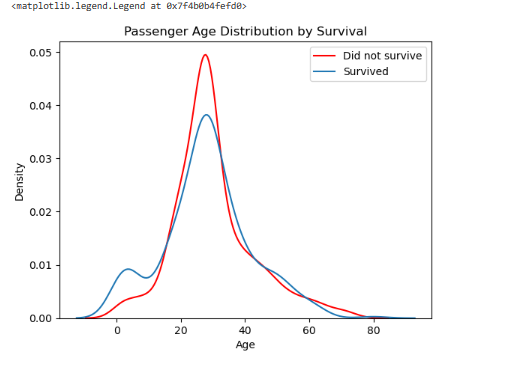
**Tools and Methods used:**

Here we are importing seaborn,matplot libraries and using kdeplot, barplot, bar charts, and correlation map to visualize and summarize data

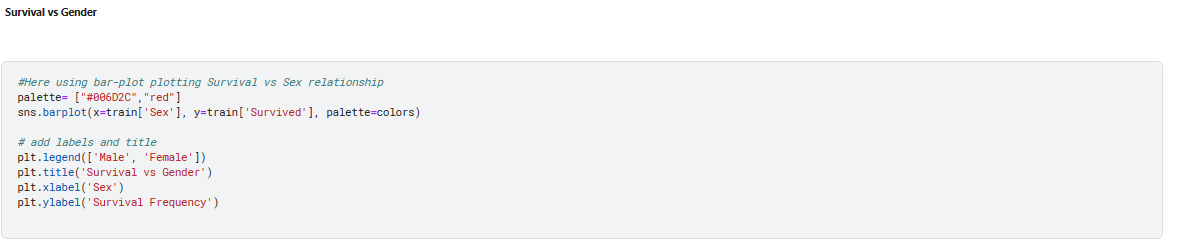
Pivot table and Pandas Library to do the statistical analysis for survival rate based on different categories and derive conclusions.

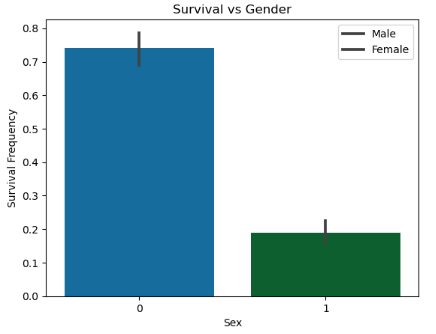
**6.1 Using Libraries visualize Dataset**



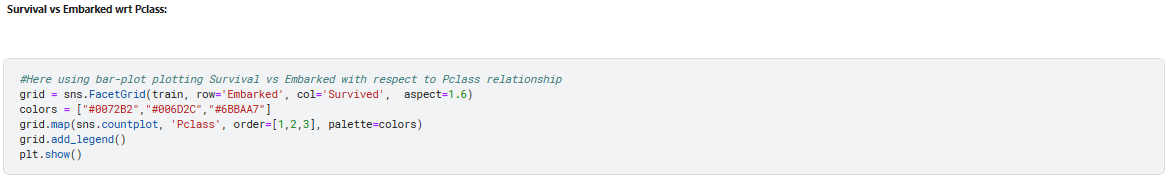


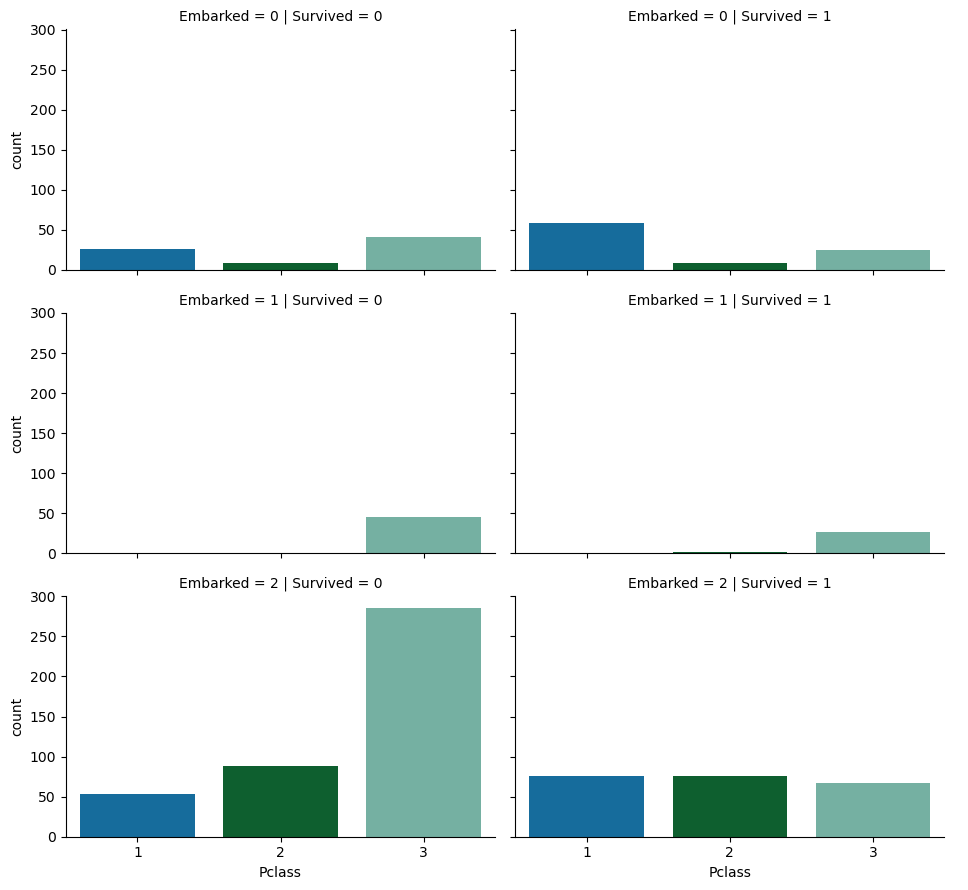
Majority of individuals onboard were aged between 25-30, they also experienced the greatest number of fatalities, while infants and elderly people had fewer onboard and consequently fewer fatalities**.**



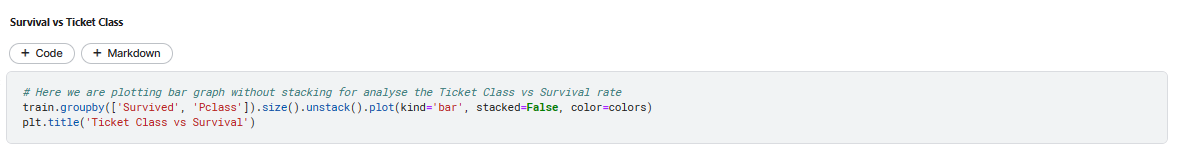


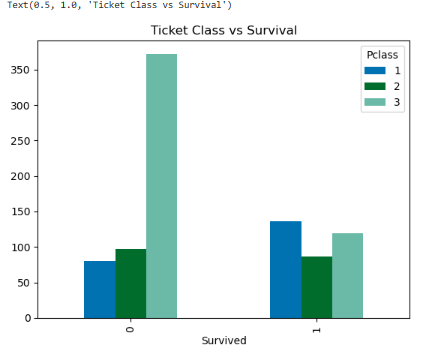
The survival rate of females is higher than men



****

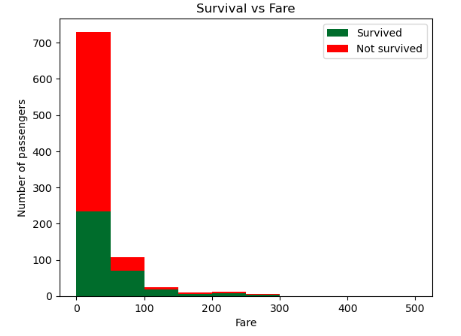






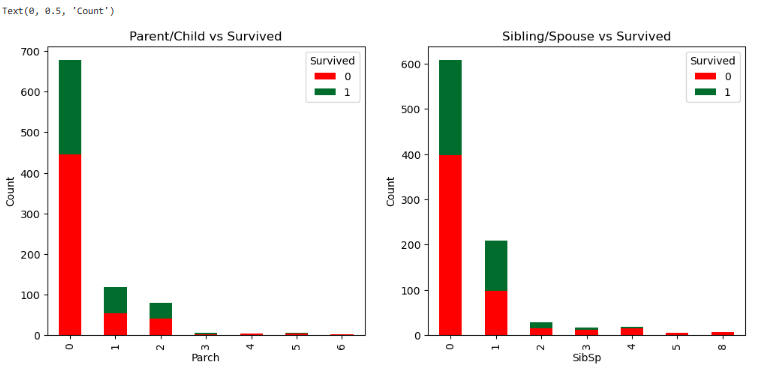
Therefore, socio-economic class (Ticket pricing) from each embarkment plays important role in survival with a lesser rate of survival of Pclass3 passengers.





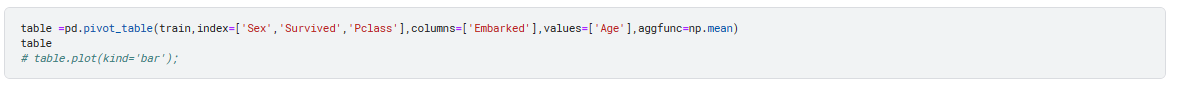
* **Survival rate vs Ticket class/Fare:** Lower the socio-economic status higher the death rate

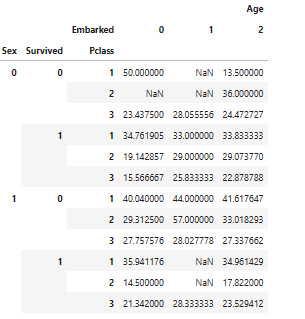




The mortality rate of the passengers travelling alone is higher than those travelling with 1 or more family members

**6.2 Using Pivot Table**





**6.3 From Pandas\_Report**



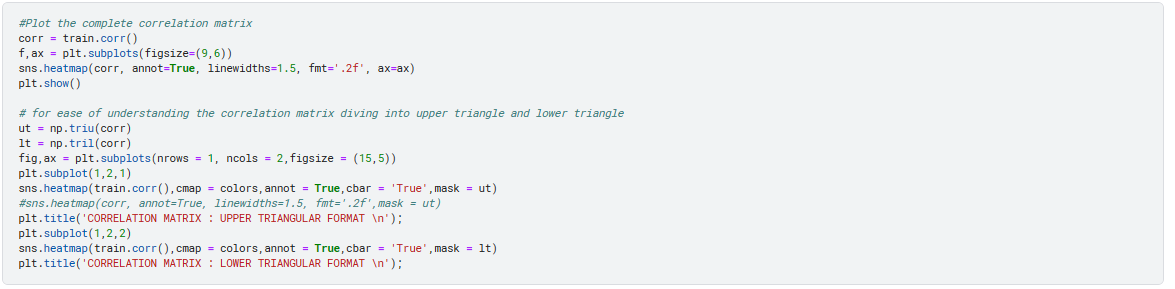


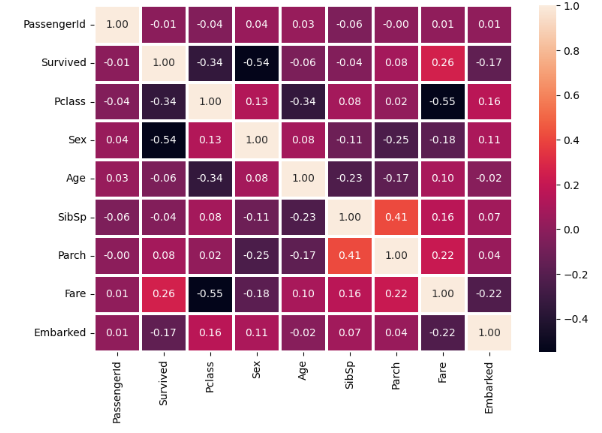
****

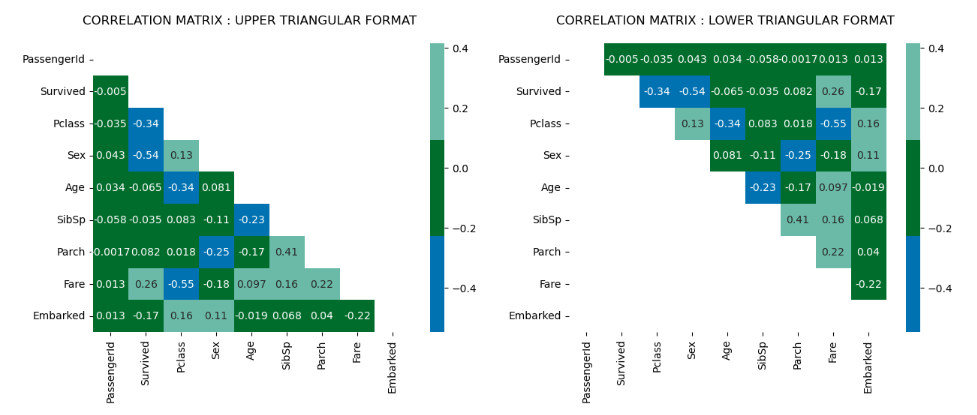
We can get information regarding the datasets from this.

**7. FEATURE EXTRACTION**

**7.1 Feature Selection using Correlation Matrix**

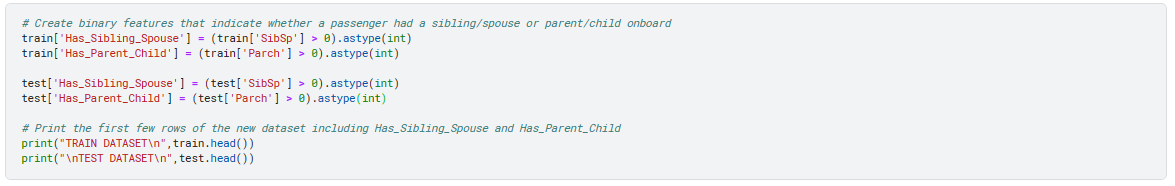


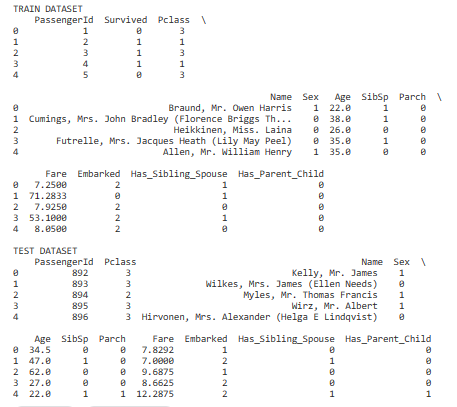
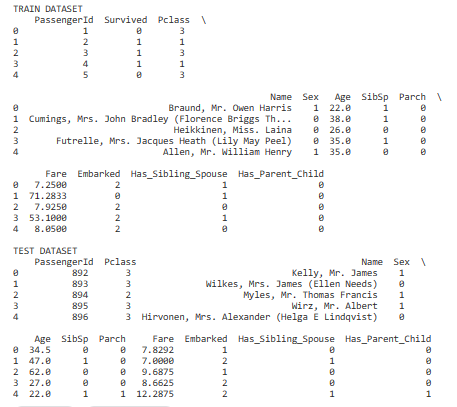




* Select features such as **Parch** and **Siblings** since they show a higher correlation.
* Extracting **Age** parameter to get more detailed information for Survival Prediction

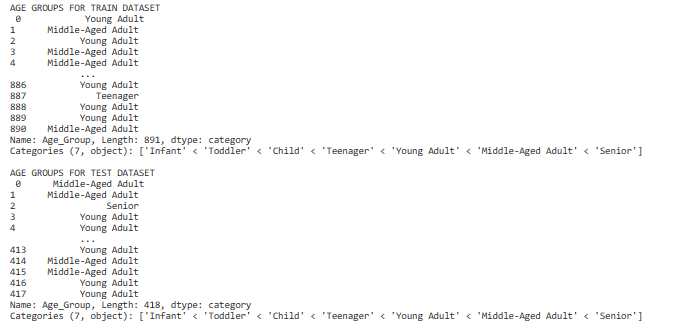
**7.2 Extracting from Sibsp and Parch - Adding Has\_Sibling\_Spouse and Has\_Parent\_Child feature**

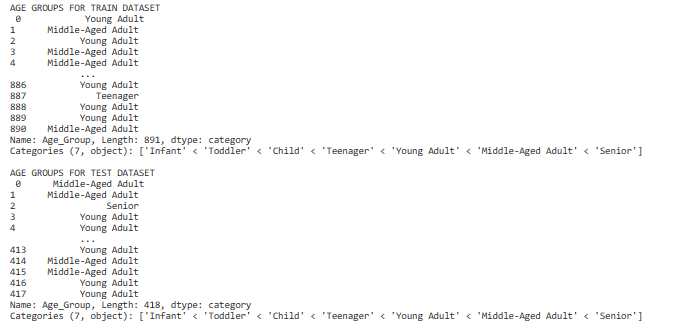


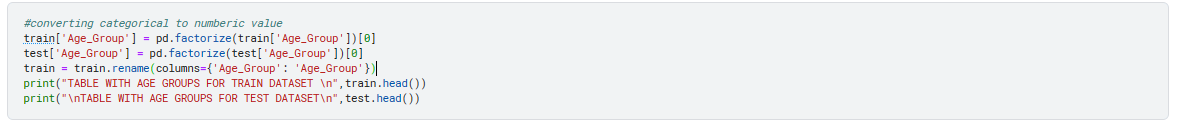


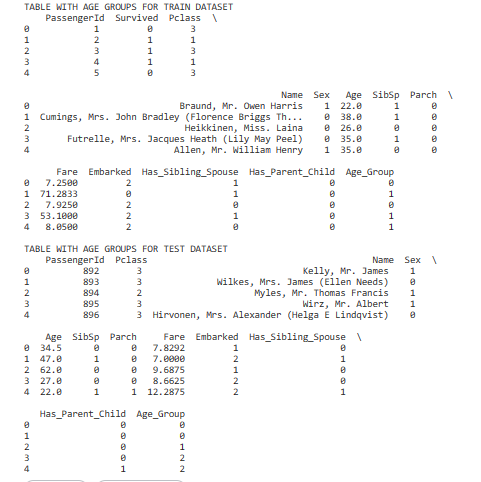
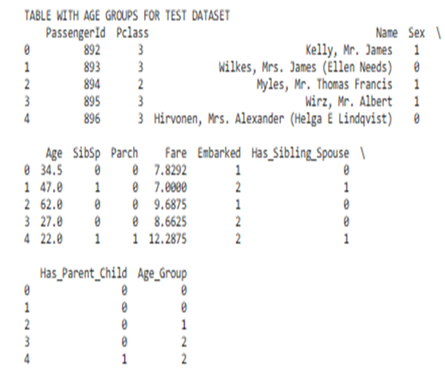
**7.3 Extracting from Age - Adding Age\_Group**





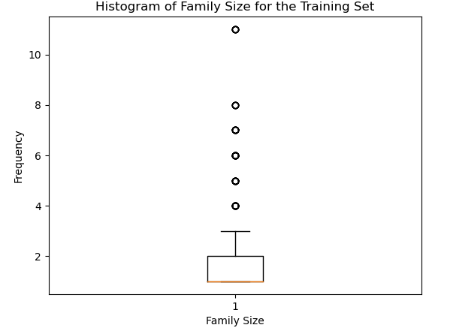


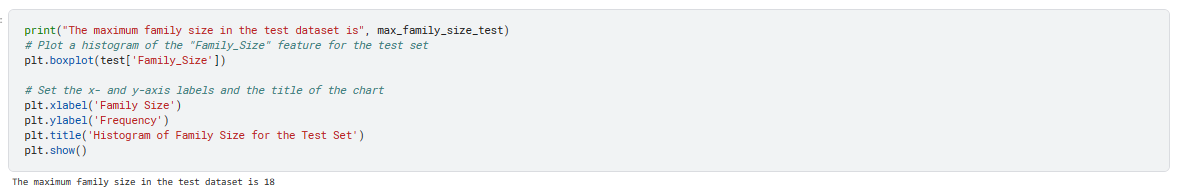


****

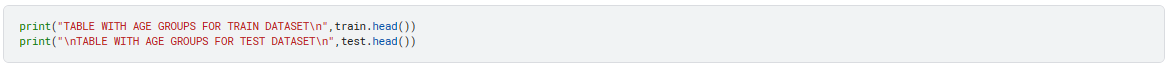
**7.4 Extracting from Age\_Group from Age**

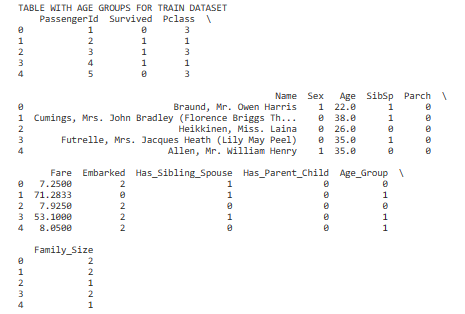
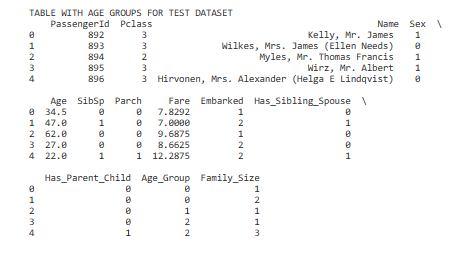




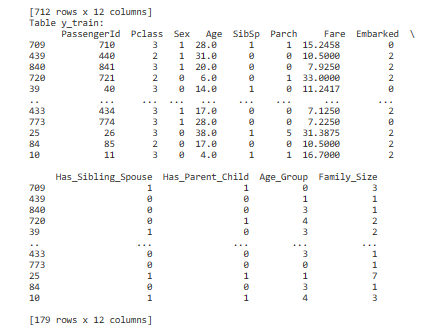
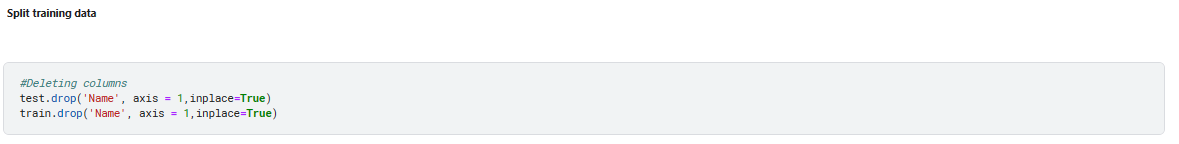


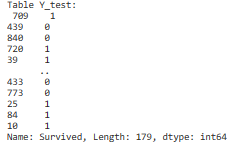
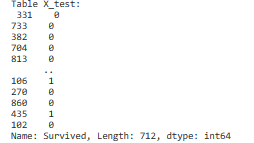






**8. Apply clustering and classification models**

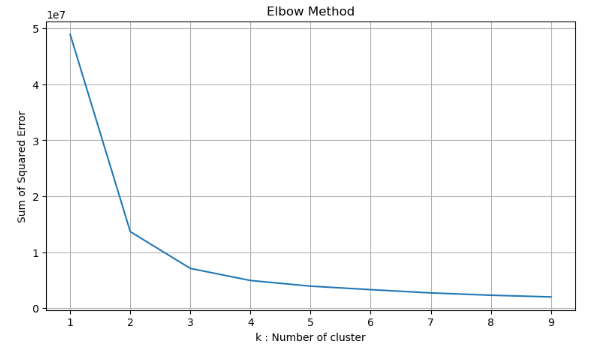




**8.1 CLUSTERING MODELS**

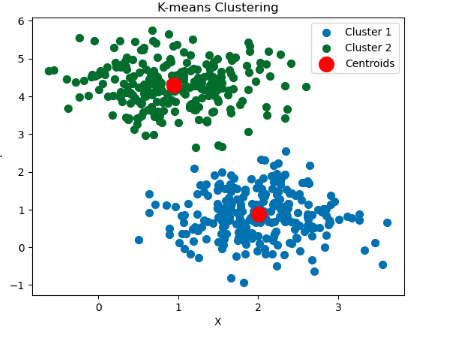
**8.1.2 k-means clustering**



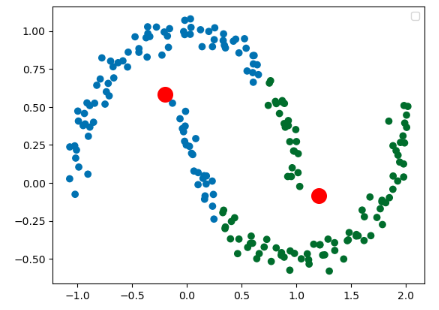


From graph, we obtain the optimal value of K=2

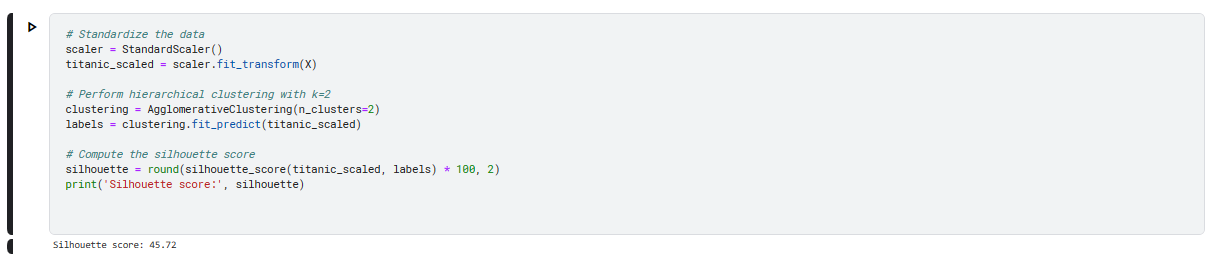


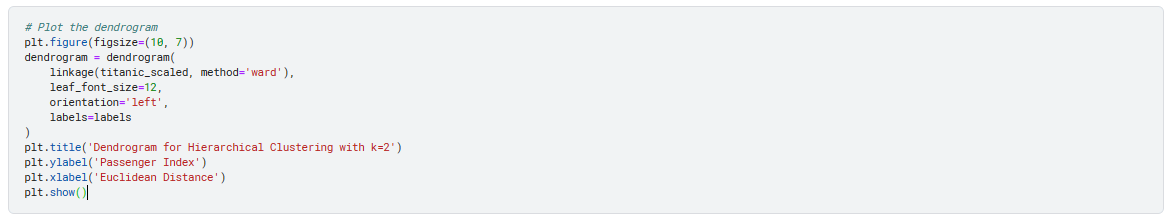


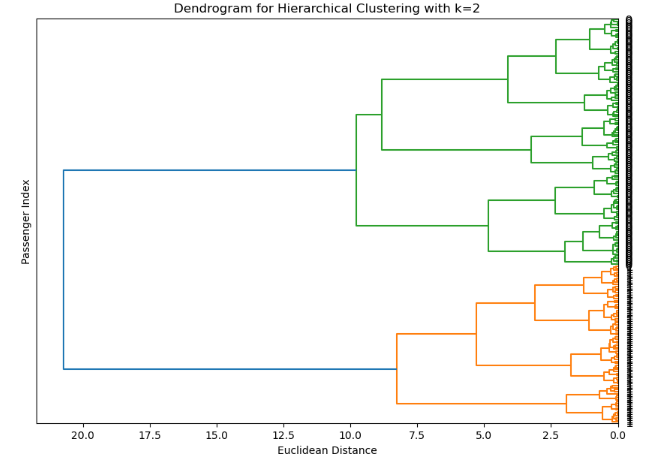


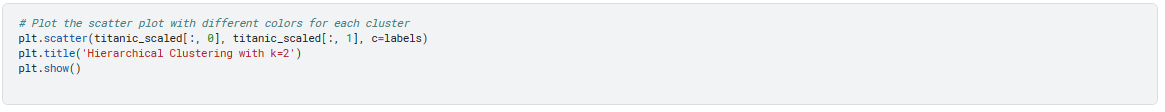


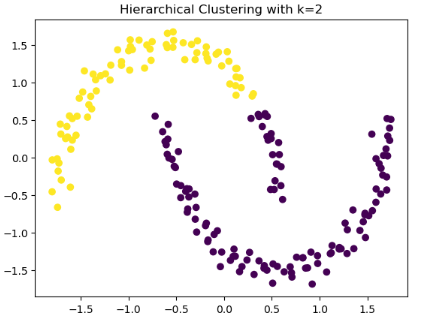
**8.2.2 Hierarchical Clustering**



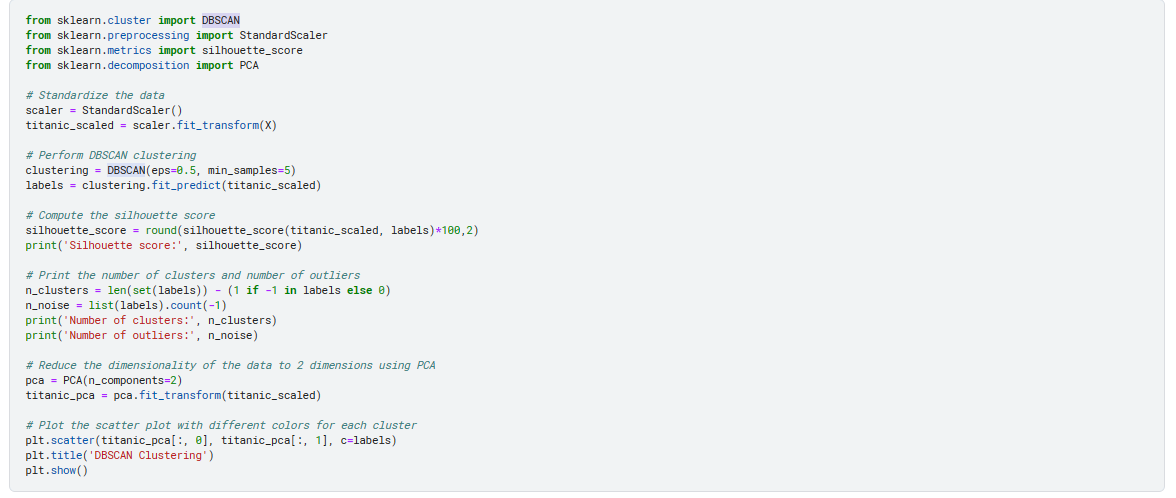


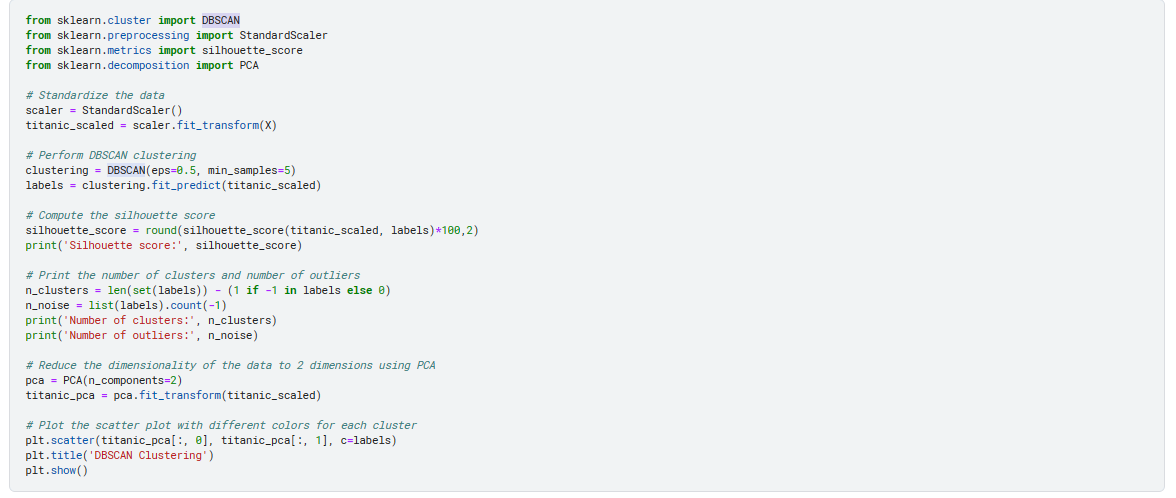


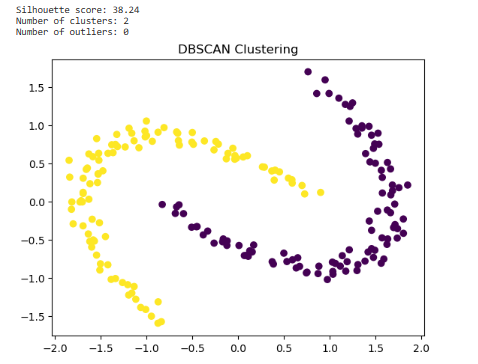




**8.1.3 DBSCAN Clustering**





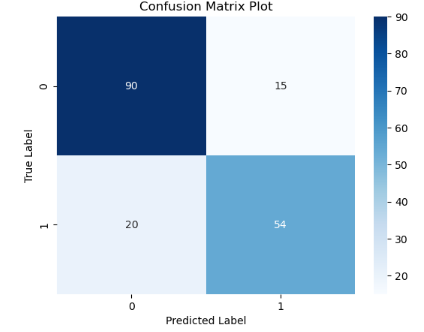
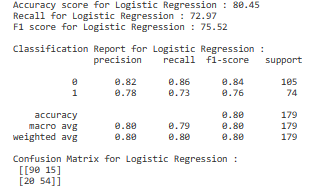


As the moon-shaped data in k-means is not ideally convex, we are exploring alternative models such as hierarchical and DBSCAN. It was discovered that the Hierarchical clustering yields a superior silhouette score of 45.2%. However, due to the presence of predetermined variables and labels within the dataset, we will opt for supervised learning when modelling.

**8.2 CLASSIFICATION MODELS**

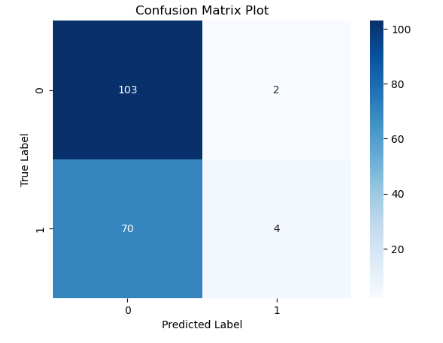
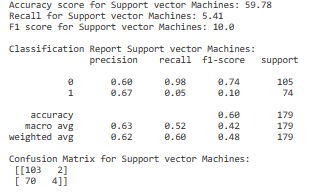
**8.2.1 Logistic Regression**



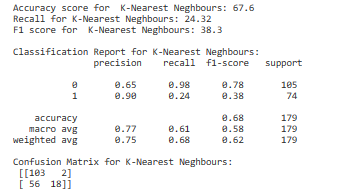


**8.2.2 Support Vector Machine**

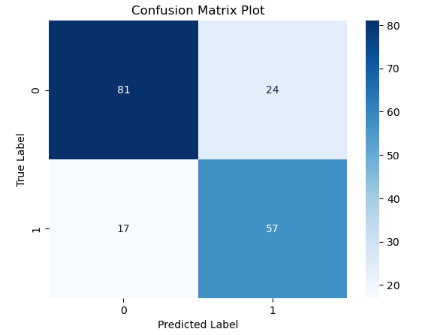
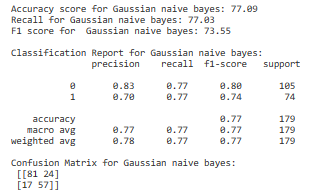




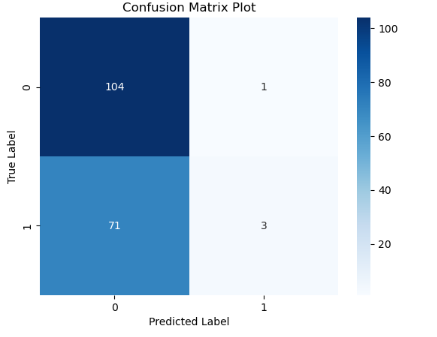
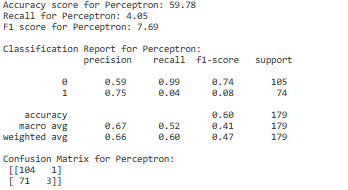


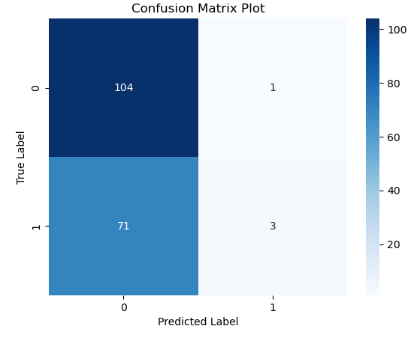
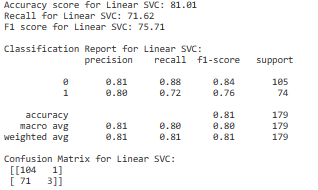




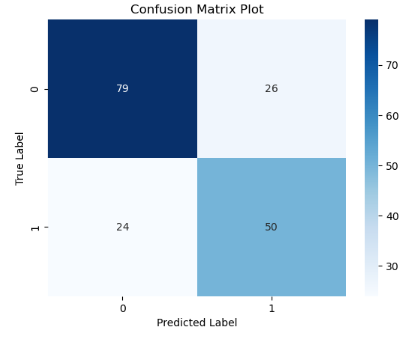
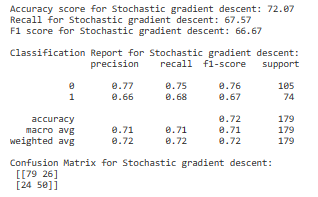




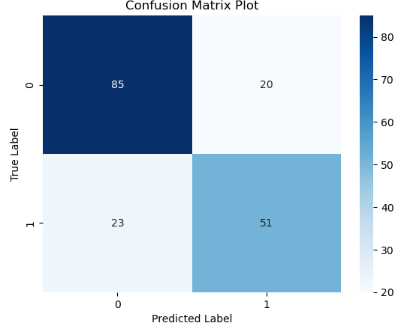
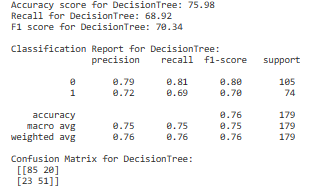




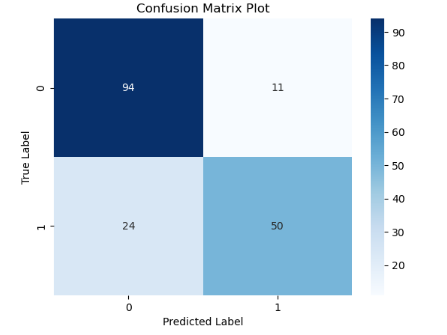
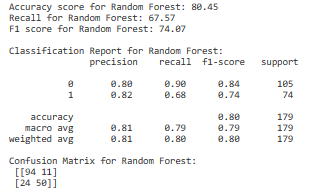




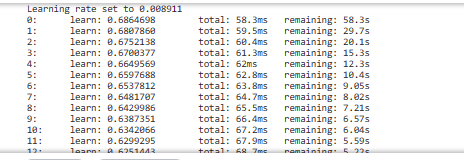


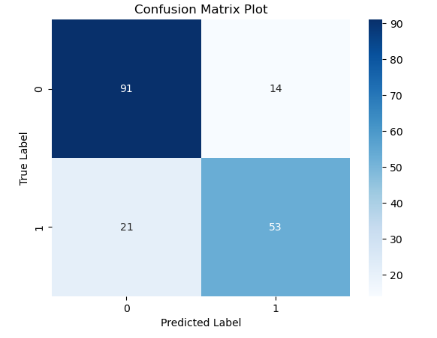


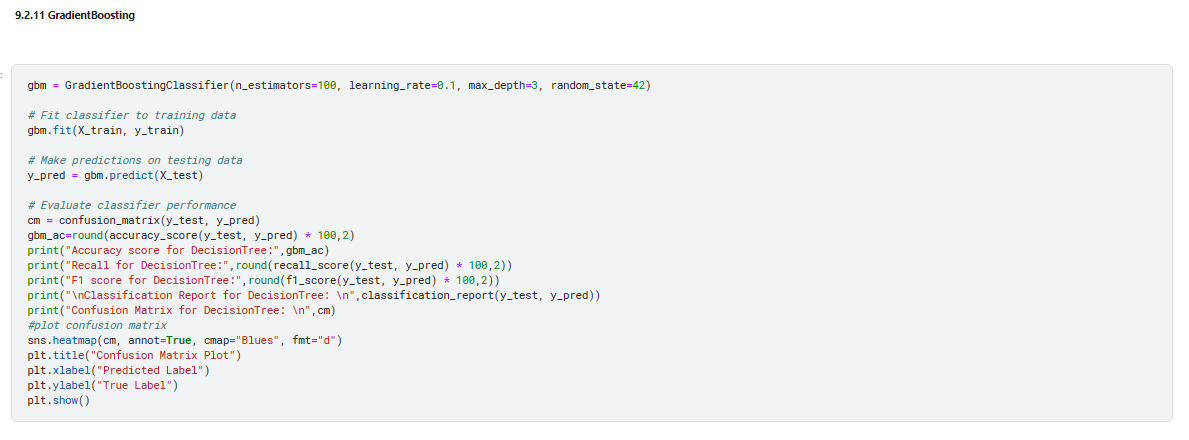


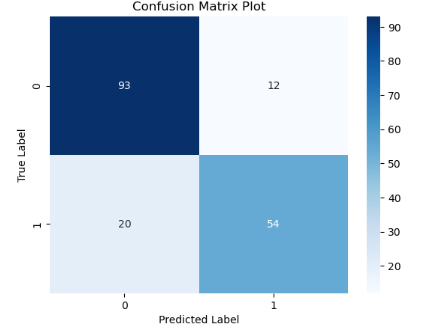
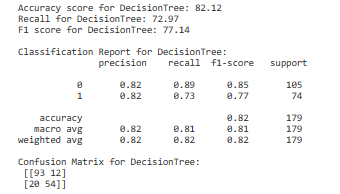








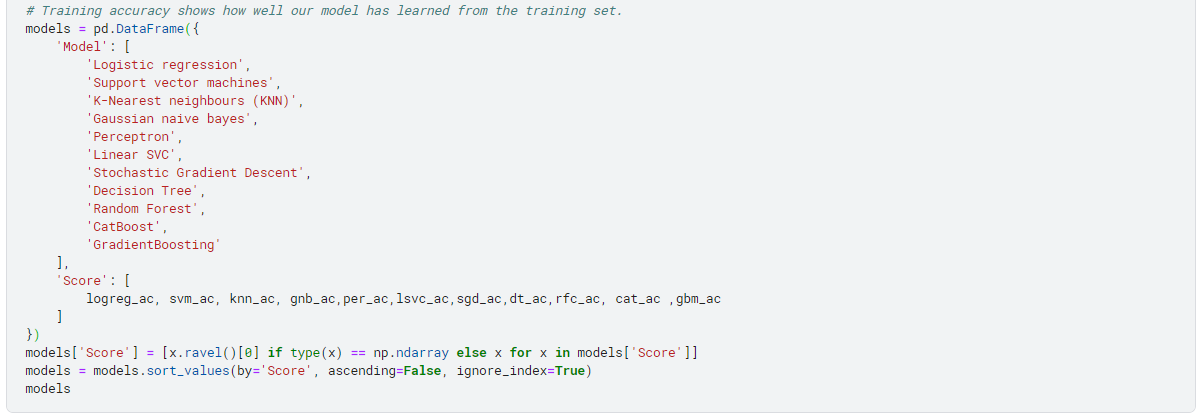


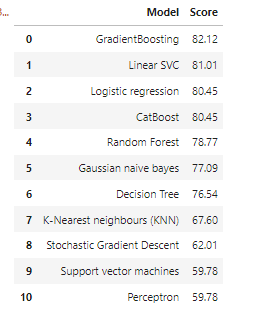


**9. EXPERIMENTAL ANALYSIS RESULT(Model evaluation)**



Training accuracy shows how well our model has learned from the training set.

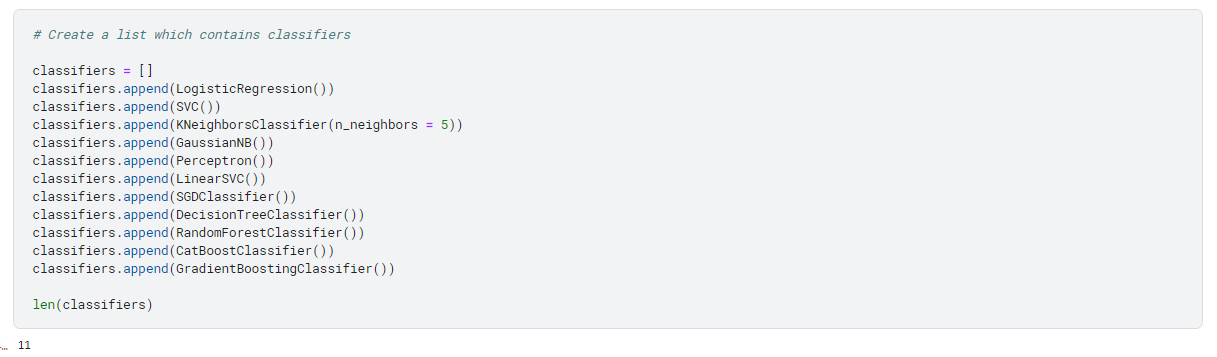


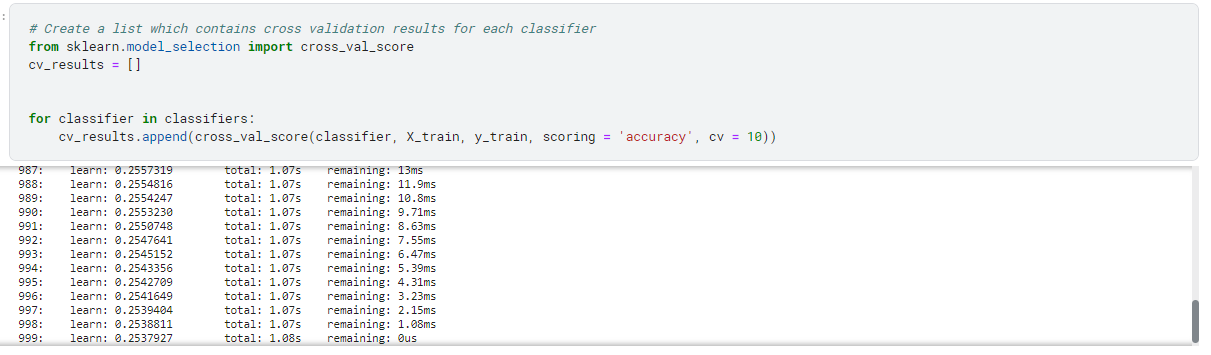


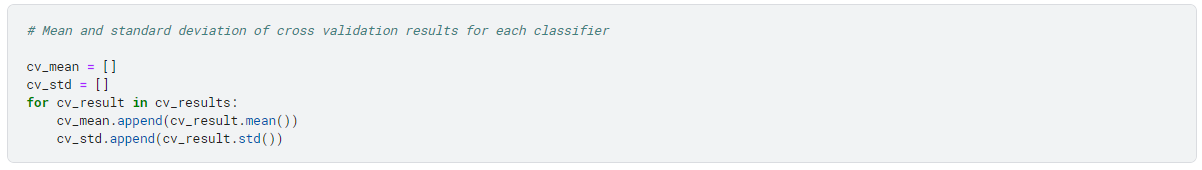
Highest accuracy score for **GradientBoosting Classifier**.

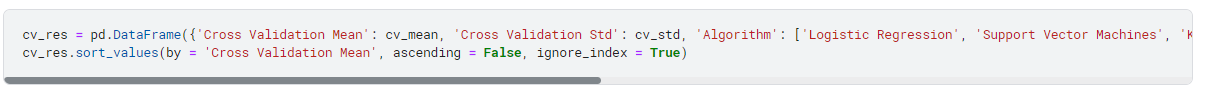
**9.2 K-fold cross-validation**

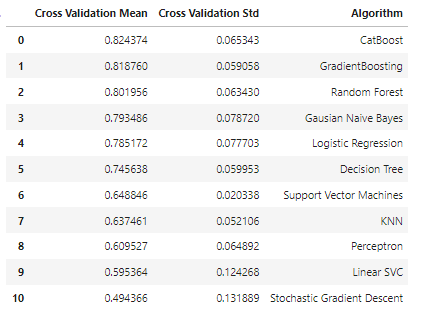
More emphasis should be on the model's capacity to **forecast data outside of the sample** because we cannot solely depend on training accuracy therefore K-fold cross-validation & feature change method done.



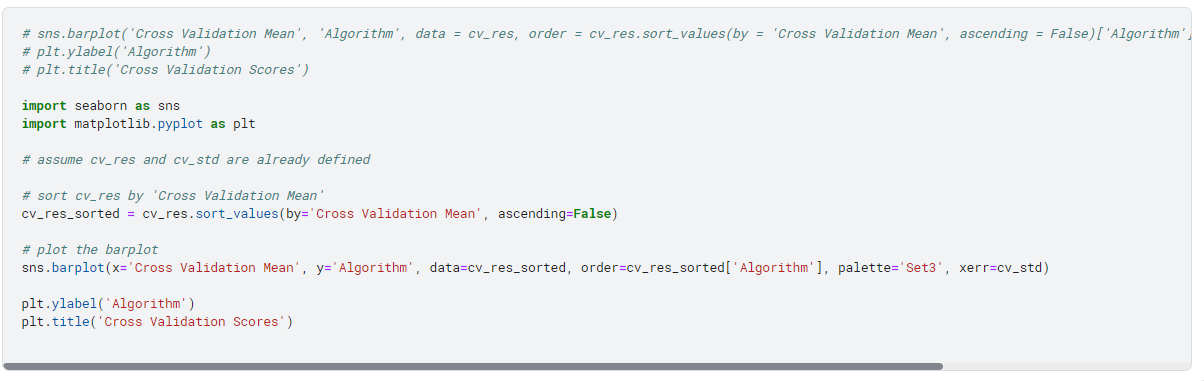


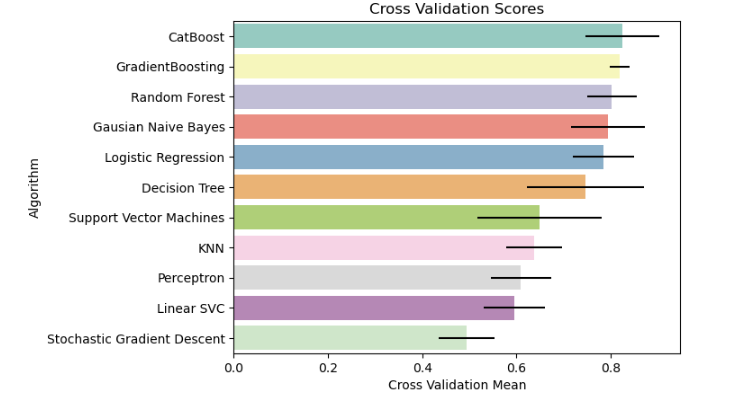






The highest cross-validation means for **CatBoost**.





The **CatBoost Classification** model leads over the others making it the best for my project.

**10. MODEL SELECTION**

I have selected the classification model (supervised learning) rather than clustering (unsupervised learning) for the dataset due to **labeled data availability.** Gradient Boosting had the highest accuracy score, but **CatBoost was ultimately selected** as the best model after further evaluation as it showed flexibility with handling changes to feature data and k-fold cross-validation as the results showed that the model is not overfitting to the training data and has a good level of generalization.

**FUTURE SCOPE**

* Do more Feature Engineering to the dataset to check if the accuracy\_score would increase
* Include features like Ticket and Name to do the modelling.
* Optimising the model hyper-parameters with GridSearchCV
* Applying techniques like boosting, stacking to the dataset.

**DISCUSSION AND CONCLUSIONS**

The main insights drawn from the analysis and experiments on Titanic Dataset are:

* The survival rate on the Titanic was low.
* Gender was a significant factor in survival - females had a much higher survival rate than males.
* The highest number of passengers onboard were aged between 20 – 30.
* Age played a role in survival: Children and seniors had a higher survival rate than adults.
* Passengers in higher classes had a higher survival rate than those in lower classes.
* Passengers with family members aboard had a higher survival rate than those alone.
* The ticket fare was also a predictor of survival: passengers who paid higher fares had a higher survival rate.
* The embarkation port seemed to have some correlation with survival: passengers who embarked in Cherbourg, France, had a higher survival rate than those who embarked in Southampton or Queenstown.

To improve the results during my implementation, I used various data pre-processing techniques, such as filling in NaN and converting categorical into numerical variables. Additionally, I performed feature engineering to extract new variables based on existing ones to improve the model's predictive accuracy. I also used different machine-learning algorithms to find the best-performing model. Finally, I evaluated the performance of the models using cross-validation and analysed their feature importance to gain insights into the factors that contribute to survival. Through this process, I was able to select the model CatBoost as the best model for my Survival Prediction.

**REFERENCES**

[1] Aljaber, N. (2018). Predicting Survival on the Titanic using Logistic Regression, Decision Trees and Random Forests. Journal of Physics: Conference Series, 1013(1), 012017.

[2] Çakmak, B. (2019). An Analysis of Feature Importance in Titanic Dataset. In 2019 7th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[3] Erol, H. (2020). Improving the Accuracy of Survival Prediction for Titanic Passengers Using Support Vector Machine. In 2020 8th International Istanbul Smart Grid and Cities Congress and Fair (ICSG) (pp. 1-6). IEEE.

[4] Kuru, S. T., & Orhan, H. (2019). Comparison of Imputation Techniques on Titanic Dataset. In 2019 7th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.