**Analysis Report**

**a. Code Details:**

1. Data Loading and Initial Exploration:

Libraries are imported, and the file path is set.

The Titanic dataset is loaded into a Pandas DataFrame named titanic\_ujjwal.

Initial exploration includes printing the first 3 rows, shape (rows and columns), and dataset information.

2. Initial Exploration:

The code identifies uninformative columns ('PassengerId', 'Name', 'Ticket', 'Cabin') for future removal.

3. Data Visualization:

Stacked bar charts are created to visualize the relationship between 'Pclass' and 'Survived' and 'Sex' and 'Survived'.

A scatter matrix is generated for features 'Survived', 'Sex', 'Pclass', 'Fare', 'SibSp', and 'Parch'.

4. Data Preprocessing:

Uninformative columns are removed from the dataset.

Categorical columns ('Sex' and 'Embarked') are one-hot encoded and added back to the dataset.

Missing values in the 'Age' column are filled with the mean age.

Data types are converted to float for consistent modeling.

5. Data Normalization:

A function, normalize\_dataframe(), scales features to the 0-1 range.

The dataset is normalized using this function, and the result is stored in titanic\_ujjwal\_final.

6. Model Training and Testing:

Data is split into features (X) and the target variable (y).

A logistic regression model is trained on the training data.

The dataset is split into training and testing sets using a 70-30 ratio.

7. Cross-validation:

A loop iterates through different test sizes (10% to 50% in 5% increments).

For each test size, the model is trained and evaluated using 10-fold cross-validation.

Minimum, mean, and maximum accuracy scores are calculated.

The best test size with the highest mean accuracy is recommended.

8. Model Evaluation:

The logistic regression model is rebuilt using the recommended test size (30% test data, 70% training data).

Predicted probabilities for survival are calculated using the predict\_proba() method.

Two sets of predictions are made based on different thresholds (0.5 and 0.75).

Accuracy, confusion matrices, classification reports, precision, and recall are calculated for each threshold.

9. Analysis and Comparison:

The code compares model performance at different thresholds.

It assesses accuracy, precision, and recall at both 0.5 and 0.75 thresholds.

It determines whether accuracy is higher on the test or training data.

It makes comparisons for precision and recall between the two threshold values.

**b. Non useful Columns:**

a. passengerId : It has 891 unique ids. It's not useful for predicting anything.

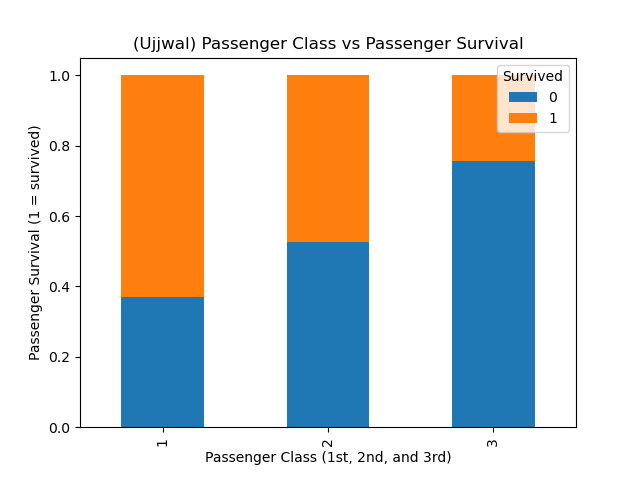
b. Name: It is similar to passengerId, has most name as unique, not useful for any predictions.

c. Ticket: Tickets does also contains mostly unique values. It does not provide meaningful predictive power.

d. Cabin: It has lot of missing values. Which makes it complex to use for predictions.

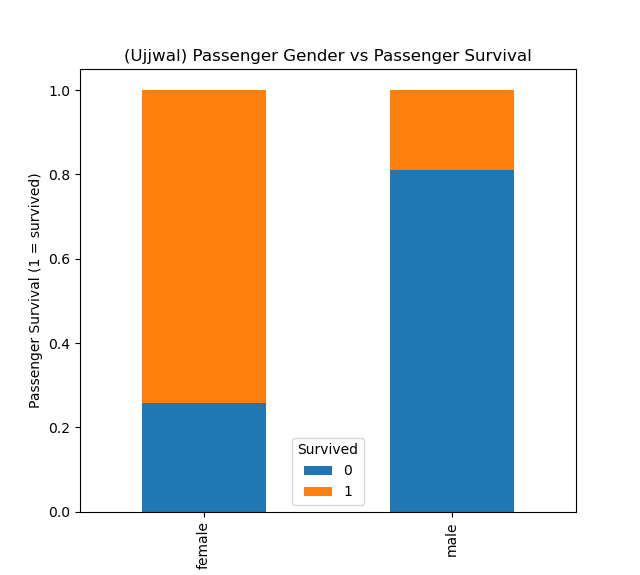
**c. Charts Analysis:**

1. Passenger Class vs Passenger Survival:



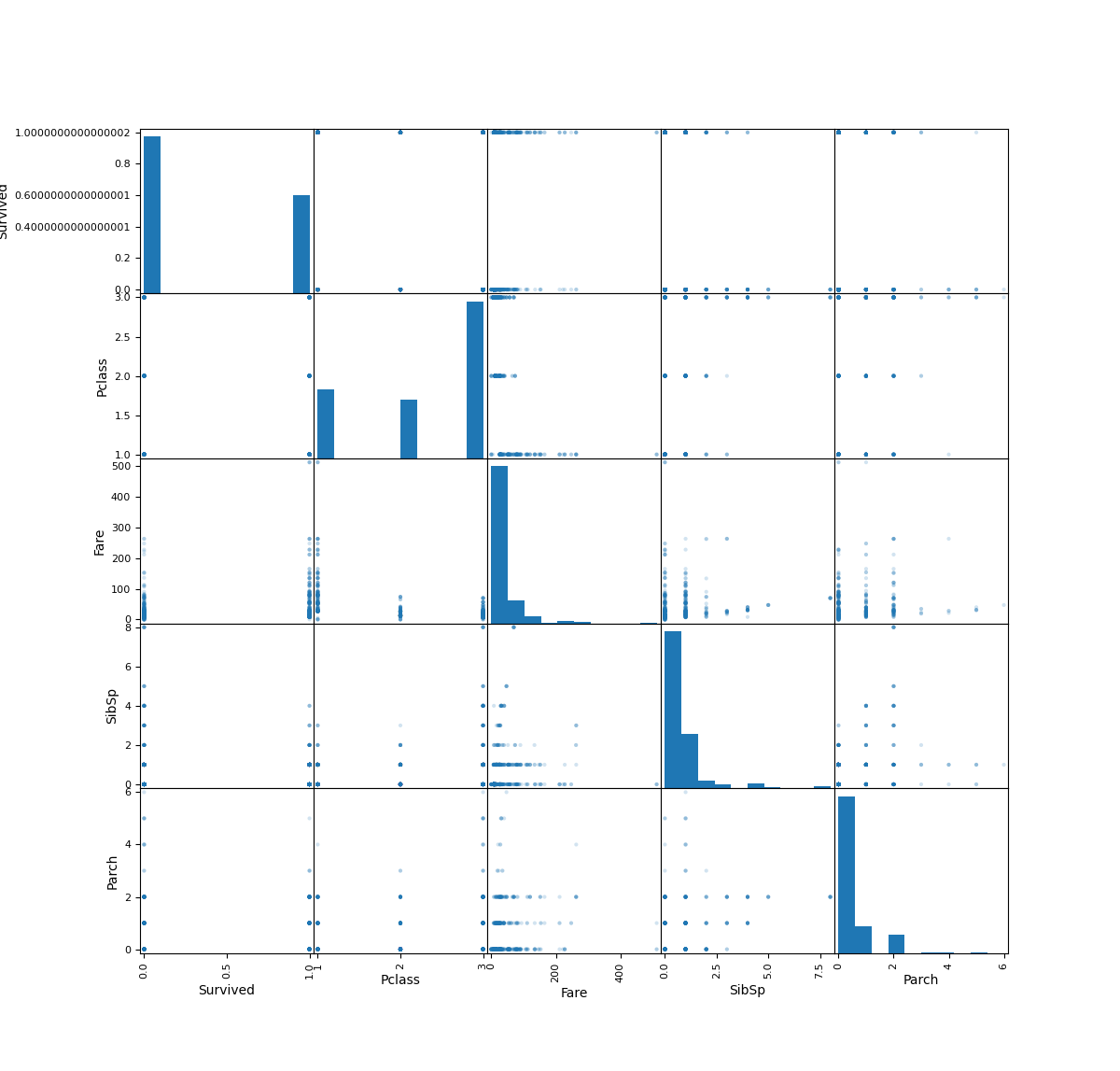
This chart shows the number of passengers survived are higher on the greater classes. For example, the proportion of passengers who were on first class were higher than second class and the proportion of passengers on second class who survived were higher than the third class.

2. Passenger Gender vs Passenger Survival:



This chart shows that the number of female survivors is higher than the male survivors.

3. Graphs of all columns:



The graph reveals insights into the relationships between passenger class, gender, and survival, it also displays potential correlations. Moreover, it provides a snapshot of ticket fare distribution across passenger classes and offers an overview of family structures on the Titanic.