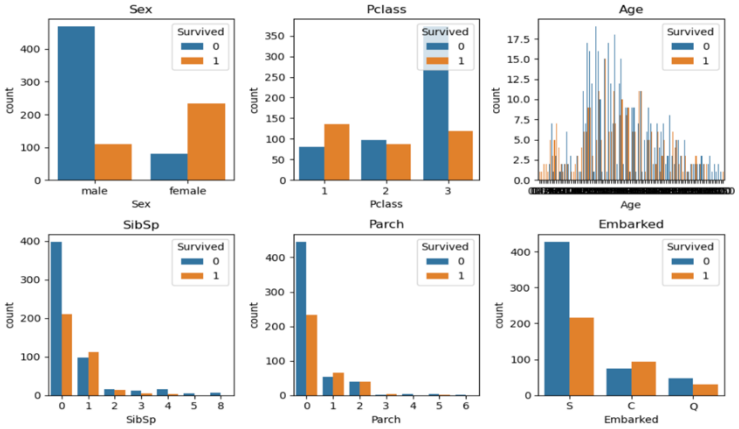
Machine learning model for the
TITANIC DATASET

The objective of this project is to develop a machine learning model that predicts the survival outcome of passengers aboard the Titanic. The Titanic dataset consists of various features such as passenger class, age, gender, and fare, which can be used to train predictive models.

Gitlab link: https://gitlab.com/Suvetaa/titanictraining.git

Description of the algorithm:

After importing all the necessary libraries such as Panda, Seaborn and Matplotlib, I did some data exploration using info(), head() to have a first overview of the dataset. I then use some other function such as describe(), value_counts(), and subplot() to know a little more better about each variable and its impact on the survival rate of a random passenger.

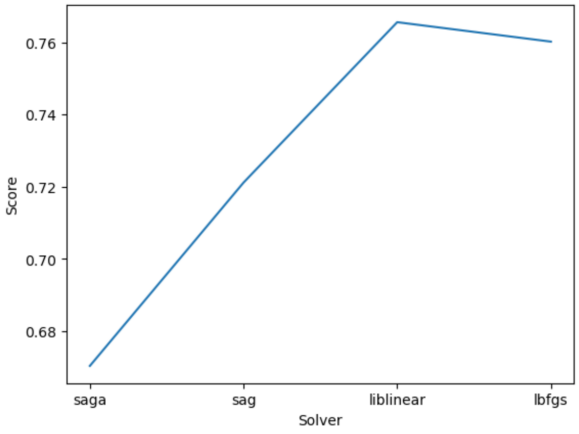


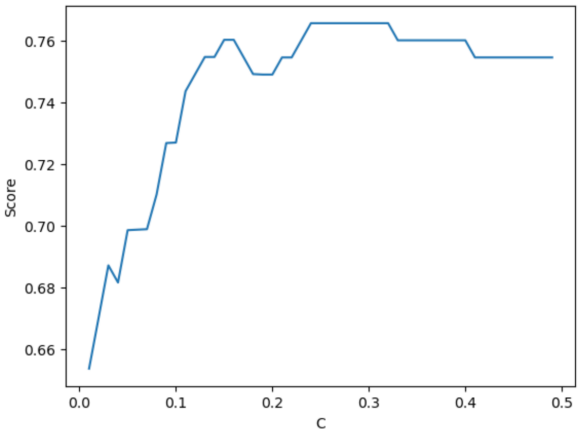
Here is the example of the graphs obtained.
This exploration helped me to understand that variable such as Name, Passenger, Ticket, Cabin did not have a big impact on the survival rate. So I drop those variables from my table.

Secondly to improve the model performance I preprocessed the data. For example, I observed missing values in the 'Age' and 'Embarked' columns. To address this, I replaced missing 'Age' values with the mean age of passengers and filled missing 'Embarked' values with the mode value. Then, the 'Sex' and 'Embarked' columns contained string values. So, I replaced the string values with numeric representations (0 and 1 for 'male' and 'female', and 0, 1, and 2 for 'S', 'C', and 'Q' in 'Embarked') to make them compatible with the machine learning algorithms.

I then split the dataset into training and testing dataset and mainly used the Logistic Regression to train the model as the output is binary (in this case, survival or not survival) and improve its performance by tunning the hyperparameters. I also tried to use another algorithm which is Random Forest to see if the testing accuracy is better. The final model, based on the best hyperparameters, was evaluated on the test dataset. The accuracy score was calculated to measure the final model's performance.

Improvement of the model performance:





To improve the model performance, I first did a model by performing hyperparameter tuning using a validation curve. I varied the "solver" hyperparameter and found that the "liblinear" solver provided the best performance. Additionally, I tuned the "C" hyperparameter and found that a value of 0.26 yielded optimal results. The tuned Logistic Regression model achieved a training accuracy of 80.5% and a test accuracy of 80.4%.

To further improve the model performance, I performed extensive hyperparameter tuning using a grid search. I explored a range of hyperparameters, including the ‘solver’, the ‘max_iter’ and ‘C’ and create a model using the best parameters given by the CvGrid.

The test dataset of the Titanic is a crucial component of the analysis performed. It serves as an independent set of data that was not used during the training phase of the models. After preprocessing, the trained models are used to predict the survival outcomes of the passengers in the test dataset. These predictions are then compared to the ground truth values provided in the "gender_submission.csv" file. The accuracy of the predictions is calculated using the accuracy_score metric for the three different model to show how the accuracy score got increased throught the project.

First model confusion matrix

	Predicted No	Predicted Yes
Actual No	247	19
Actual Yes	12	140

```
Training accuracy: 0.9258373205741627
```

Final model confusion matrix

	Predicted No	Predicted Yes
Actual No	258	8
Actual Yes	9	143

```
Training accuracy: 0.9593301435406698
```