Titanic Data Analysis & Prediction by Balancing Data

1. Introduction

1.1 Introduction to Data Set.

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

Variable Name	Description	Sample Data
PassengerId	ID of passenger (unique ID)	1; 2; ...
Survived	Survival status (0 = No, 1 = Yes)	0; 1; ...
Pclass	Ticket class (1 = 1st/Upper, 2 = 2nd/Middle, 3 = 3rd/Lower)	1; 3; ...
Name	Passenger name (unique)	Braund, Mr. Owen Harris; Heikkinen, Miss. Laina; ...
Sex	Passenger gender (male or female)	male; female; ...
Age	Passenger age (in years)	22; 38; ...
SibSp	No of siblings / spouses aboard the Titanic	0; 3; ...
Parch	# of parents / children aboard the Titanic	1; 2; ...
Ticket	Ticket number (unique)	A/5 21171; PC 17599; ...
Fare	Passenger fare	7.25; 71.2833; ...
Cabin	Cabin number	C85; C123; ...
Embarked	Embarkation port (C = Cherbourg, Q = Queenstown, S = Southampton)	C; S; ...

This dataset has been referred from Kaggle: https://www.kaggle.com/c/titanic/data

1.2 Problem Statement.

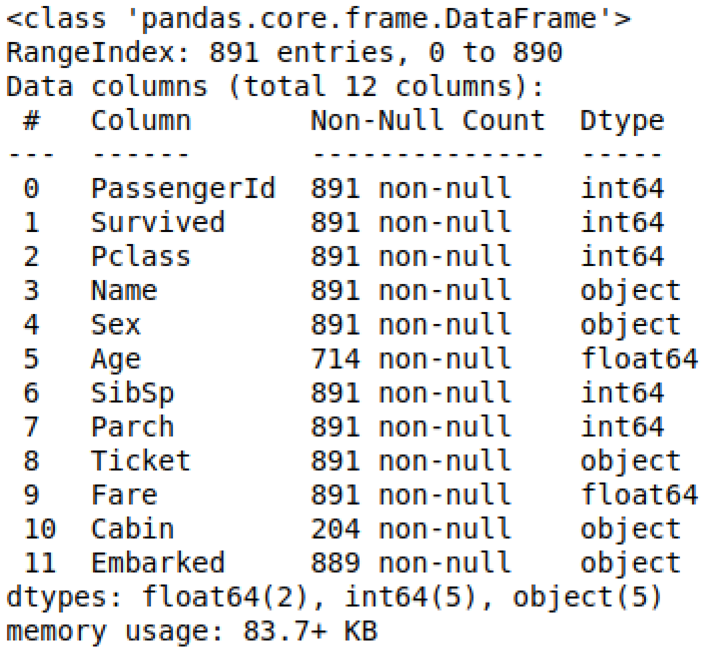
In this challenge, first we’re going to perform exploratory analysis to fetch the insights from the data. And second, we are going to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e. name, age, gender, socio-economic class, etc.).

2. Goal of the Project

- Understand the Dataset & clean it and perform exploratory analysis.
- Dataset exploration using various types of data visualization.
- Feature Selection
- Build classification model to predict weather the passenger survives or not.
- Use different classification model to check the accuracy.
- Also fine-tune the hyper-parameters & compare the evaluation metrics of various classification algorithms.

3. Basic Analysis:

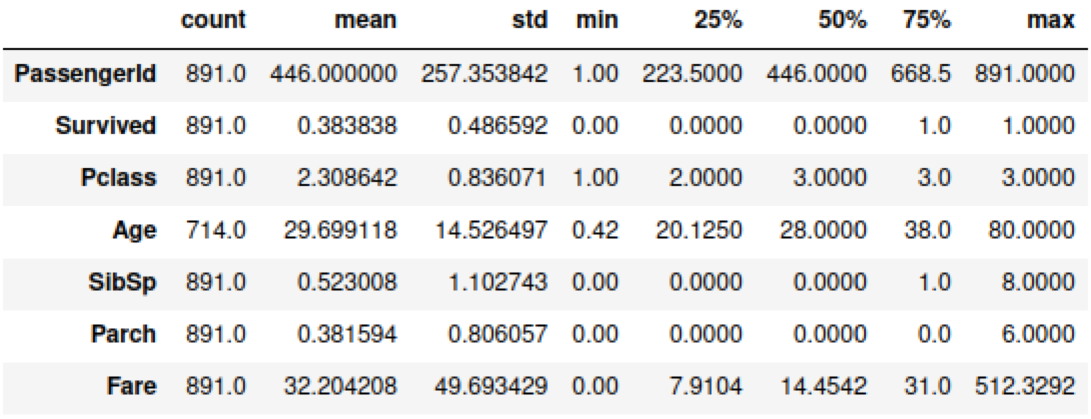
There are 891 numbers of rows and 12 numbers of columns.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

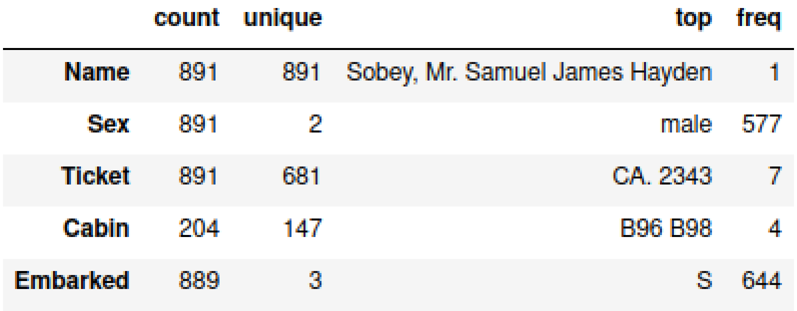
- From the above observation we can see there are some null values in few columns.
- Let see how many exact null values in which columns

Numerical Summary:



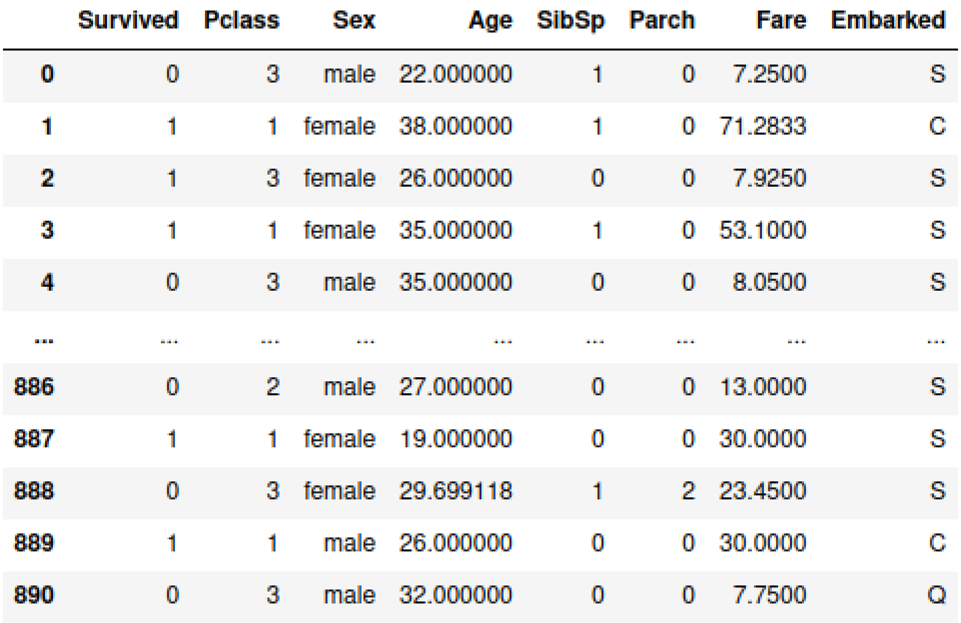
	count	mean	std	min	25%	50%	75%	max
PassengerId	891.0	446.000000	257.353842	1.00	223.5000	446.0000	668.5	891.0000
Survived	891.0	0.383838	0.486592	0.00	0.0000	0.0000	1.0	1.0000
Pclass	891.0	2.308642	0.836071	1.00	2.0000	3.0000	3.0	3.0000
Age	714.0	29.699118	14.526497	0.42	20.1250	28.0000	38.0	80.0000
SibSp	891.0	0.523008	1.102743	0.00	0.0000	0.0000	1.0	8.0000
Parch	891.0	0.381594	0.806057	0.00	0.0000	0.0000	0.0	6.0000
Fare	891.0	32.204208	49.693429	0.00	7.9104	14.4542	31.0	512.3292

Categorical Summary:



	count	unique	top	freq
Name	891	891	Sobey, Mr. Samuel James Hayden	1
Sex	891	2	male	577
Ticket	891	681	CA. 2343	7
Cabin	204	147	B96 B98	4
Embarked	889	3	S	644

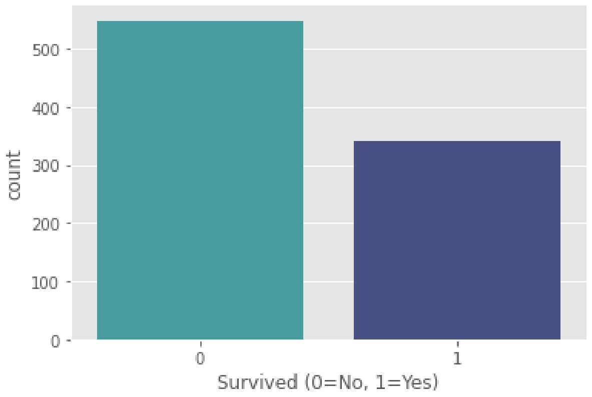
- From the above observation we can see that we have total 891 number of rows
- In the age column we have 177 null values and in the embarked column we have 2 null values.
- But in cabin we have 687 null values.
- We can remove missing values by filling the age with the mean of age columns from both train and test data
- And Embarked column with the mode of embarked column in the train data
- But it's not a good idea to fill a column with mean or mode which has more than 80 % null values.
- So, we'll remove cabin column from both train and test data
- Also remove passenger id, Name and Ticker because these are not going to help us in exploration and prediction
- Since Name, Ticket, and PassengerId contains unique data, these columns will be removed.

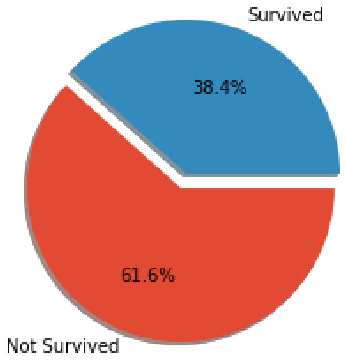


	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.000000	1	0	7.2500	S
1	1	1	female	38.000000	1	0	71.2833	C
2	1	3	female	26.000000	0	0	7.9250	S
3	1	1	female	35.000000	1	0	53.1000	S
4	0	3	male	35.000000	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.000000	0	0	13.0000	S
887	1	1	female	19.000000	0	0	30.0000	S
888	0	3	female	29.699118	1	2	23.4500	S
889	1	1	male	26.000000	0	0	30.0000	C
890	0	3	male	32.000000	0	0	7.7500	Q

4. Exploratory Analysis:

4.1 Survived Distribution:

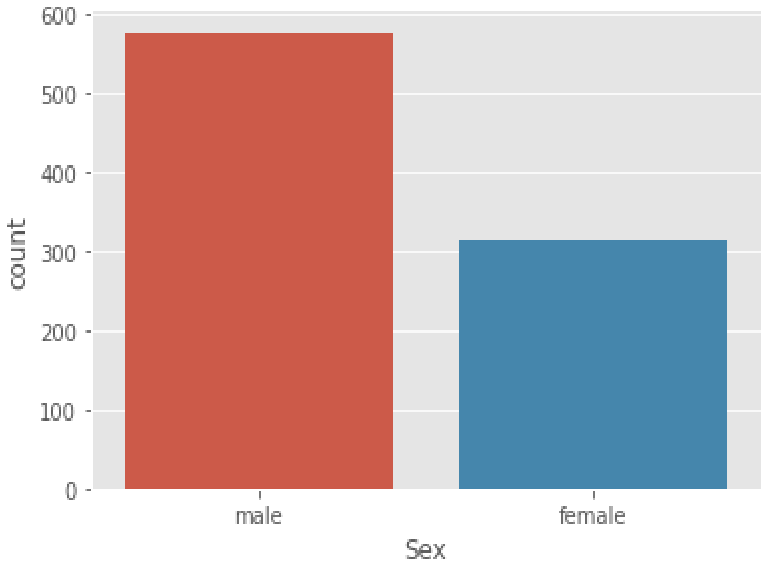




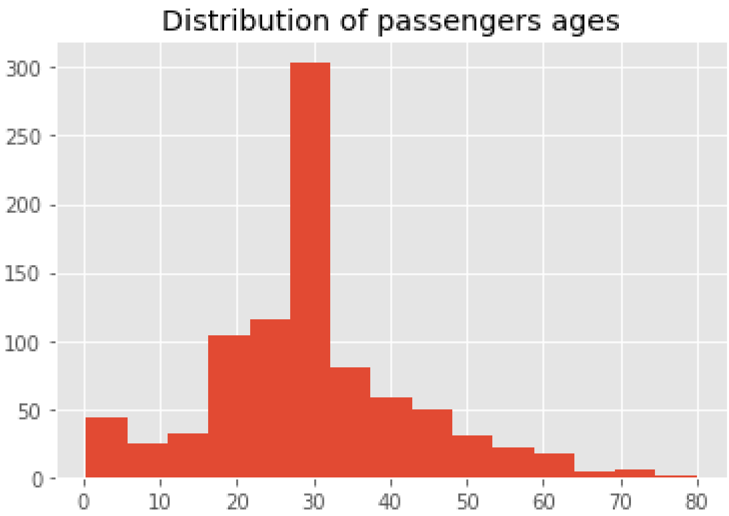
- Not Survive Percentage: 61.62%
- Survive Percentage: 38.38%
- It can be seen that **most passengers are not survived.**

4.2 Gender Distribution

- Female Percentage: 35.24%
- Male Percentage: 64.76%
- **The percentage of male passengers is higher** than female passengers.

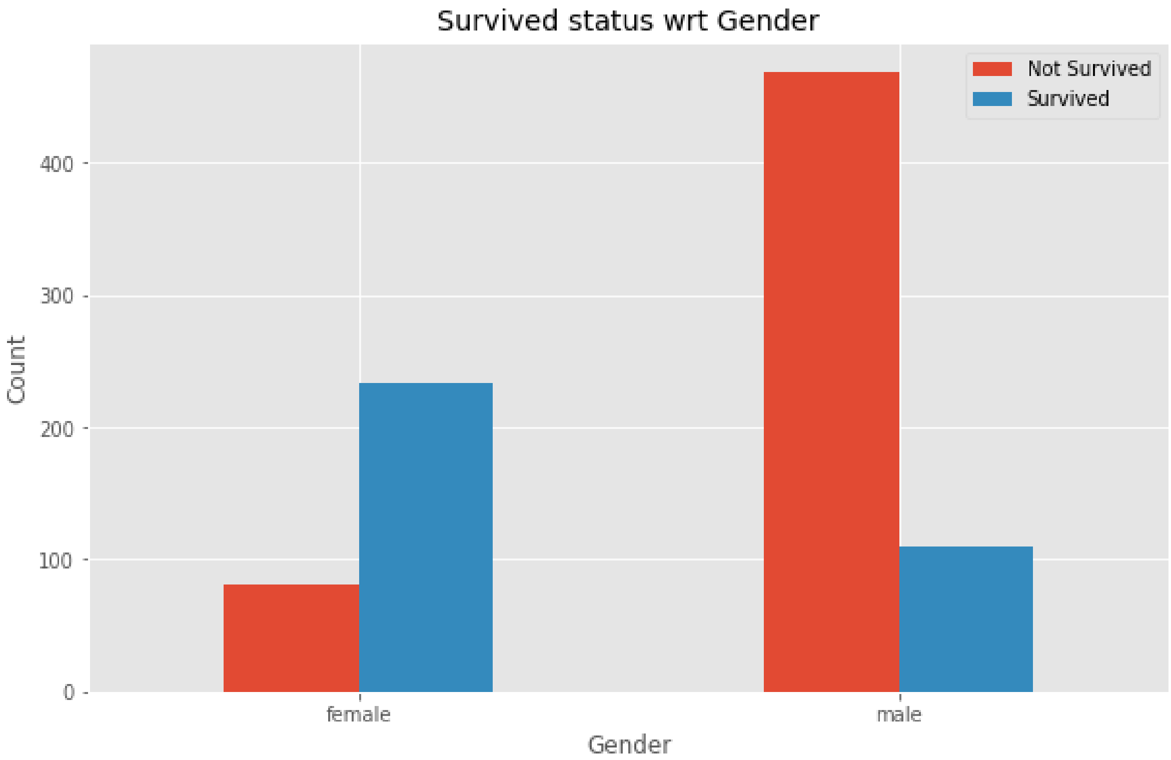


4.3 Age Distribution:



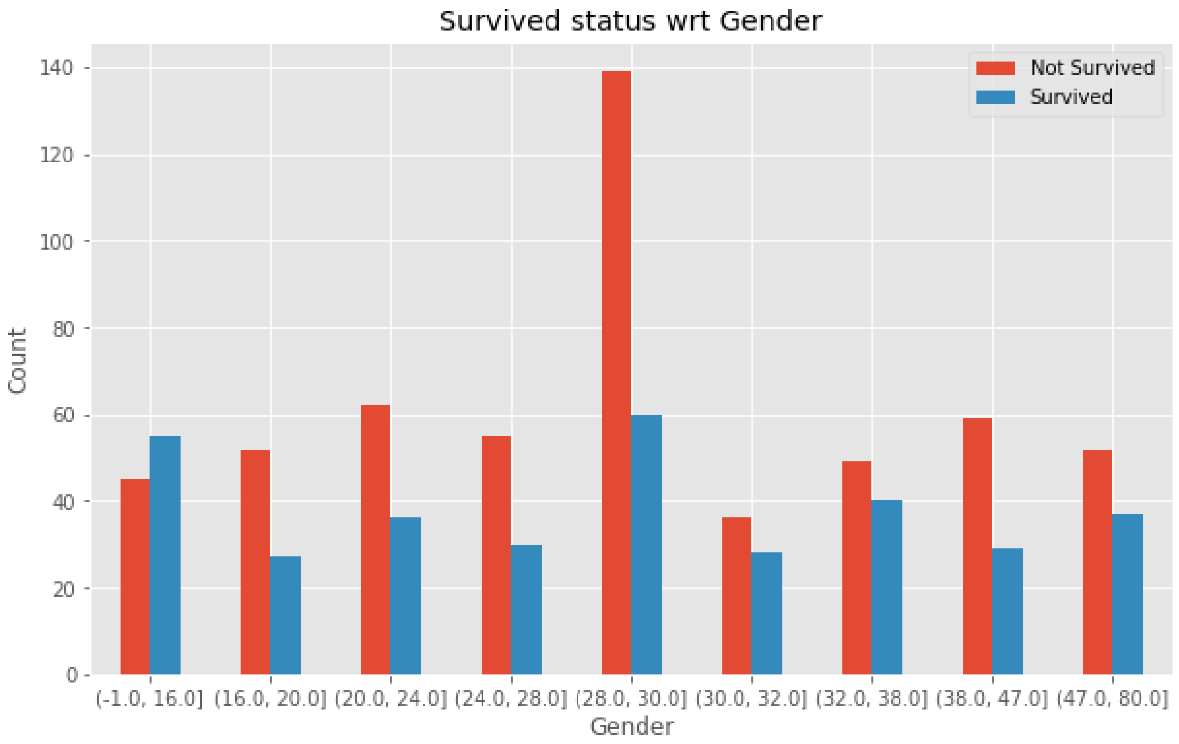
- The average age is approximately 30
- maximum age is near 80
- And minimum is 0.42 years old. It means may be some are kids and having age less than year.

4.4 Survived Distribution based on Gender:



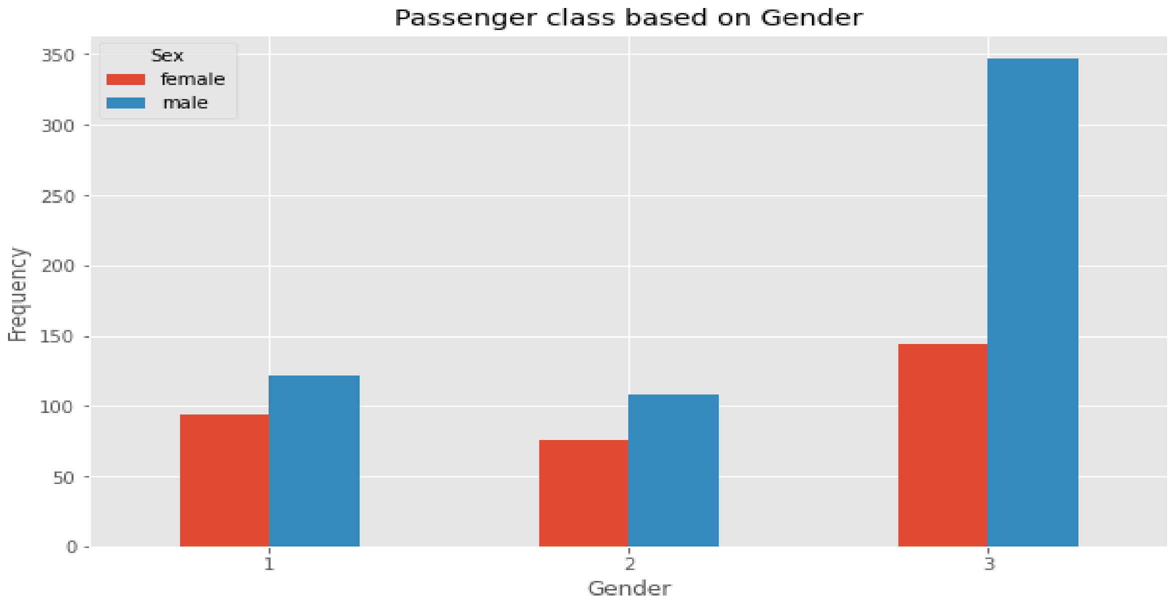
The number of female passengers that are survived is higher than male not survived.

7.5 Survived Distribution based on Age:



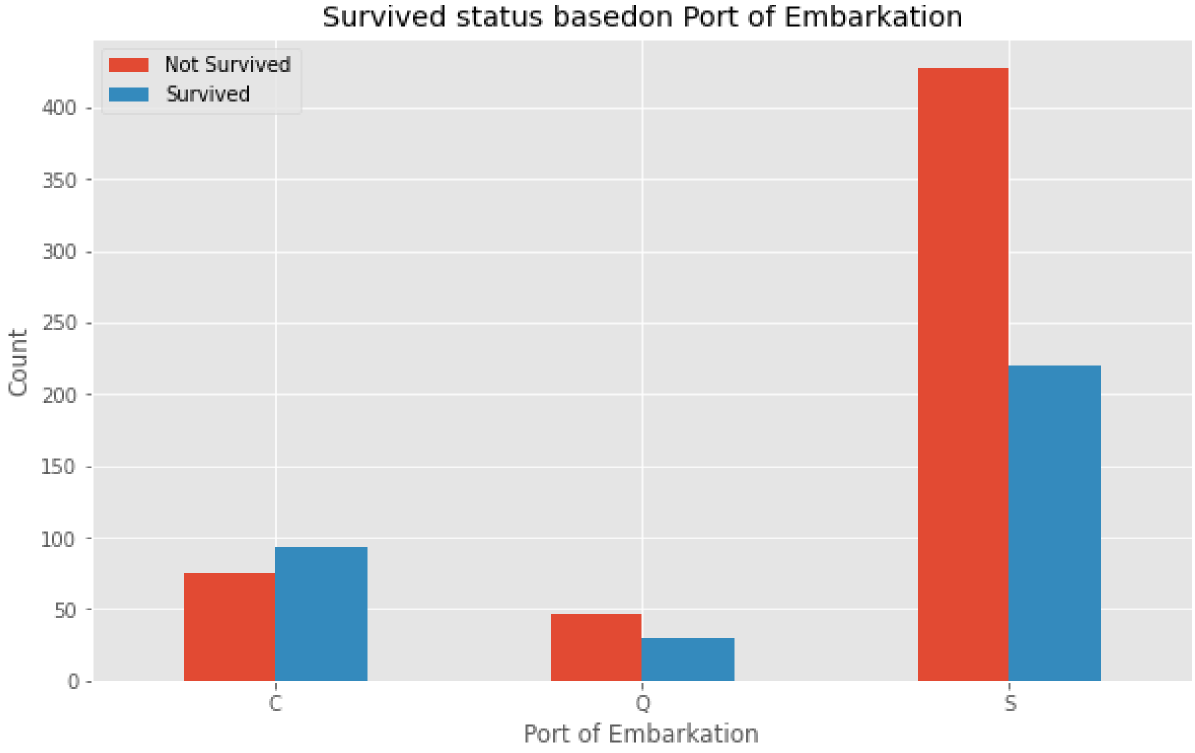
It can be seen that the majority of not survived passengers are between 16-30 y.o, quite similar to survived passengers in the same age range .Mostly people who survived have the age between 0 to 16 years old range.

4.6 Passenger Class based on Gender:



- Most of passengers from both genders prefer to choose 3rd class rather than other classes.

7.7 Survived Distribution based Port of Embarkation



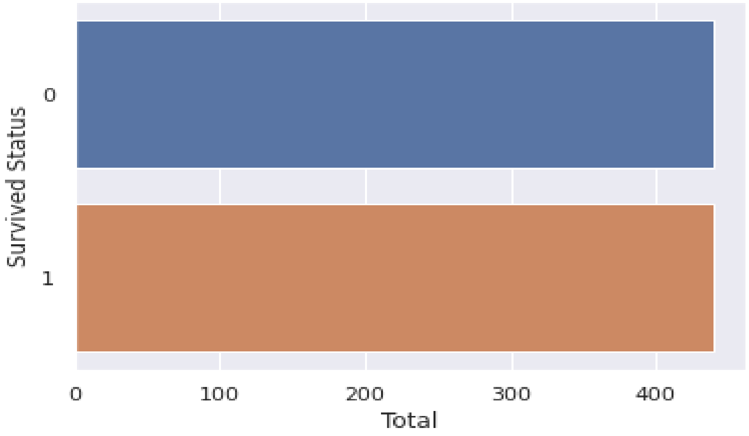
- C = Cherbourg, Q = Queenstown, S = Southampton
- Mostly survived and died people on Southampton port

5 Feature Engineering

The Feature Engineering method that used is one-hot encoding, which is transforming categorical variables into a form that could be provided to ML algorithms to do a better prediction.

5.1 Over Sampling

- Since the number of not survived passengers is more than survived passengers, oversampling is carried out to avoid over-fitting.
- The number of Not survived passengers are more than survived. So there is a chance that model will show the biasness.
- That's why we are going to handle imbalanced data.



- Now the data is balance we can see from the above graph that the survived and not survived ratio is same.

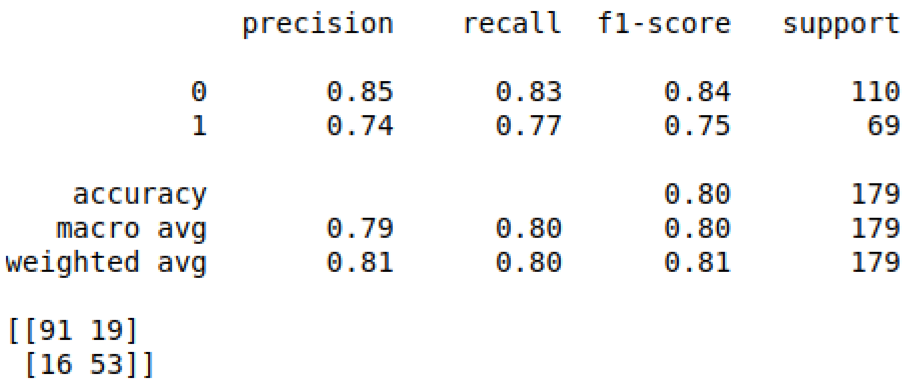
6. Model Building:

The machine learning models used in this project are:

- Logistic Regression
- SVC
- K Neighbors Classifier
- Decision Tree
- Random Forest
- Gradient Boosting

6.1 Logistic Regression

In the regression model, the model which is used for classification is logistic regression model. We’re going to implement it. And let see the accuracy Score.



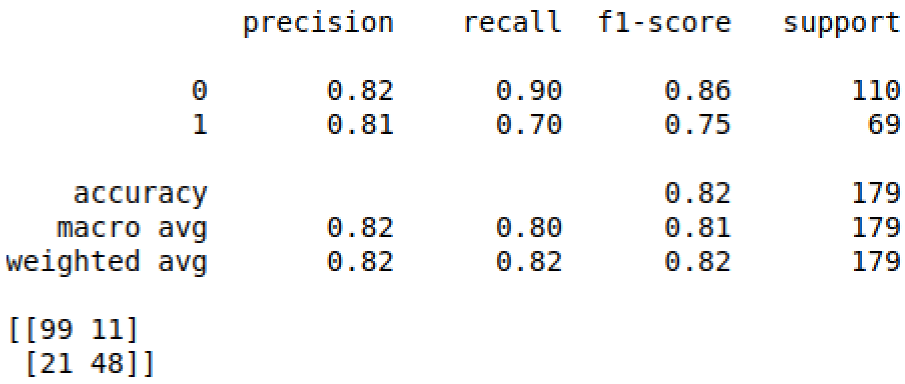
```
              precision    recall  f1-score   support

           0       0.85      0.83      0.84       110
           1       0.74      0.77      0.75        69

    accuracy                           0.80       179
   macro avg       0.79      0.80      0.80       179
weighted avg       0.81      0.80      0.81       179

[[91 19]
 [16 53]]
```

- Logistic regression accuracy: 80.45%

6.2 Decision tree Classifier

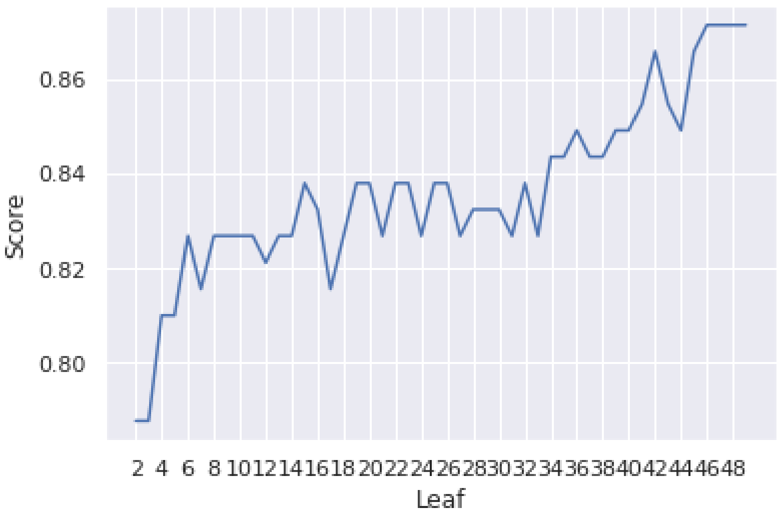
The second model which I'm going to use is Decision tree Classifier.

- I have randomly selected the max_leaf_nodes attribute to 10.
- Let's see the accuracy



```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86       110
           1       0.81      0.70      0.75        69

    accuracy                           0.82       179
   macro avg       0.82      0.80      0.81       179
weighted avg       0.82      0.82      0.82       179

[[99 11]
 [21 48]]
```

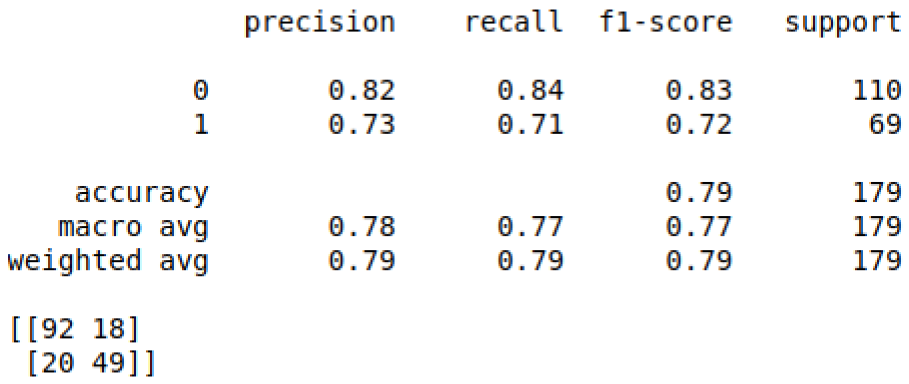
- Decision tree accuracy: 82.68%
- Let's check the accuracy of first 50 max_leaf_nodes and find the maximum accuracy.



The maximum Accuracy of Decision Tree Classifier is: 87.15%
The above graph show the maximum accuracy is 87.15% and the max_leaf_nodes attribute's value is 6.

6.3 Support Vector Classifier

Now let’s implement Support Vector Classifier to test the classification results and check the accuracy.



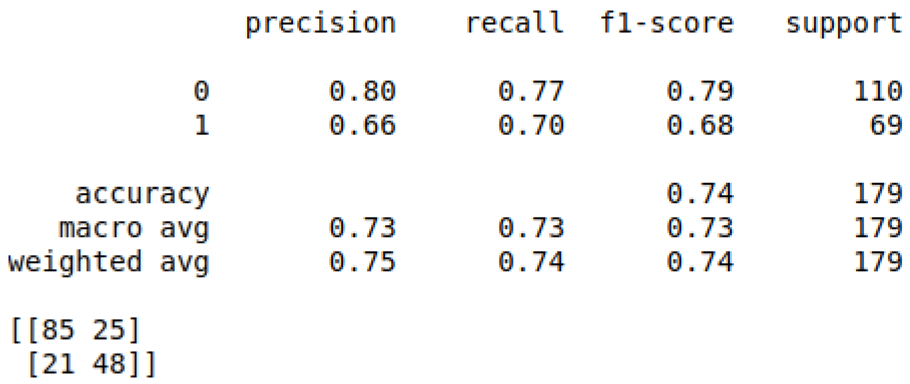
```
              precision    recall  f1-score   support

           0       0.82      0.84      0.83       110
           1       0.73      0.71      0.72        69

    accuracy                           0.79       179
   macro avg       0.78      0.77      0.77       179
weighted avg       0.79      0.79      0.79       179

[[92 18]
 [20 49]]
```

- SVC Classifier has accuracy: 79%

6.4 K Neighbors Classifier

KNN is very famous machine learning model. It can be used for both classification and regression problem.

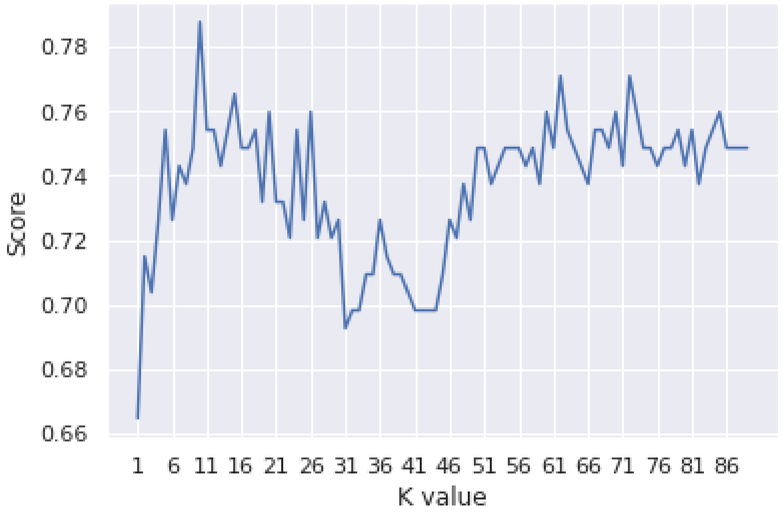
- Let's implement it and test the accuracy.
- I have randomly set n_neighbors to 90.



```
              precision    recall  f1-score   support

           0       0.80      0.77      0.79       110
           1       0.66      0.70      0.68        69

    accuracy                           0.74       179
   macro avg       0.73      0.73      0.73       179
weighted avg       0.75      0.74      0.74       179

[[85 25]
 [21 48]]
```

- K Neighbors Classifier accuracy: 74.30%

Let's test first 90 n_neighbors and see the maximum accuracy

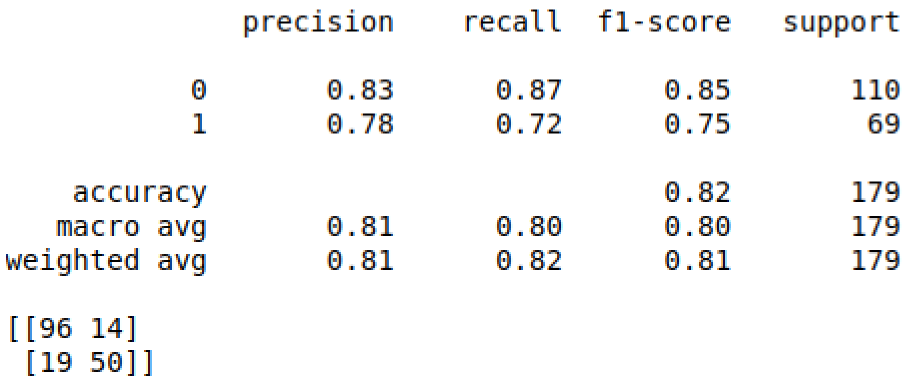


KNN Maximum Accuracy is: 78.77%

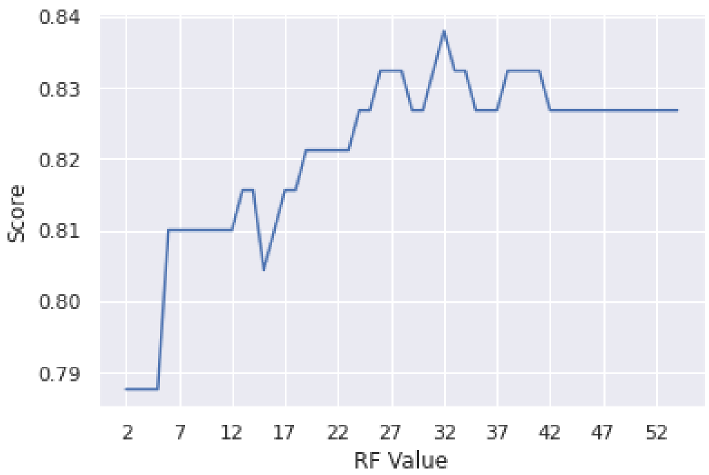
6.5 Random Forest Classifier

The next Model is Random Forest Classifier. It’s also a good algorithm for classification.

- I have randomly selected the max_leaf_nodes to 10.
- Let’s see the accuracy.



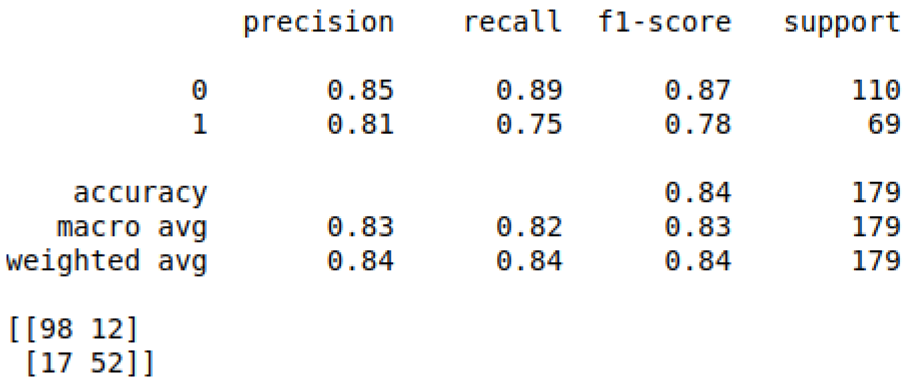
```
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       110
           1       0.78      0.72      0.75        69

    accuracy                           0.82       179
   macro avg       0.81      0.80      0.80       179
weighted avg       0.81      0.82      0.81       179

[[96 14]
 [19 50]]
```



- Random Forest accuracy is: 81.56%

Random Forest has maximum Accuracy: 83.80%

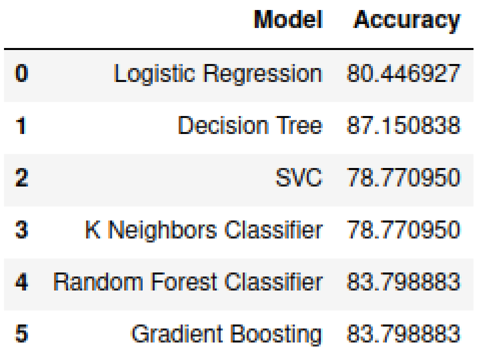
6.6 Gradient Boosting Classifier



```
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       110
           1       0.81      0.75      0.78        69

    accuracy                           0.84       179
   macro avg       0.83      0.82      0.83       179
weighted avg       0.84      0.84      0.84       179

[[98 12]
 [17 52]]
```

- Gradient Boosting accuracy is: 84%

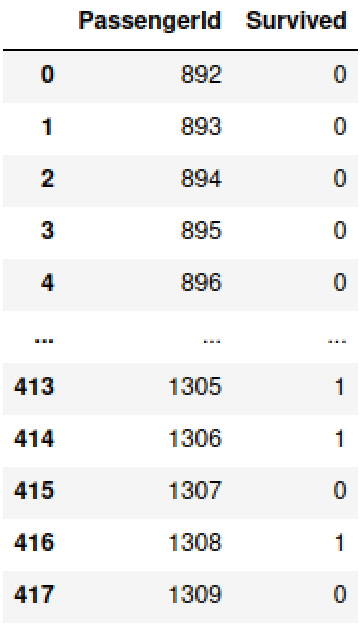
7. Model Comparison



	Model	Accuracy
0	Logistic Regression	80.446927
1	Decision Tree	87.150838
2	SVC	78.770950
3	K Neighbors Classifier	78.770950
4	Random Forest Classifier	83.798883
5	Gradient Boosting	83.798883

- Some models can achieve up to 80% accuracy
- Decision Tree has highest Accuracy

Now let's test Our Actual test data on a model having highest Accuracy. This data is predicted using Gradient Boosting Classifier.



	PassengerId	Survived
0	892	0
1	893	0
2	894	0
3	895	0
4	896	0
...	...	...
413	1305	1
414	1306	1
415	1307	0
416	1308	1
417	1309	0

8. Conclusion

An Exploratory Analysis was performed on Titanic dataset and multiple models have been built to predict the survived people. The preprocessing is performed on data to clean the data and give them structure and prepared them enough to be able to pass through machine learning models. The data was imbalanced, so Over sampling methodology was used to balance the data so that models don't show biased results and give better production. 6 machine learning algorithms have been applied to dataset to get better accuracy for better results. The Decision tree classifier is giving best accuracy.

9. Future work

The model can be enhanced using other models such as neural network. Also, different sampling methods rather than SMOTE Can be applied. Feature engineering has to be applied to reduce the over-fitting.

10. References.

[1] Dataset Retrieved from https://www.kaggle.com/c/titanic/data

[2] Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems April 2017

[3] Understanding random forest. Retrieved from official Scikit-learn documentation https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier

[4] Brownlee, J. (2021, March 16). SMOTE for imbalanced classification with Python. Retrieved May 06, 2021, from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/