TITANIC SURVIVAL PREDICTION USING MACHINE LEARNING

Predict passenger survival based on attributes like age, class, sex, fare, etc.

INDEX

DATASET OVERVIEW

- Source : Seaborn inbuilt Dataset
- Shape of Dataset :891 rows, 15 columns
- Dataset :

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True
6	0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
7	0	3	male	2.0	3	1	21.0750	S	Third	child	False	NaN	Southampton	no	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False

DATASET DICTIONARY

Variable	Definition
survival	Survival {0 = No, 1 = Yes}
pclass	Ticket class {1 = Class A, 2 = Class B, 3 = Class C}
sex	Sex
Age	Age in years
sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
cabin	Cabin number
embarked	Port of Embarkation {C = Cherbourg, Q = Queenstown, S = Southampton, O= Others}

EXPLORATORY DATA ANALYSIS (EDA)

Information table

- Having total 15 features.
- **Age** have 177 (~20%) null values
- **Embarked** have 2 (~0.22%) null values
- **Deck** have 688 (~77%) null values

	ColumnName	DataType	DistinctCount	Null	Percentage
0	survived	int64	2	0	0.00
1	pclass	int64	3	0	0.00
2	sex	object	2	0	0.00
3	age	float64	89	177	19.87
4	sibsp	int64	7	0	0.00
5	parch	int64	7	0	0.00
6	fare	float64	248	0	0.00
7	embarked	object	4	2	0.22
8	class	category	3	0	0.00
9	who	object	3	0	0.00
10	adult_male	bool	2	0	0.00
11	deck	category	8	688	77.22
12	embark_town	object	4	2	0.22
13	alive	object	2	0	0.00
14	alone	bool	2	0	0.00

Dtypes table

- Having total 6 data types.
- **Numeric** feature :6
- **Non-Numeric** :9

Passenge in each class

Class 1	24.2%
Class 2	20.7%
Class 3	55%

Data Types

Object	5
Int64	4
Float64	2
Bool	2
Category	2

Describe table

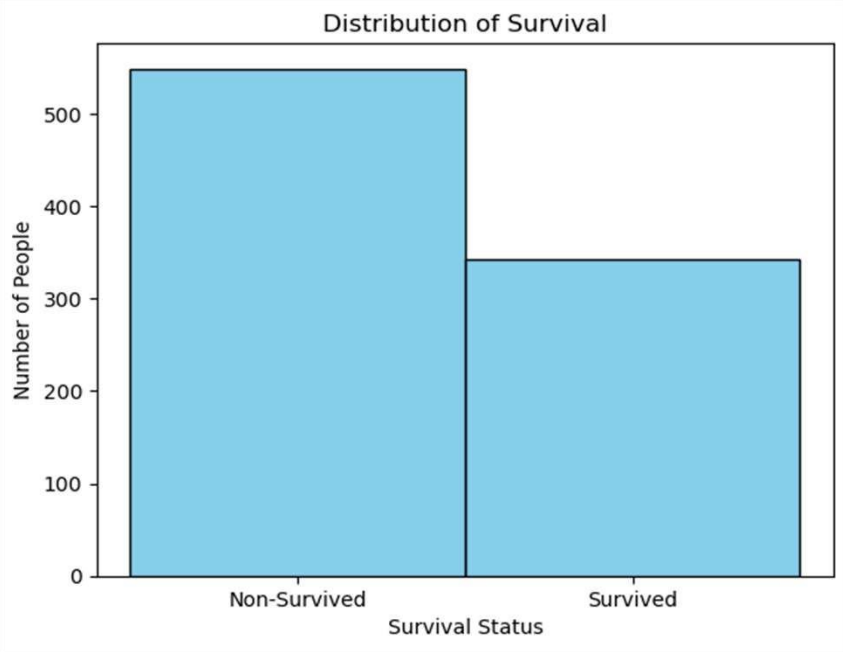
- We can see that **~38% people survived.**
- Most people are from **2nd and 3rd class.**
- Most people are of **age 30** on ship.

	ColumnName	mean	std	max	min
0	survived	0.383838	0.486319	1.0000	0.00
1	pclass	2.308642	0.835602	3.0000	1.00
2	age	29.699118	14.516321	80.0000	0.42
3	sibsp	0.523008	1.102124	8.0000	0.00
4	parch	0.381594	0.805605	6.0000	0.00
5	fare	32.204208	49.665534	512.3292	0.00

VISUAL ANALYSIS

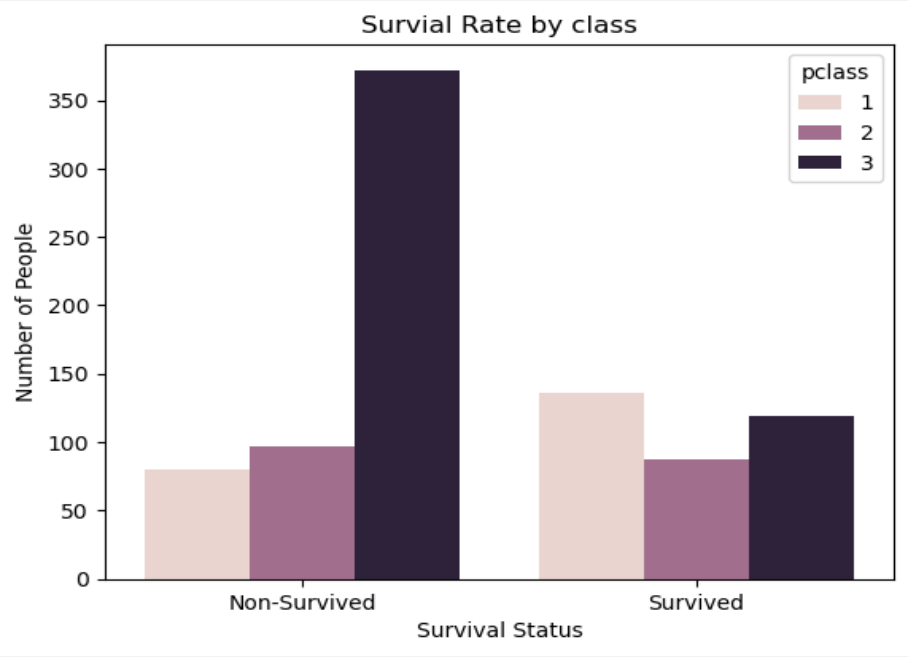
Distribution of survived vs non-survived.

- Approximately 549 (~61%) died out of 891 peoples.
- According to our data we can say that most number of people who died are from 3rd class.



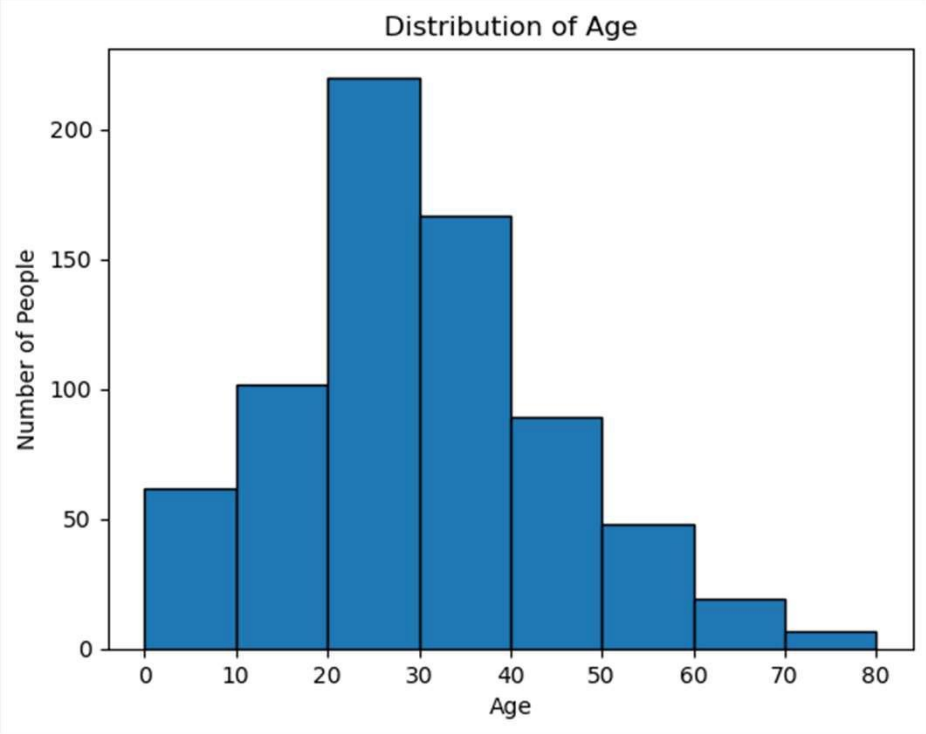
Survival by class

- From class 1 approx 63% of people got survived.
- From class 3 approx 76% of people died.



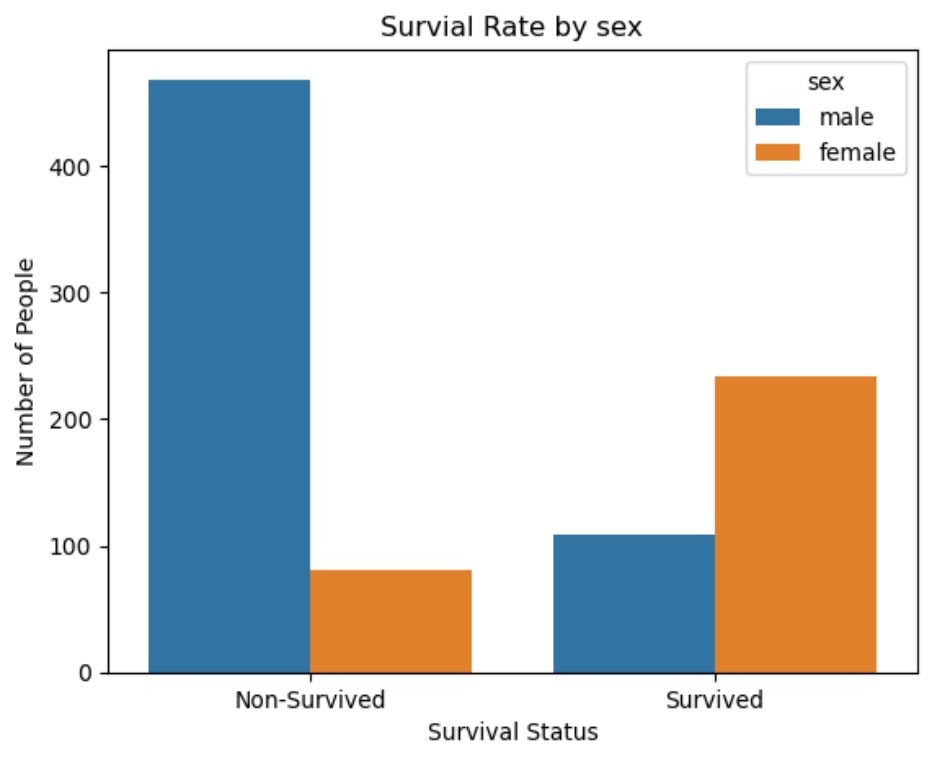
Age distribution

- Most of the people on the ship belongs to **20-30 age** group.
- There are **65% males** and **35% females.**
- So, we can say that there are many male adults on the ship.



Survival rate by sex

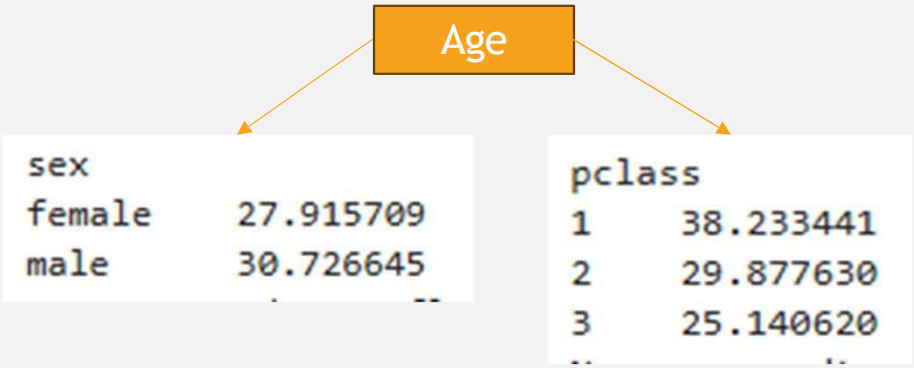
- According to data, most people died are male.
- Also, females are the most who survived.



DATA CLEANING & PREPROCESSING

Checking and handle missing values

- Most of null values are in age column so I handled by using MEAN.
- So, I preferred to fill null values by the **mean of age** from each class.



Convert categorical columns to numeric using encoding

- Used label encoder to encode categorical columns to numeric because I don't want to increase features.

Drop irrelevant columns for simplicity

- alive, class, embark_town, deck features are dropped.

MODEL SELECTION

Logistic Regression Model

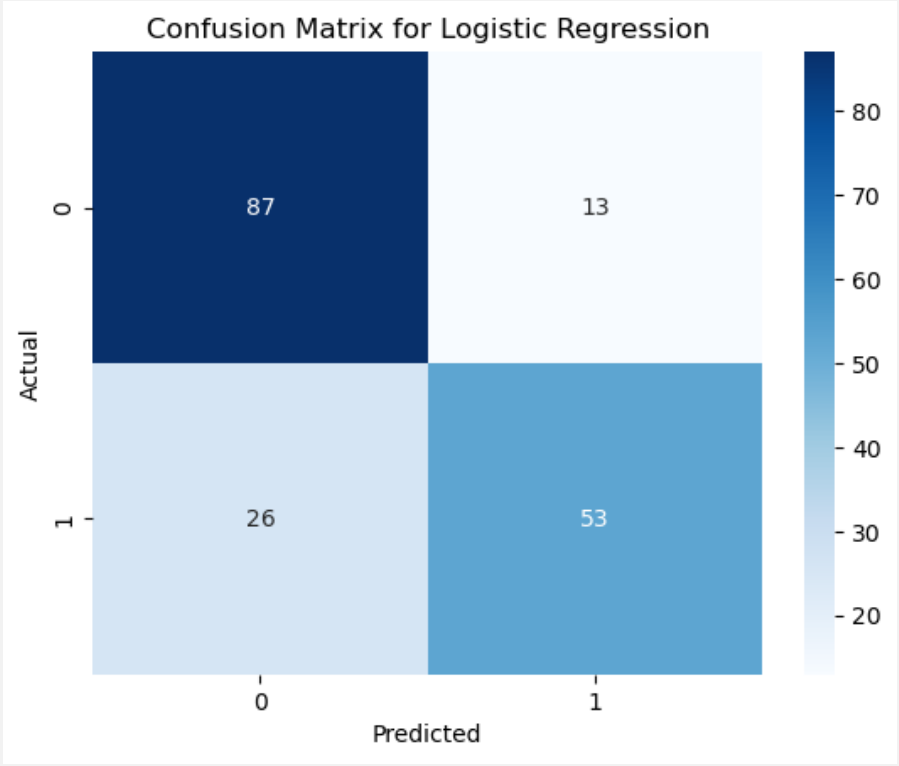
	Without HP	With HP
Train Score	83%	84%
Test Score	78%	78%
AUC score	83%	84%

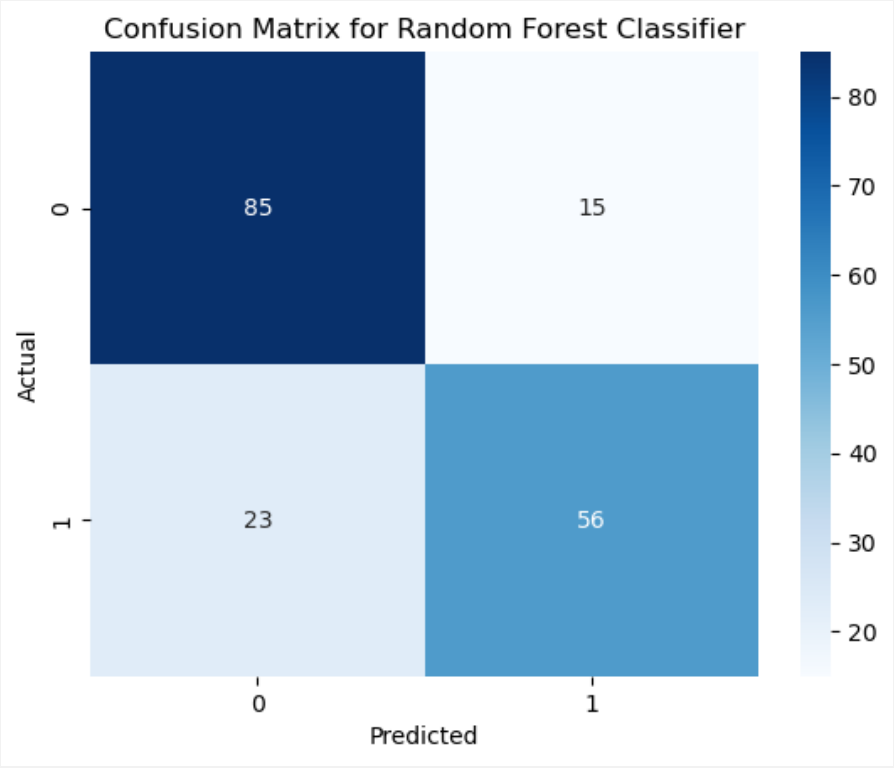
Random Forest Classifier Model

	Without HP	With HP
Train Score	99%	83%
Test Score	80%	80%
AUC score	86%	86%

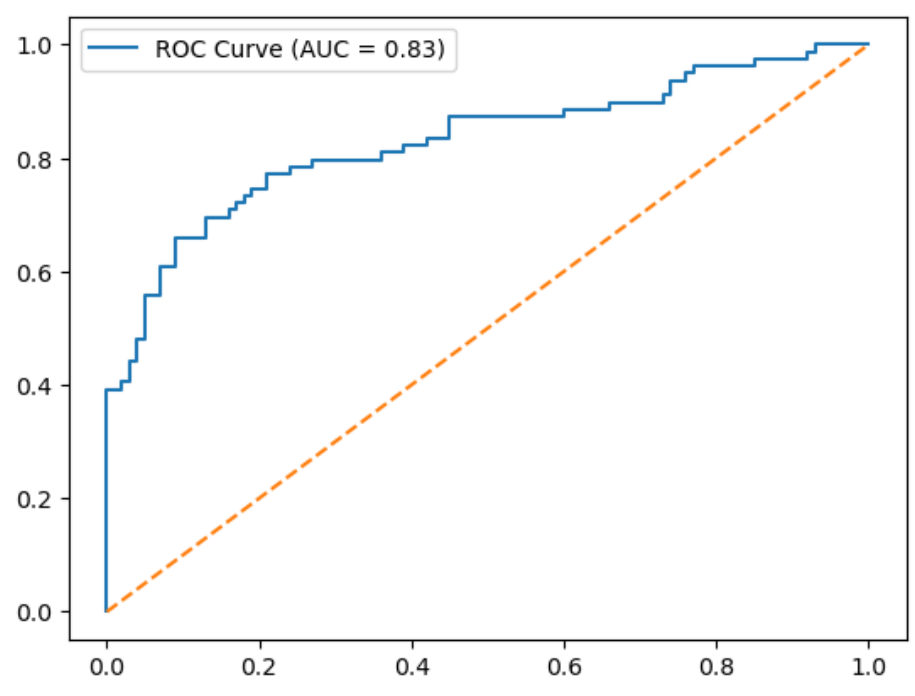
On the basis of above data we can say that Random Forest classifier model works better.

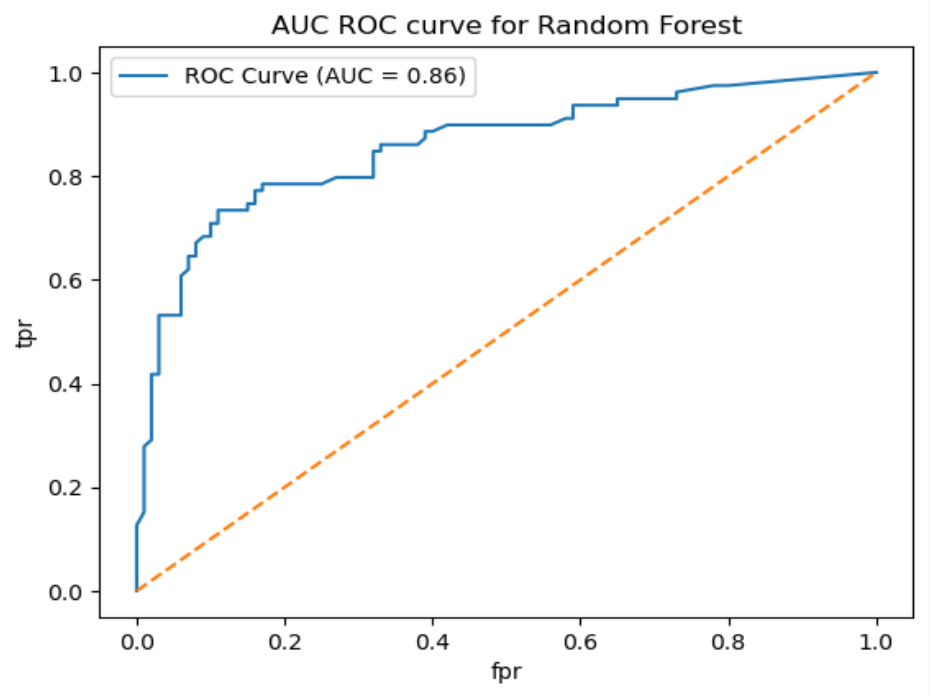
CONFUSION MATRIX BEFORE PARAMETERS TUNING



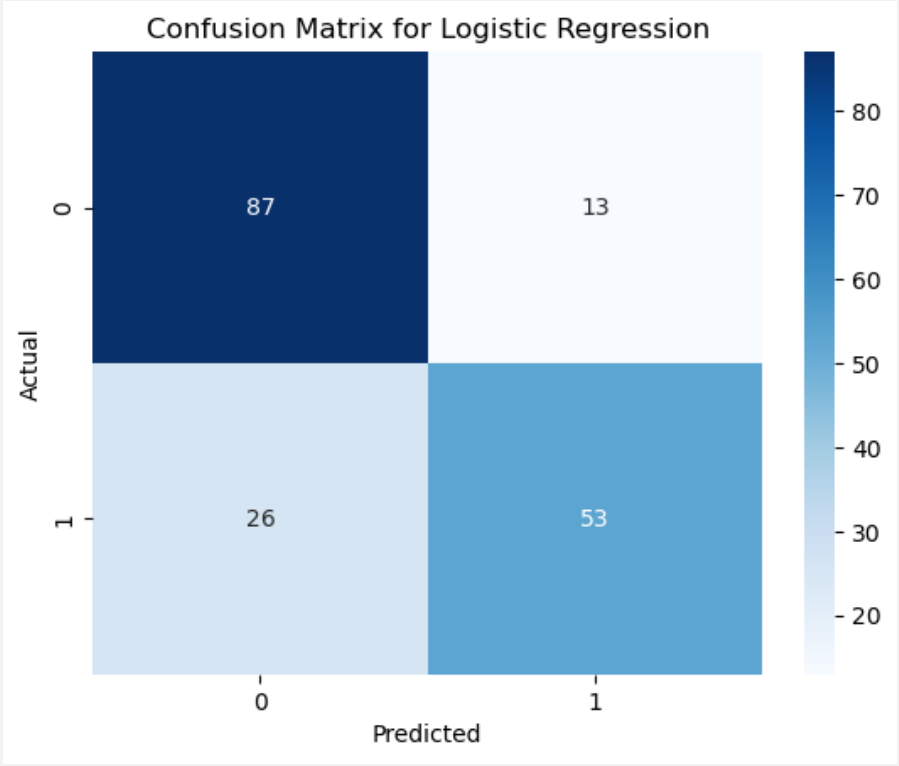


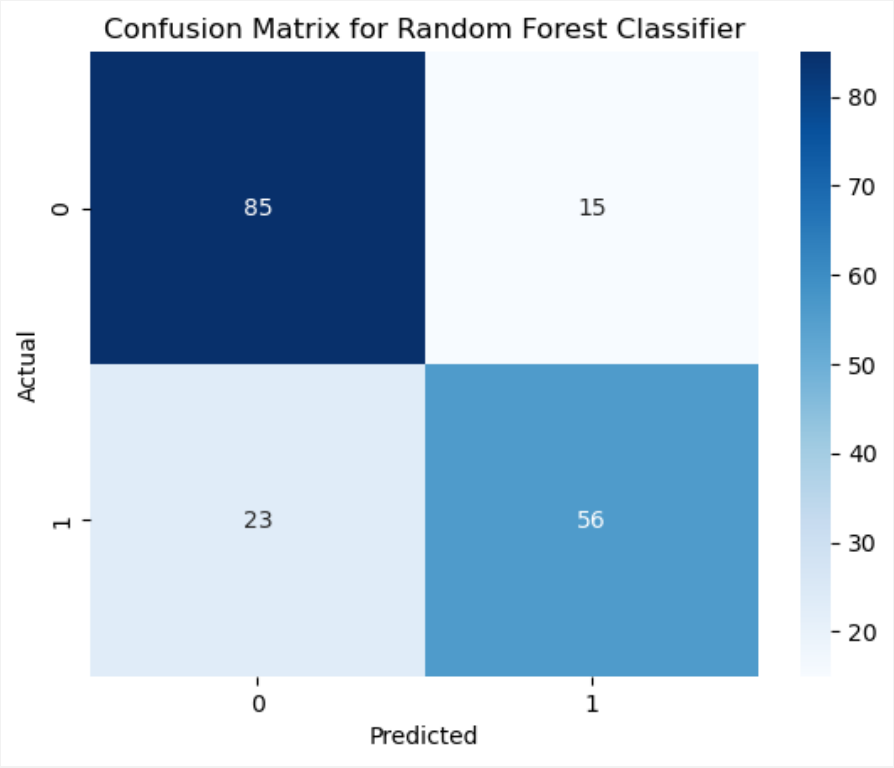
AUC AND ROC CURVE BEFORE TUNING



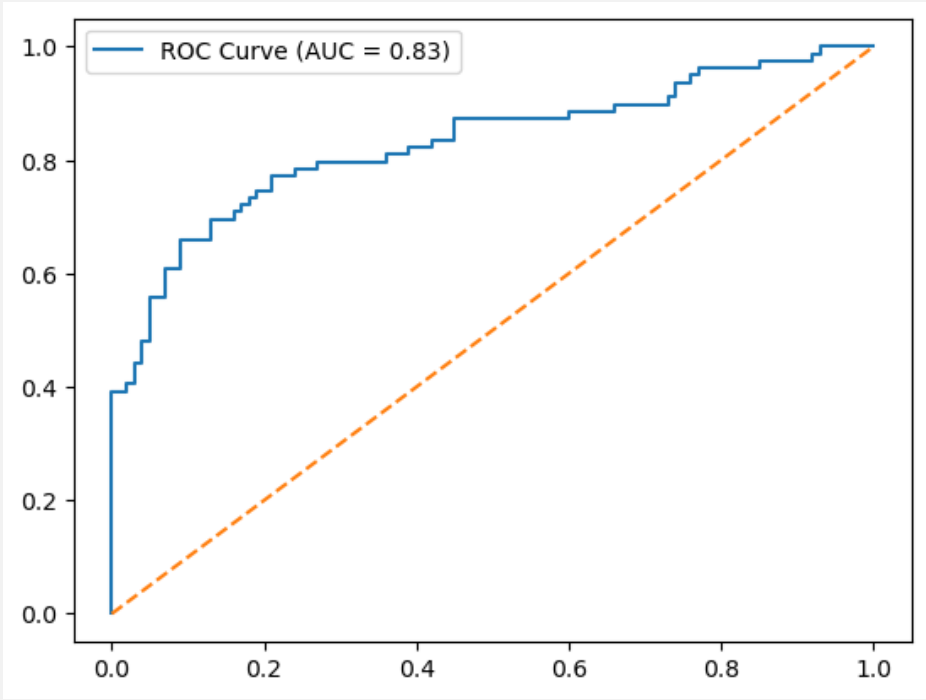


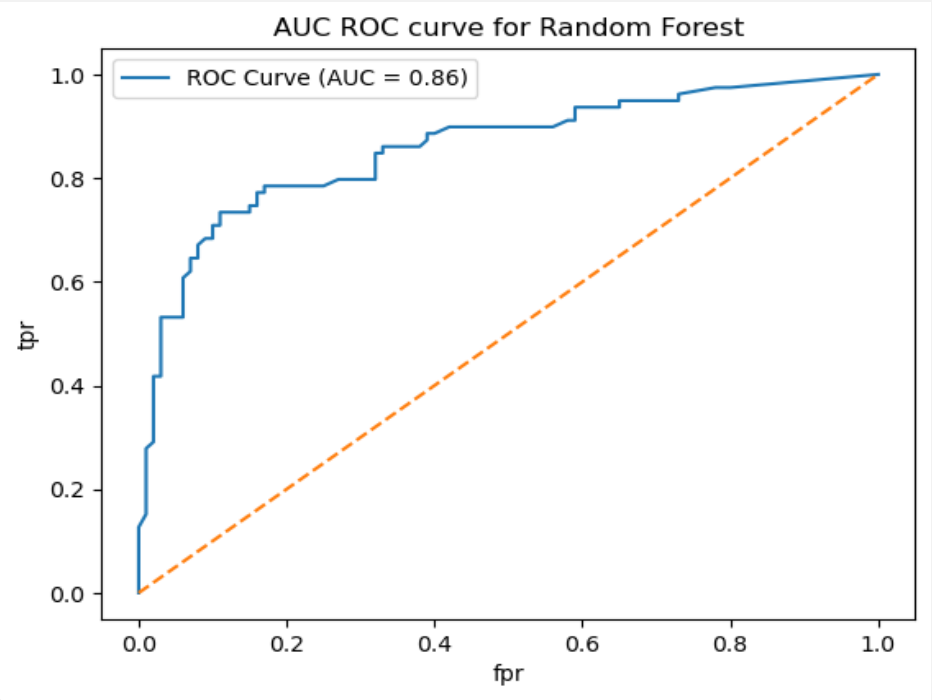
CONFUSION MATRIX AFTER PARAMETERS TUNING

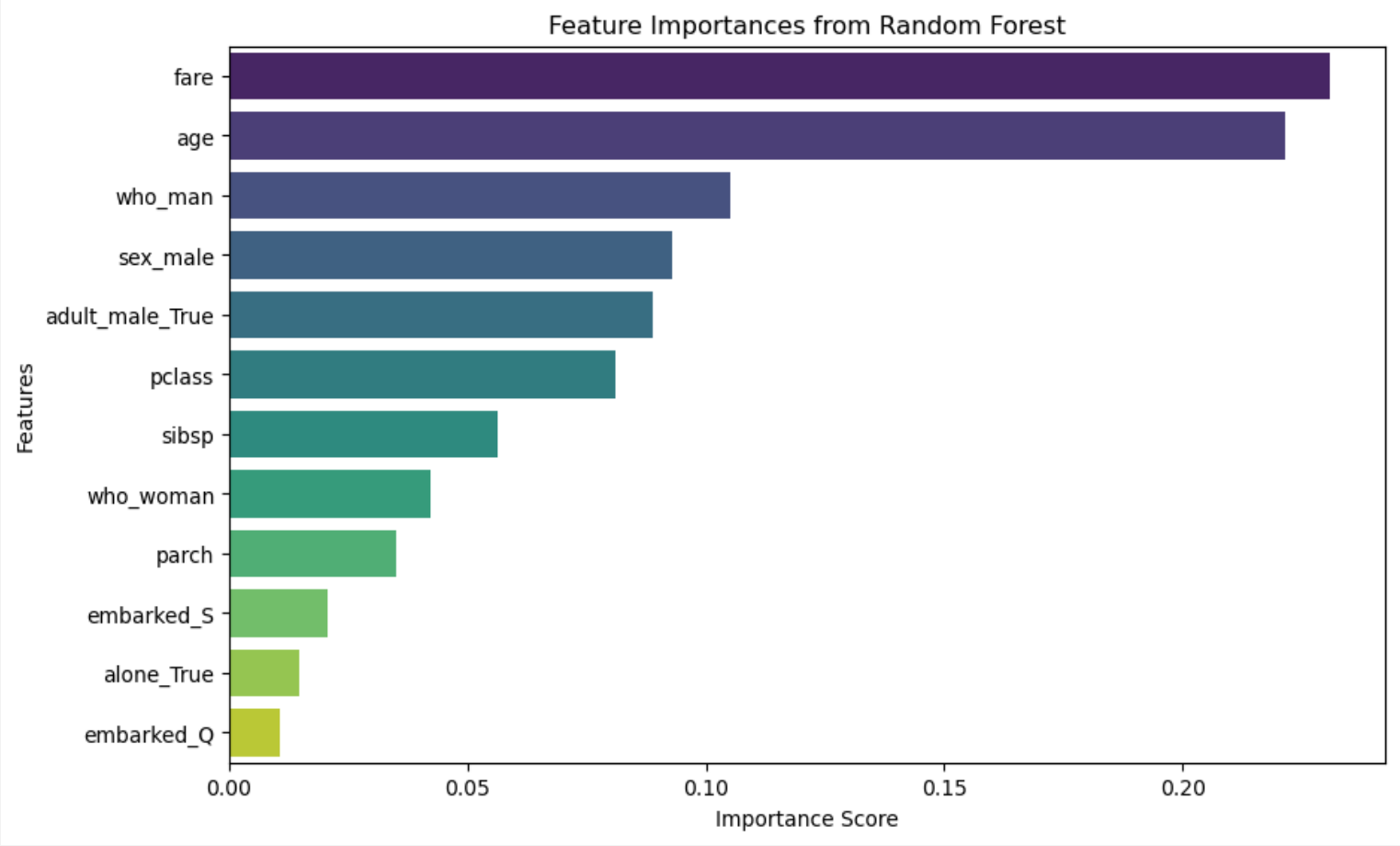




AUC AND ROC CURVE AFTER TUNING






Feature Importances from Random Forest
fare
age
who_man
sex_male
adult_male_True
pclass
sibsp
who_woman
parch
embarked_S
alone_True
embarked_Q
Features
0.00
0.05
0.10
0.15
0.20
Importance Score

CHALLENGES FACED

1. Concept and implementation of Optum parameter tuning is difficult.
2. To understand the relation between features.
3. To choose the right graph.

NOTE

1. Optum params tuning is very advance.
2. Uses the previous accuracy score to predict the next HP using TPESampler Algo.
3. Also used to find the best algo for our data.

CONCLUSION

We can say that Random Forest Classifier model fits better and gives the best result.
By doing more parameter tuning we can definitely increase the accuracy.