Titanic Dataset Analysis

Digital Skill Fair 30.0

FAZRIN MEILA AZZAHRA SOFYAN

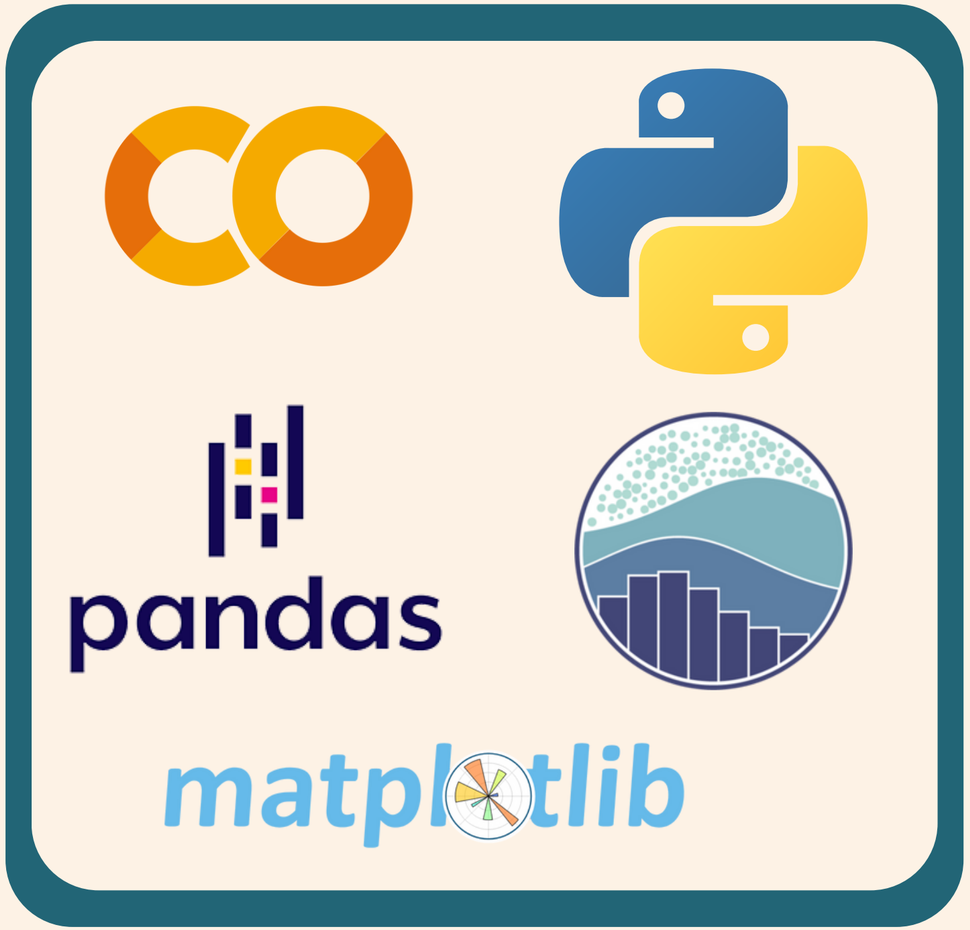
OVERVIEW

Titanic, British luxury passenger liner that sank on April 14–15, 1912, during its maiden voyage, en route to New York City from Southhampton, England, killing about 1,500 passengers and ship personnel. One of the most famous tragedies in modern history, it inspired numerous stories, several films, and a musical and has been the subject of much scholarship and scientific speculation.

Dataset & Tools

The dataset used can be accessed at this link

https://www.kaggle.com/competitions/titanic/data?select=train.csv



DATASET OVERVIEW

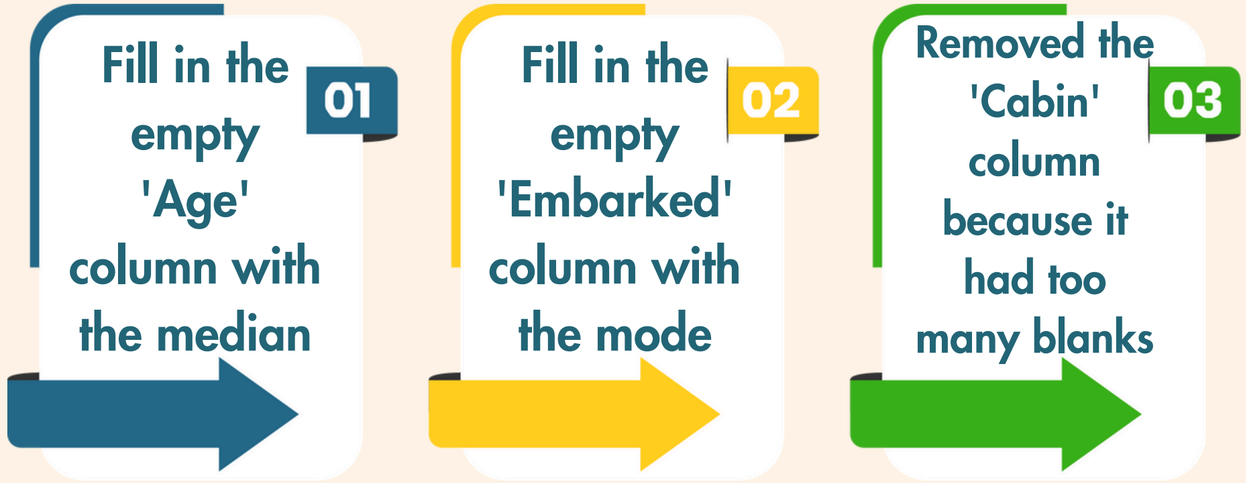
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

The Titanic dataset has 891 rows of data and 12 columns of information that include various important variables, such as passenger age, passenger class, and safety status, and there are some columns that have missing values such as Age, Cabin and Embarked.

DATA PREPROCESSING

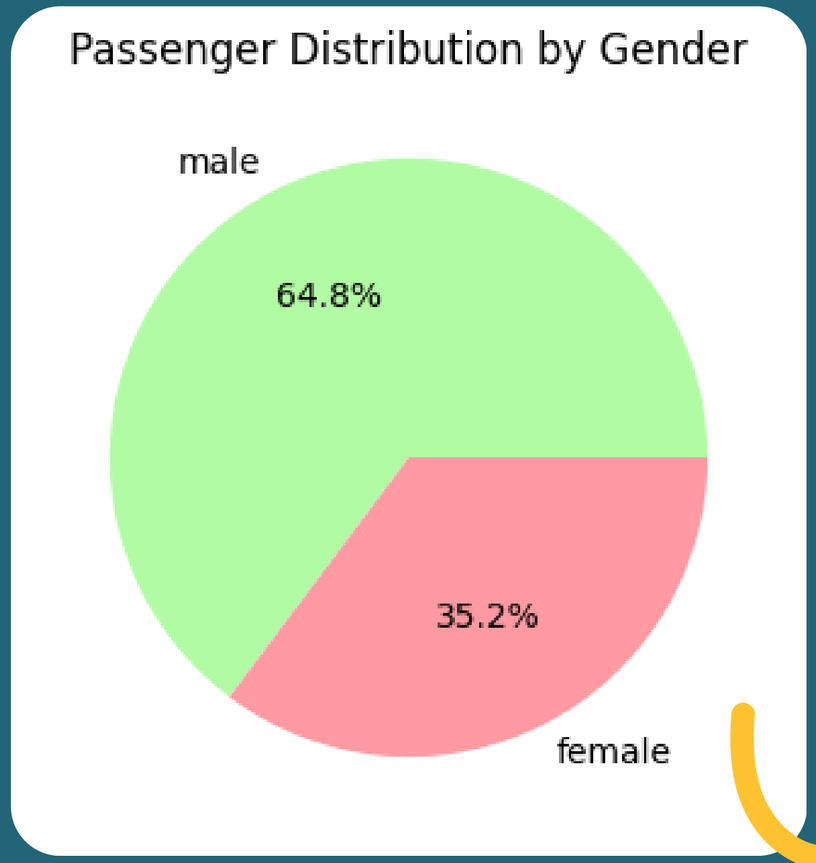


PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S

the data is clean and ready to be explored and analyzed

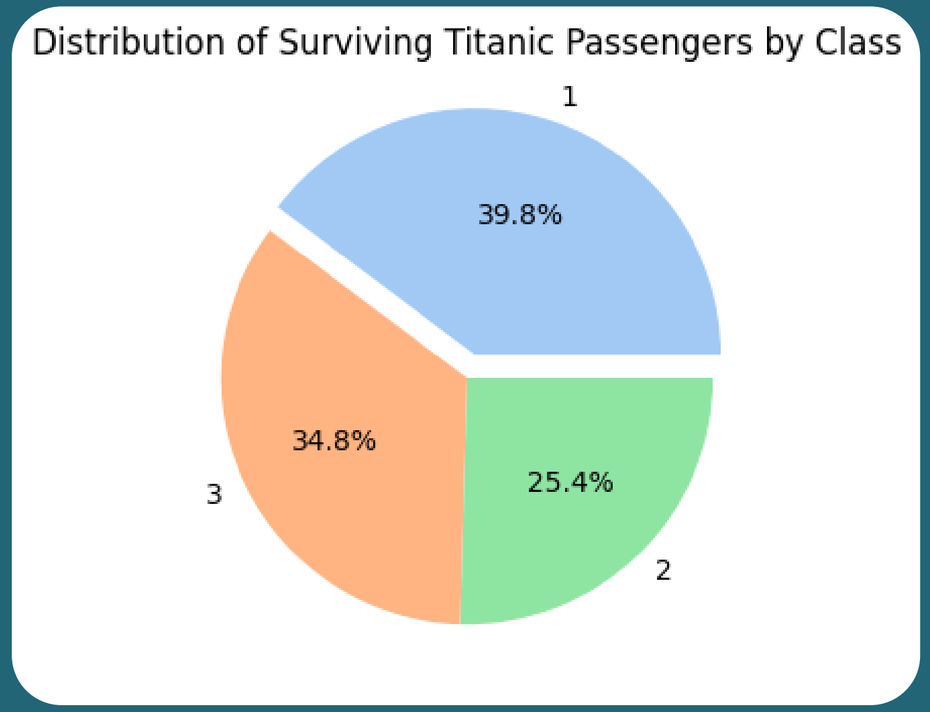
Exploratory Data Analysis

Passenger Distribution by Gender



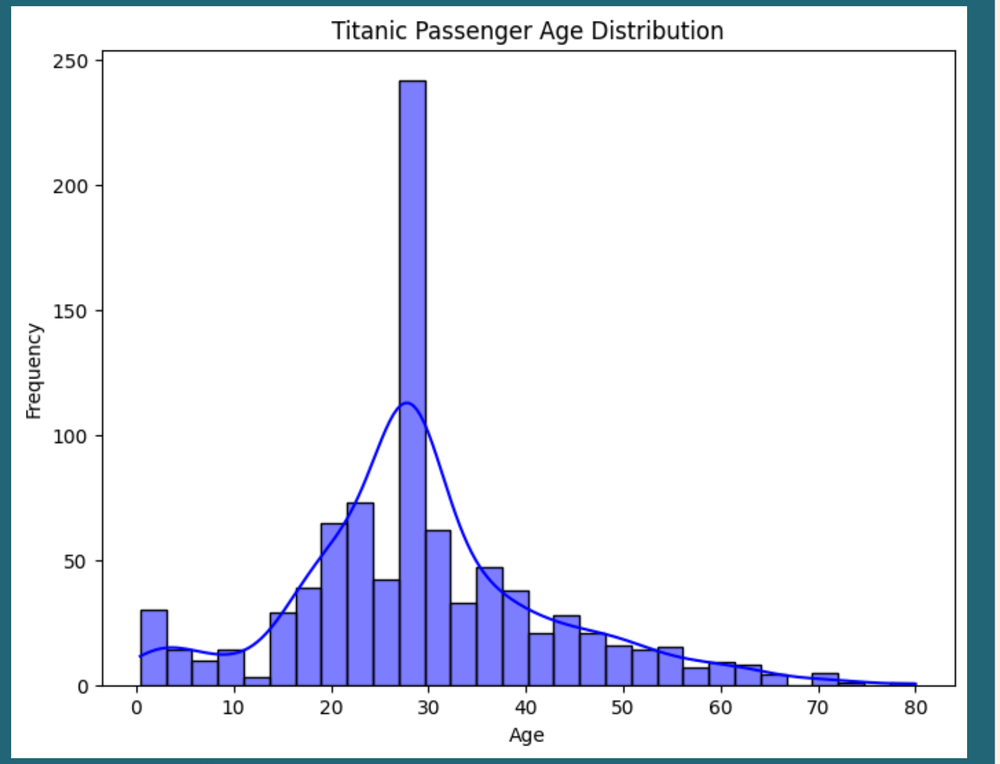
The distribution of Titanic passengers by gender shows that of the total passengers, about 65% were male and 35% were female. The pie chart clearly illustrates the dominance of the number of male passengers on board compared to women.

Distribution of Surviving Titanic Passengers by Class



Class 1 passengers appear to have the most survivors at 39.8%. Followed by class 3 with 34.8% and class 2 with 25.4%.

Titanic Passenger Age Distribution



Titanic's passengers showed that the majority of passengers were in the age range of 20 to 30 years old, with lower concentrations in the very young (under 10 years old) and old (over 60 years old), indicating that most passengers were young adults.

LABEL ENCODER

convert categorical data into numerical data

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.250	2
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	71.283	0
3	1	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.925	2
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.100	2
5	0	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.050	2

the columns that were changed are the ‘Sex’ column and the ‘Embarked’ column, both columns are now numeric data.

FEATURE SELECTION & SPLITTING DATA

```
features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked']
X = df[features]
y = df['Survived']
```

Feature selection is selected to continue the next process, the columns selected as features are the 'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', and 'Embarked' columns as X variables. As for the 'Survived' column, it is selected as the y-variable.

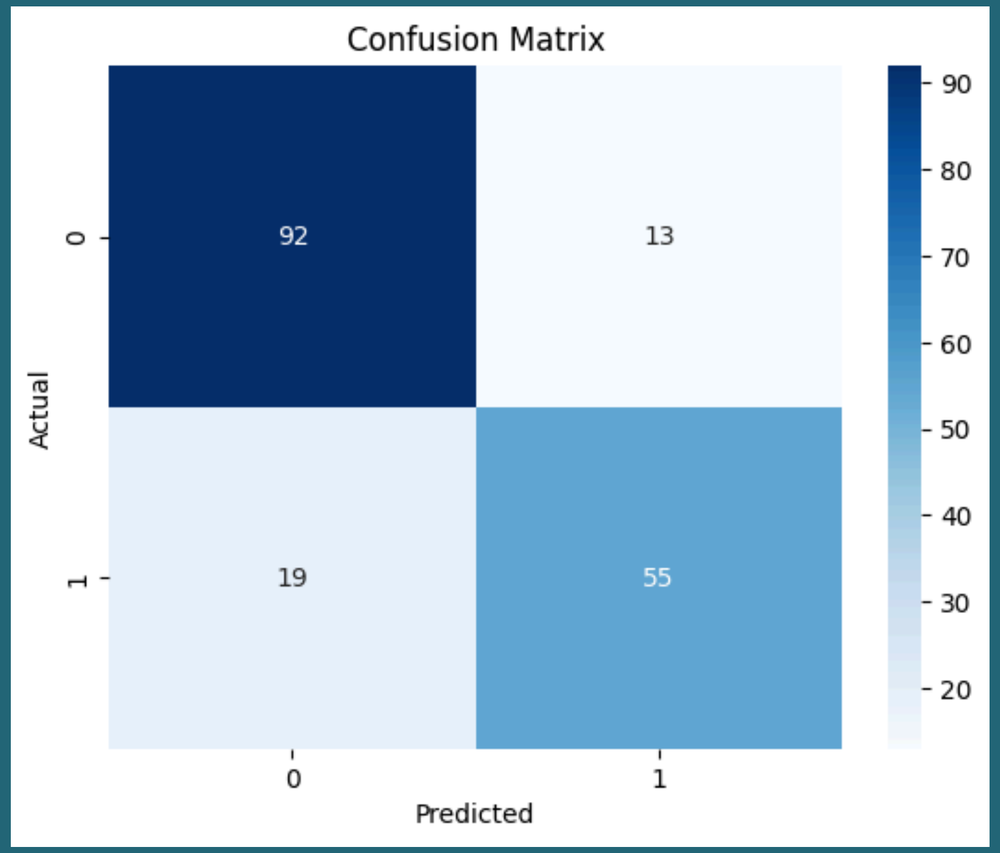
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
jumlah_data_training = len(X_train)
jumlah_data_testing = len(X_test)

print(f"Jumlah data training: {jumlah_data_training}")
print(f"Jumlah data testing: {jumlah_data_testing}")
```

```
Jumlah data training: 712
Jumlah data testing: 179
```

The entire Titanic dataset will be split in an 80:20 ratio, with 712 data as training data and 179 data as testing data.

MODELLING

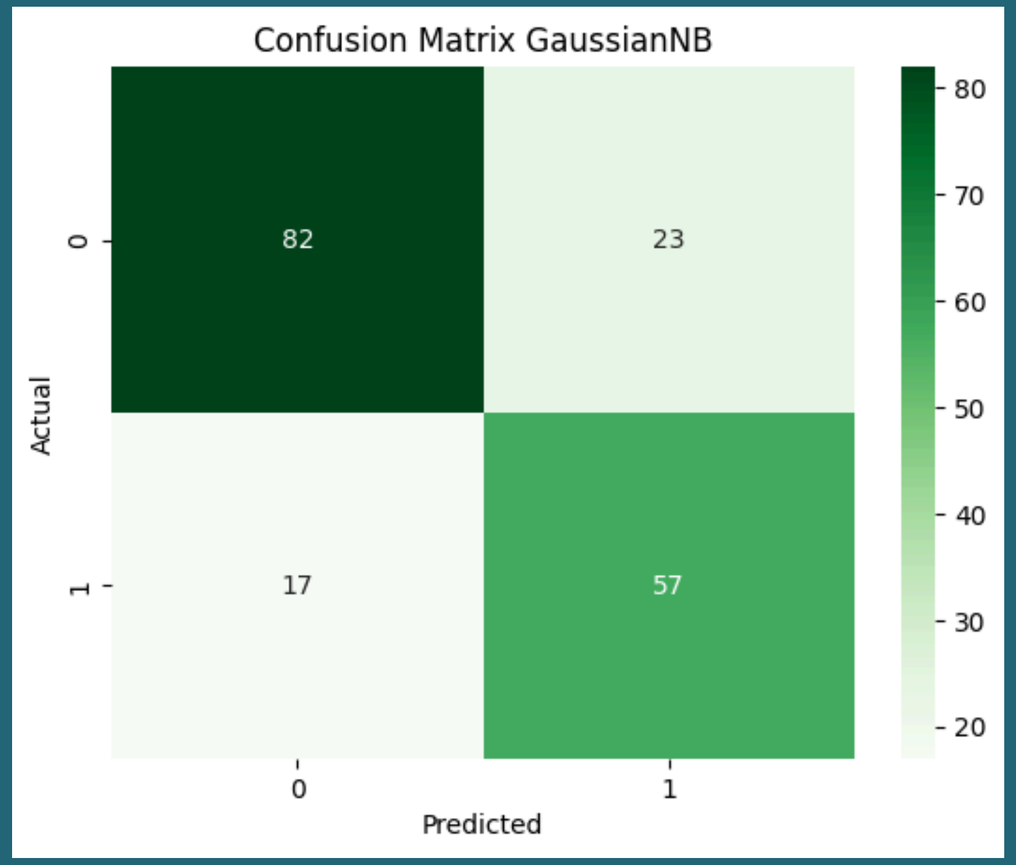
Random Forest



	precision	recall	f1-score	support
0	0.83	0.88	0.85	105
1	0.81	0.74	0.77	74
accuracy			0.82	179
macro avg	0.82	0.81	0.81	179
weighted avg	0.82	0.82	0.82	179

Modeling using the random forest algorithm has an **accuracy of 82%**, **recall of 81%** and **precision of 82%.**

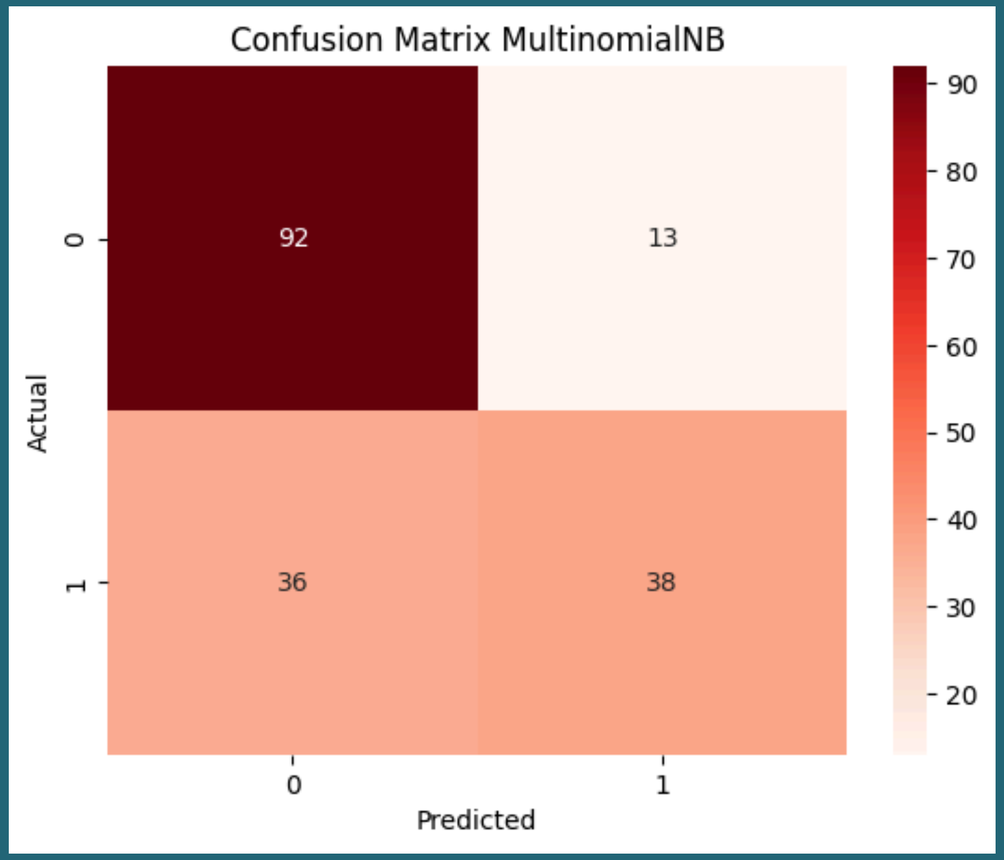
Naive Bayes - Gaussian

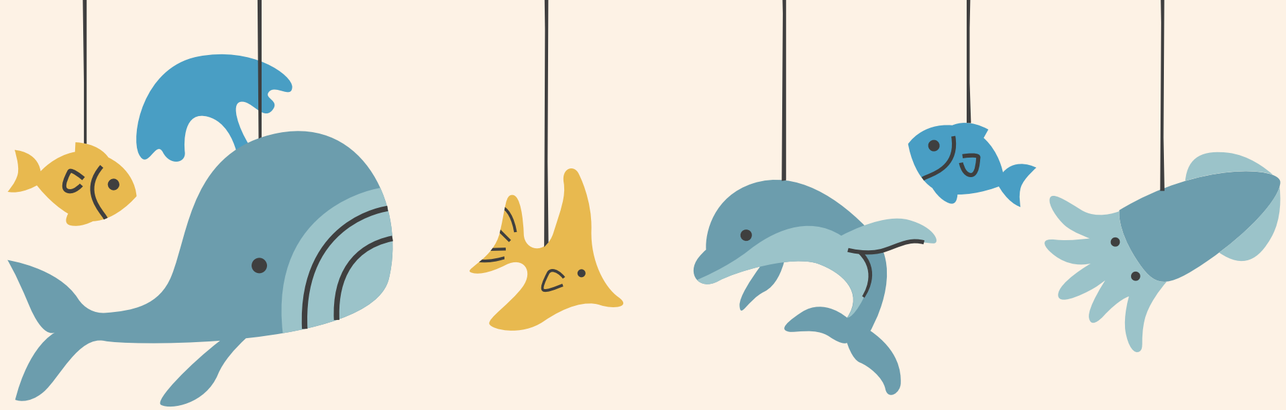


	precision	recall	f1-score	support
0	0.83	0.78	0.80	105
1	0.71	0.77	0.74	74
accuracy			0.78	179
macro avg	0.77	0.78	0.77	179
weighted avg	0.78	0.78	0.78	179

Modeling using the Gaussian Naive Bayes algorithm has an **accuracy of 78%**, **recall of 78%** and **precision of 77%.**

Naive Bayes - Multinomial

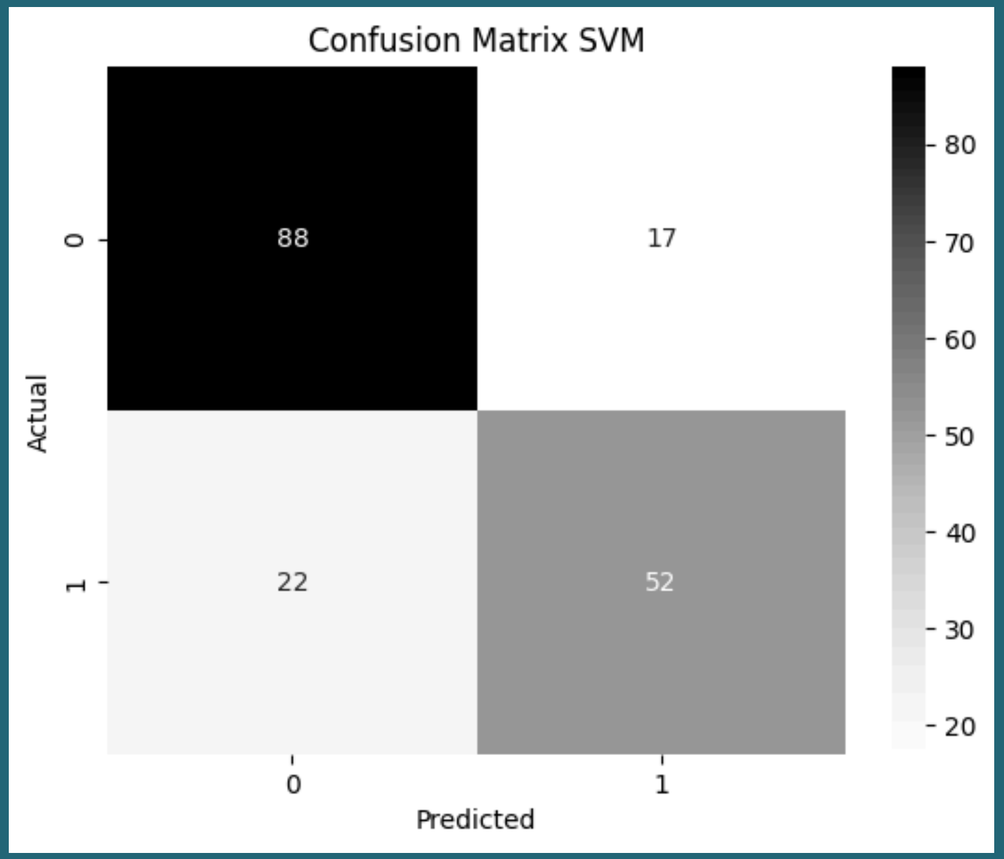




	precision	recall	f1-score	support
0	0.72	0.88	0.79	105
1	0.75	0.51	0.61	74
accuracy			0.73	179
macro avg	0.73	0.69	0.70	179
weighted avg	0.73	0.73	0.71	179

Modeling using the Gaussian Naive Bayes algorithm has an **accuracy of 73%**, **recall of 69%** and **precision of 73%.**

Support Vector Machine (SVM)



	precision	recall	f1-score	support
0	0.80	0.84	0.82	105
1	0.75	0.70	0.73	74
accuracy			0.78	179
macro avg	0.78	0.77	0.77	179
weighted avg	0.78	0.78	0.78	179

Modeling using the Gaussian Naive Bayes algorithm has an **accuracy of 78%**, **recall of 77%** and **precision of 78%.**

CONCLUSION

- Male passengers outnumber female passengers
- More class 1 passengers survived followed by class 3 then class 2
- The majority of passengers are in the age range of 20 to 30 years old

Several algorithms were tested to find out the results of good accuracy values, the algorithms tested were random forest, naive bayes (Gaussian and Multinomial), and SVM. The **algorithm** that has **good accuracy** is **Random Forest with an accuracy of 82%**, Gaussian Naive Bayes and SVM both have the same accuracy of 78% while multinomial naive bayes has an accuracy of 73%.

THANK YOU