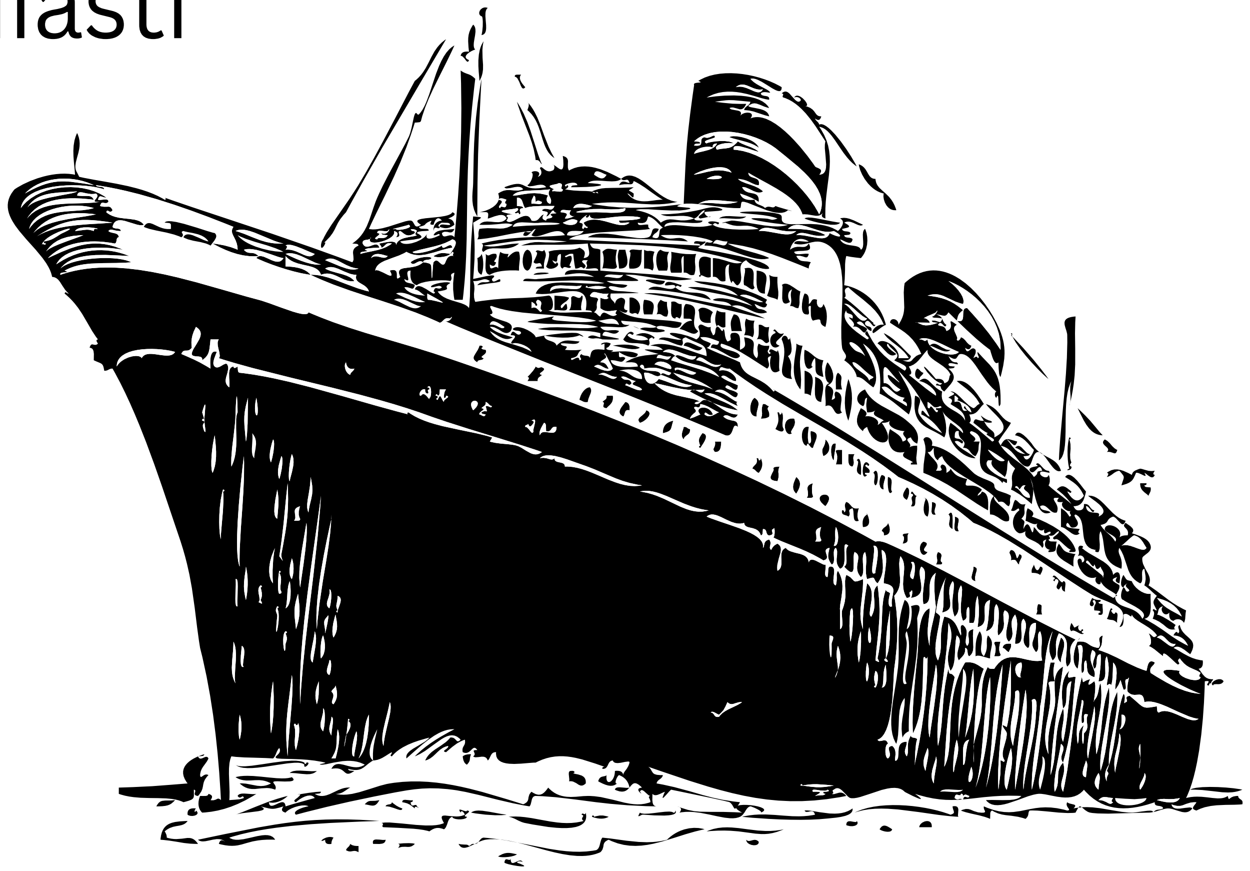
TITANIC SURVIVAL

Prediction using Machine Learning Models

by Maresha WIdya Muliadiasti



OVERVIEW

Titanic, launched on May 31, 1911, and set sail on its maiden voyage from Southampton on April 10, 1912, with 2,240 passengers and crew on board. On April 15, 1912, after striking an iceberg, Titanic broke apart and sank to the bottom of the ocean, taking with it the lives of more than 1,500 passengers and crew.

Source : NOAA gov

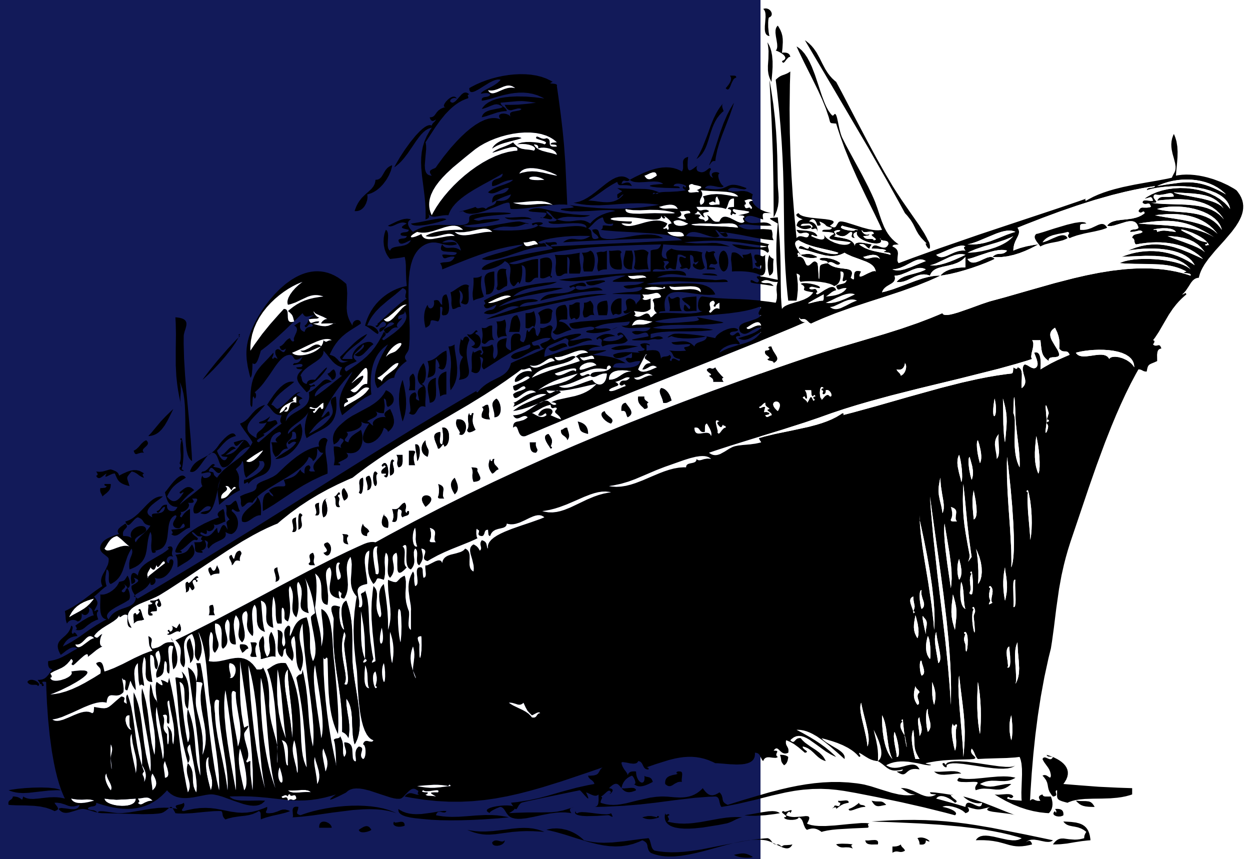
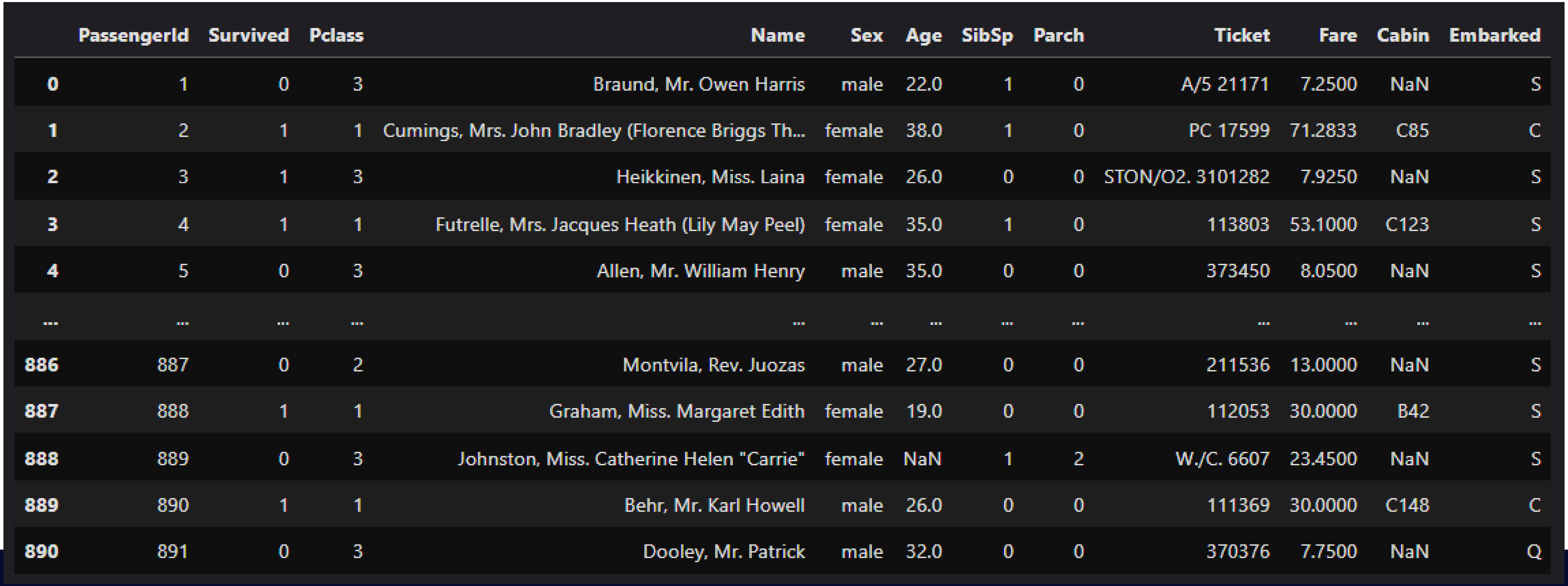


Table Of Contents

1. Data Description
2. Preprocessing Data
3. Feature Engineering
4. Data Modelling and Evaluation
5. Conclusion

Data Description

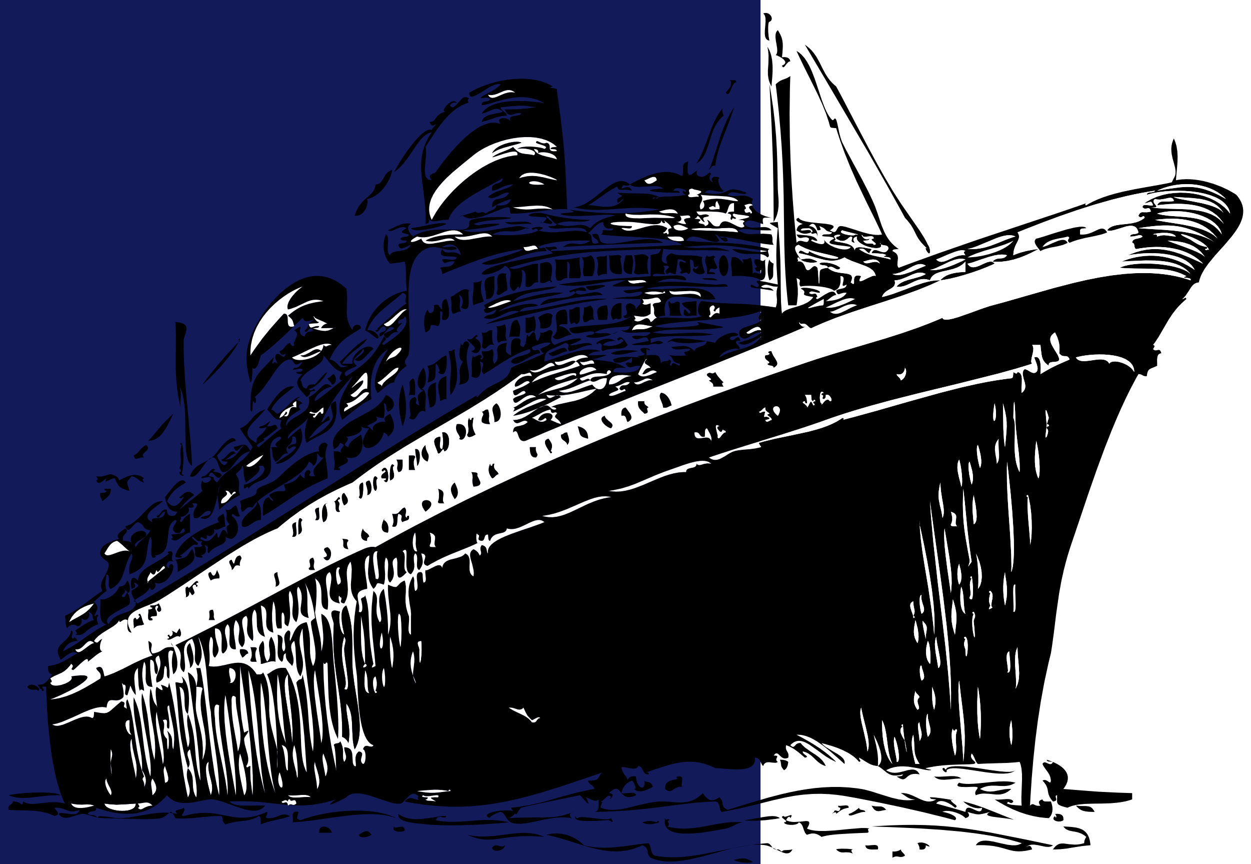


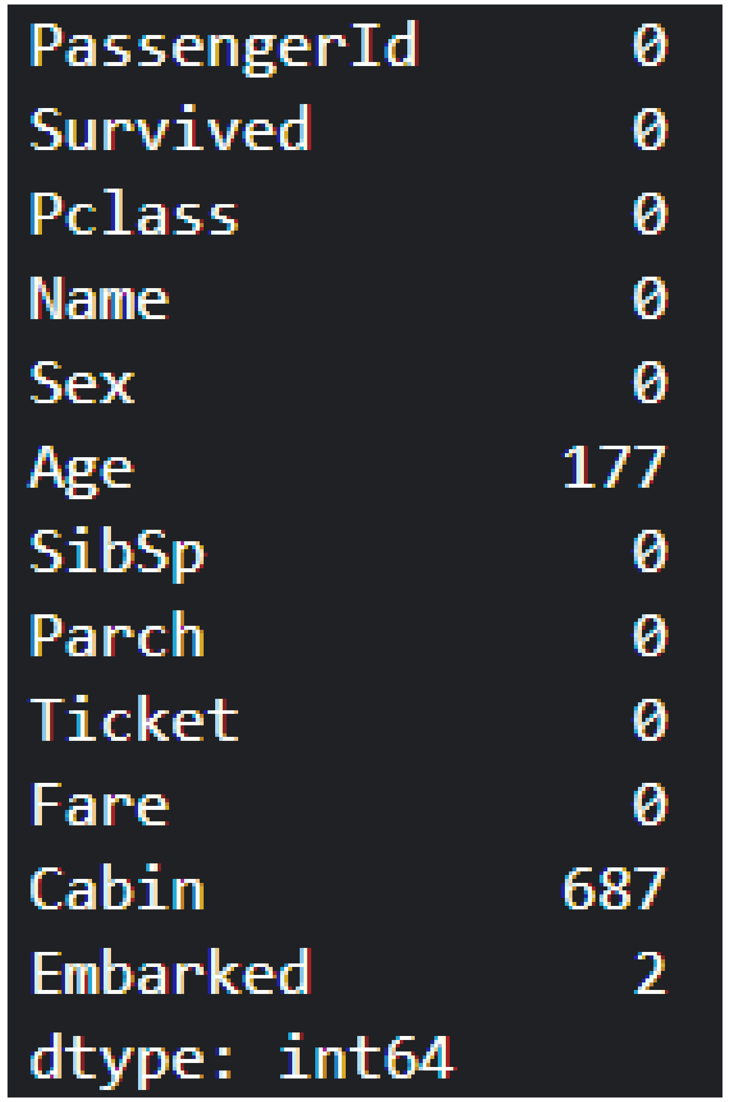
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 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

In this case, we used Titanic Survival Datasets from Kaggle. This data contains PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.

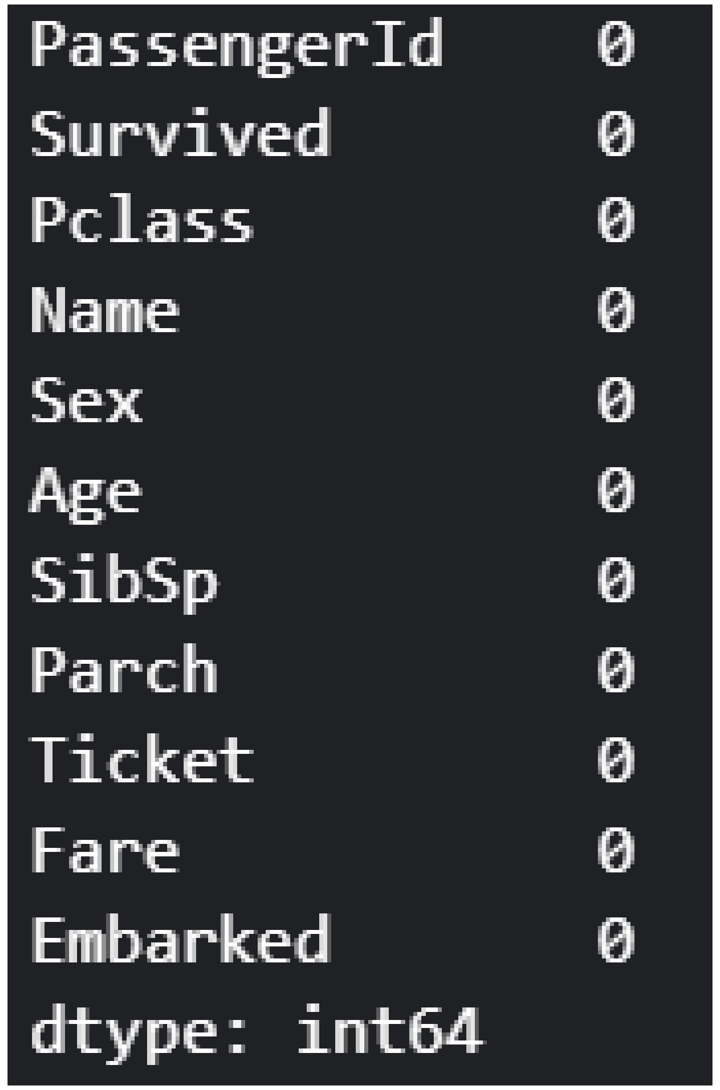
Data Preprocessing

Data preprocessing is the process of transforming raw data into a clean and usable format, making it suitable for analysis and model training in fields such as data mining, machine learning, and artificial intelligence. This essential step involves various techniques to handle issues like missing values, inconsistencies, and noise in the data, ensuring high-quality inputs for analytical tasks. In this case, there are three columns had missing data (Age, Cabin, and Embarked). These value need to be handled, we can imputing using the mean or median of the value, that prediction model can function optimally.





```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```



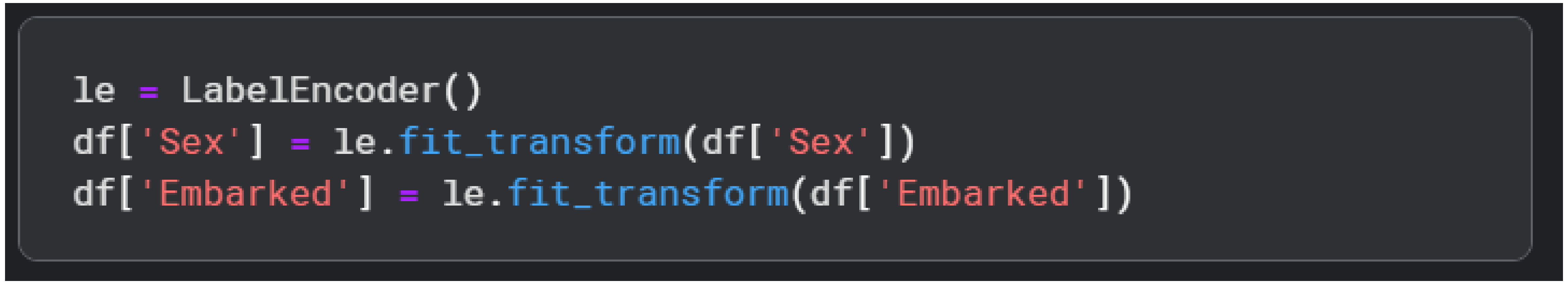
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Age was filled with the median

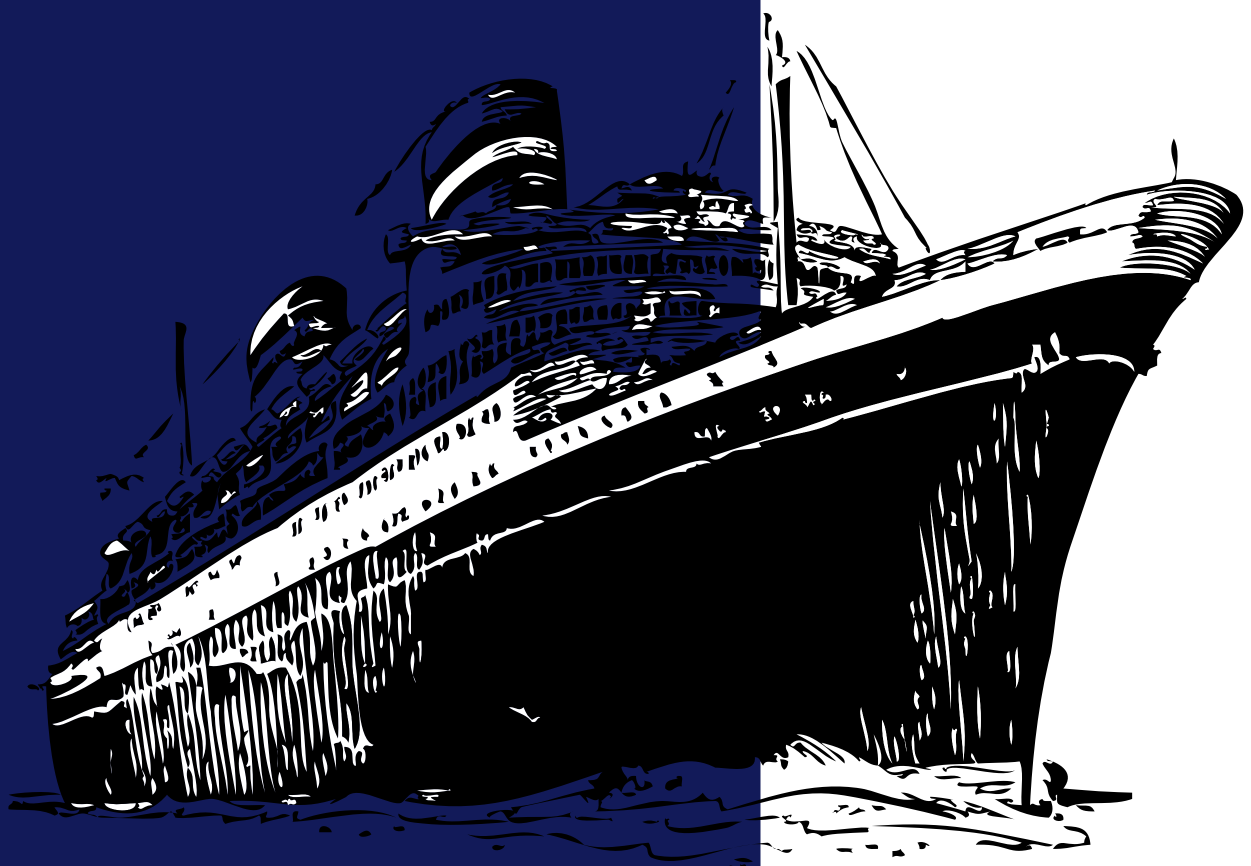
Embarked was filled with the mode

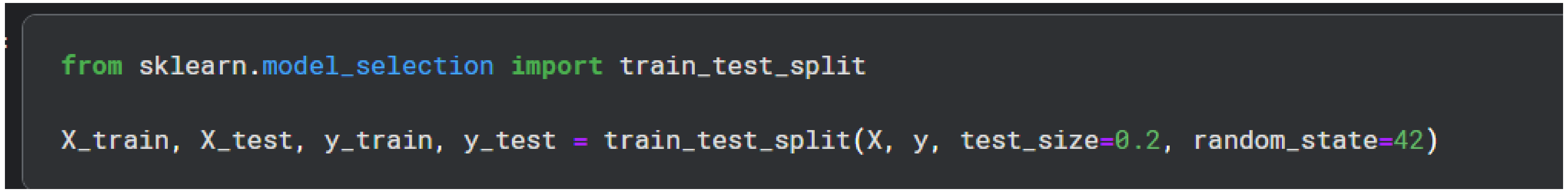
Feature Engineering

in this case, we using Label Encoder for changing categoric data into numeric data and splitting data divided by training data and testing data with an 80:20 ratio.



```
le = LabelEncoder()
df['Sex'] = le.fit_transform(df['Sex'])
df['Embarked'] = le.fit_transform(df['Embarked'])
```

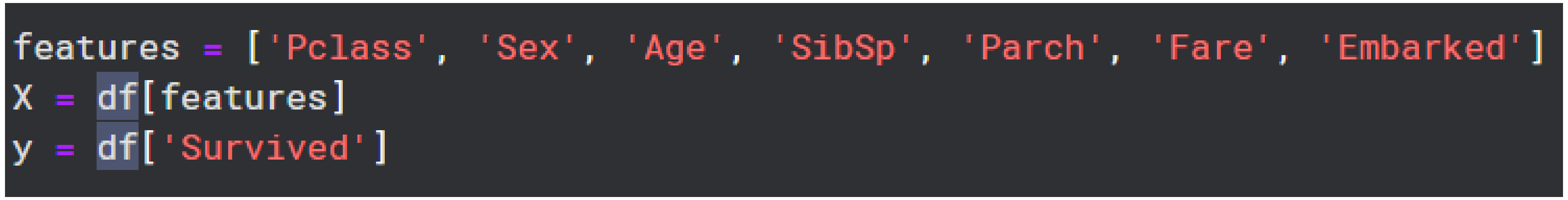




```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

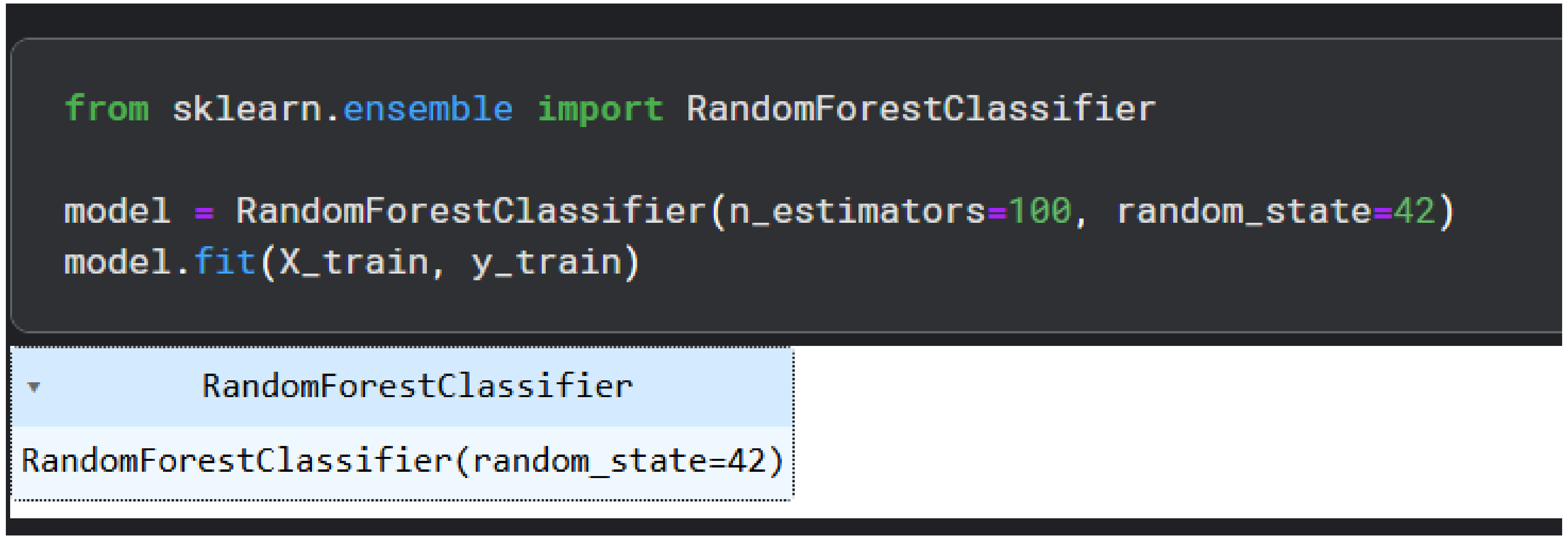
Feature Selection is used for identifying and selectiong the most relevant features from a datasets for use in building a machine learning model. In this case, we use **features** for X variables and Survived as Y Variable.



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

Data Modelling and Evaluation

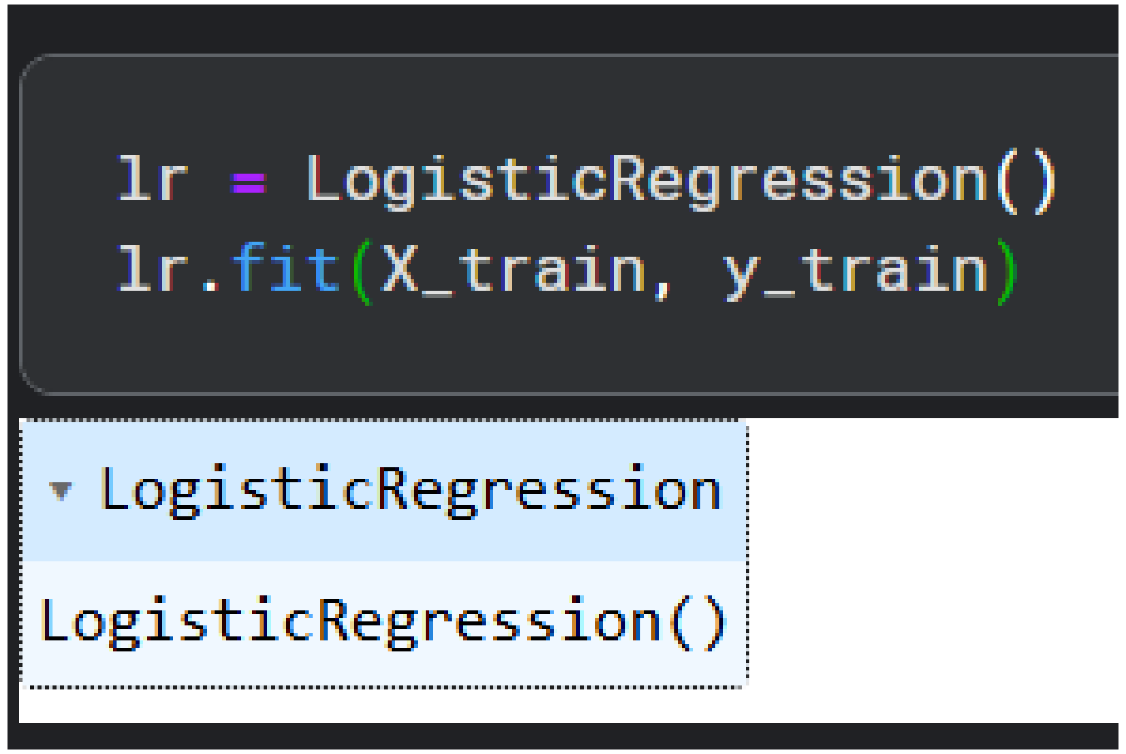
in this case, we comparing Random Forest Classification and Logistic Regression model for knowing the best model to predicting Titanic passenger survival.



```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
```

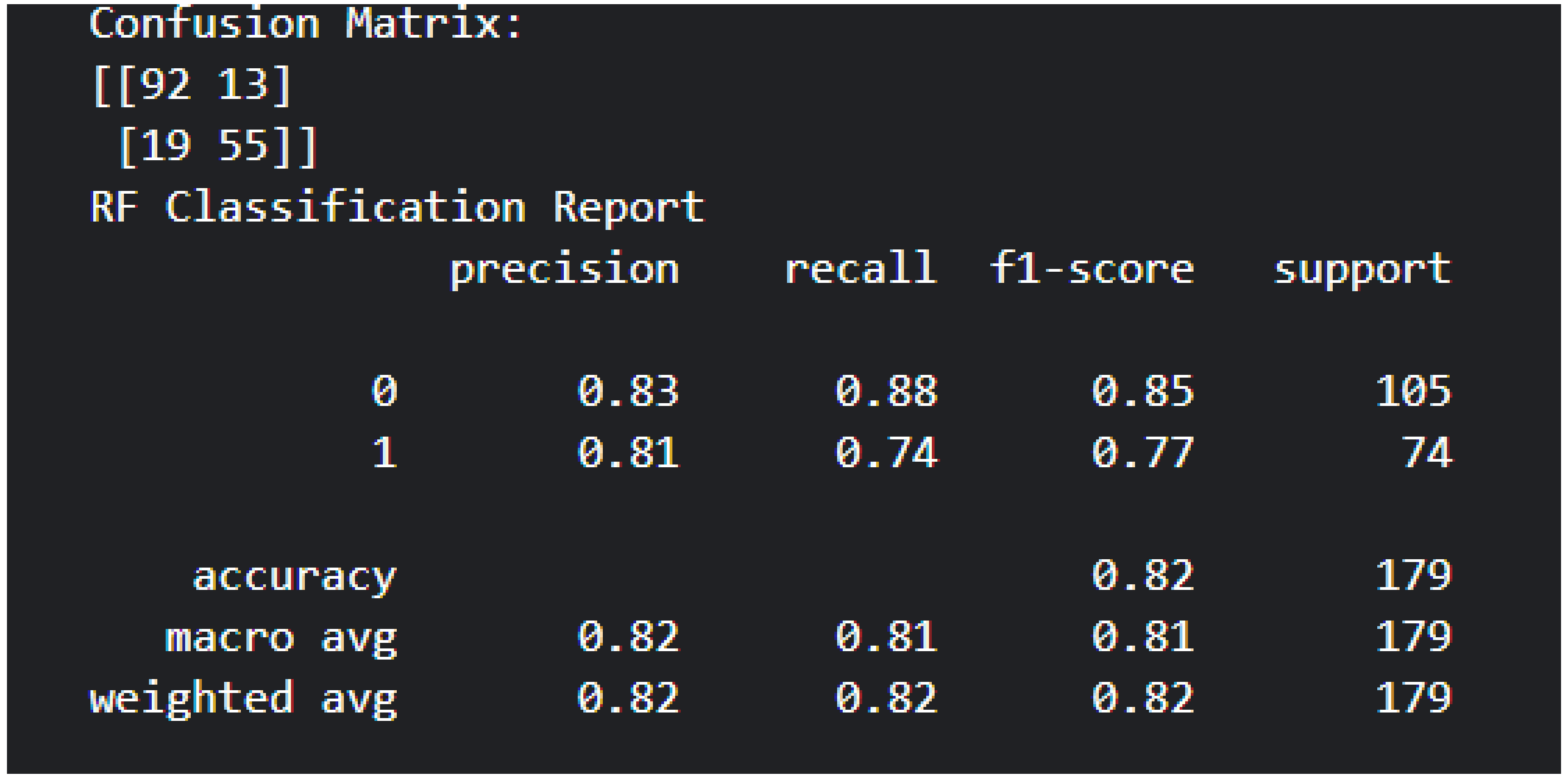
```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```



```
lr = LogisticRegression()
lr.fit(X_train, y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

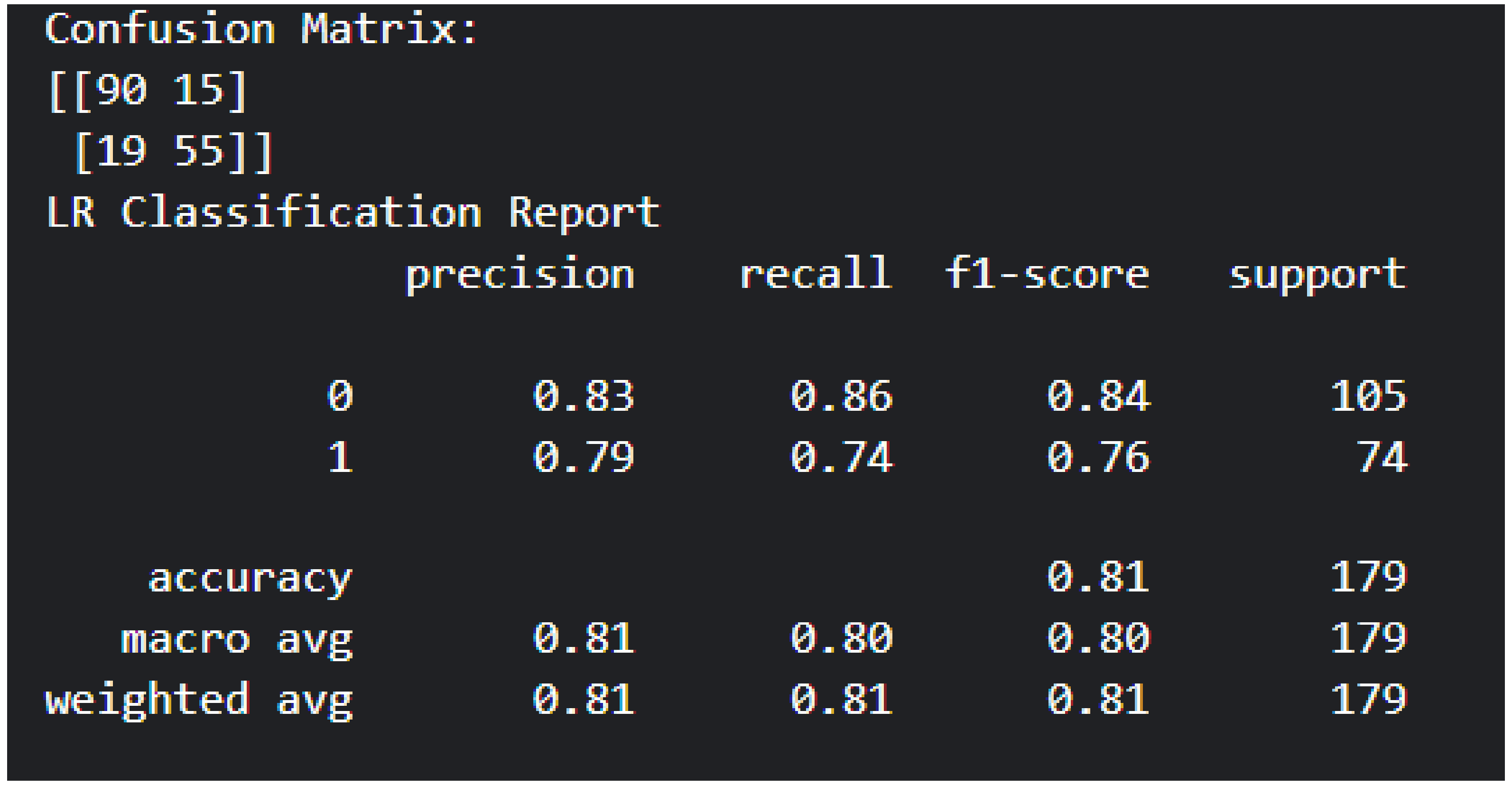
Data Modelling and Evaluation



```
Confusion Matrix:
[[92 13]
 [19 55]]
RF Classification Report
              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
```



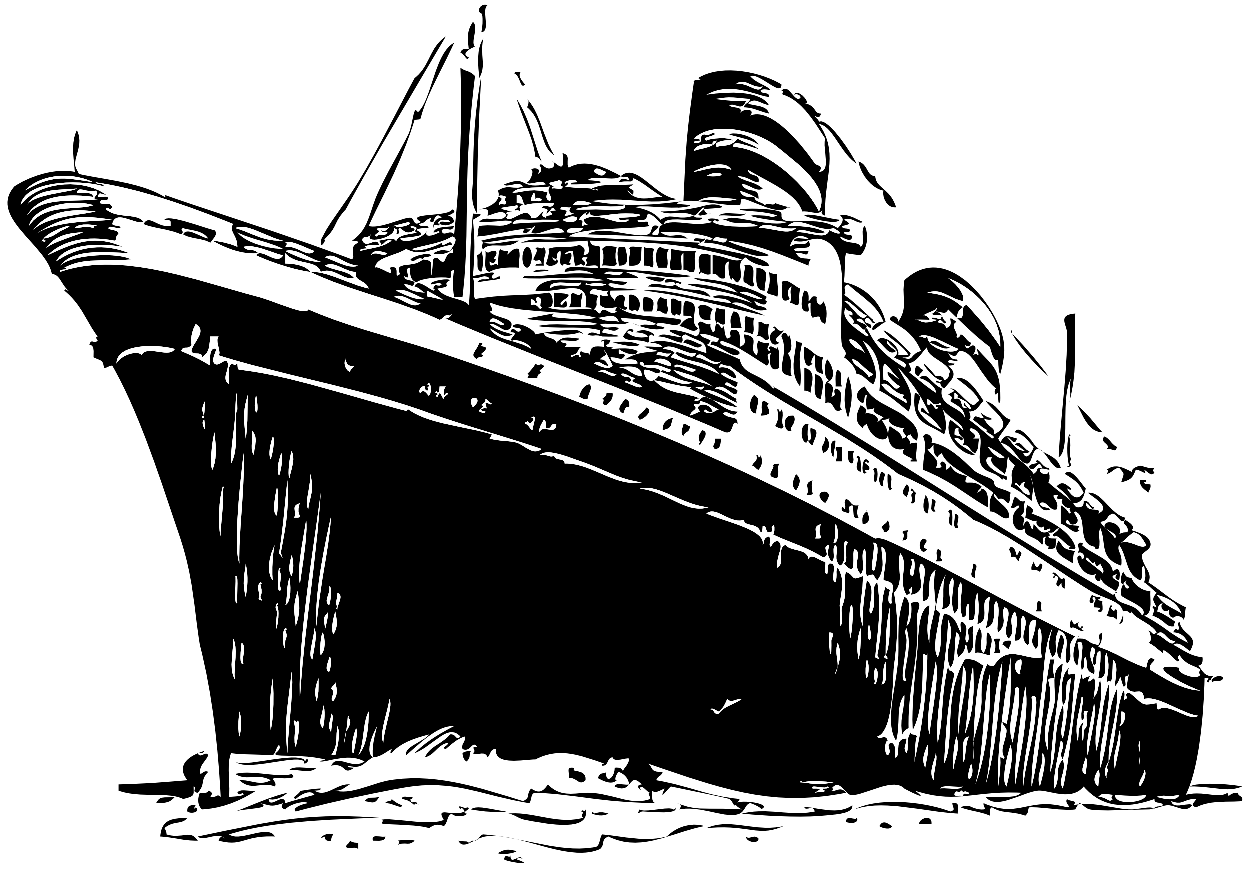
```
Confusion Matrix:
[[90 15]
 [19 55]]
LR Classification Report
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       105
           1       0.79      0.74      0.76        74

    accuracy                           0.81       179
   macro avg       0.81      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179
```

Conclusion

Random Forest Method is recommended for predicting Titanic Survival Passengers and have bigger accuracy score (82%) than Logistic Regression who has 81% accuracy score.



THANK YOU