A

MINI PROJECT REPORT

ON

“Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data.”

In the partial fulfillment of the requirement for the Degree in

Computer Engineering

Submitted By

4133 Dighe Gauri Vishwanath

4134 Dighe Pratik Sitaram

Under the Guidance of

Guide:- Prof. Pandit S. R.

Submitted for the Course of Final Year in Computer Engineering for the Practical Lab “Laboratory Practice III(Machine Learning)”

Department of Computer Engineering

AMRUTVAHINI COLLEGE OF ENGINEERING, SANGAMNER

For the academic year -2025

1. Title of Mini-Project: Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data

2. Aims/Benefits of Mini-Project: Building ML model to predict who survived the Titanic shipwreck using Random Forest Classifier.Build a machine learning model that predicts the type of people who survived the Titanic shipwreck using passenger data (i.e. name, age, gender, socio-economic class, etc.).

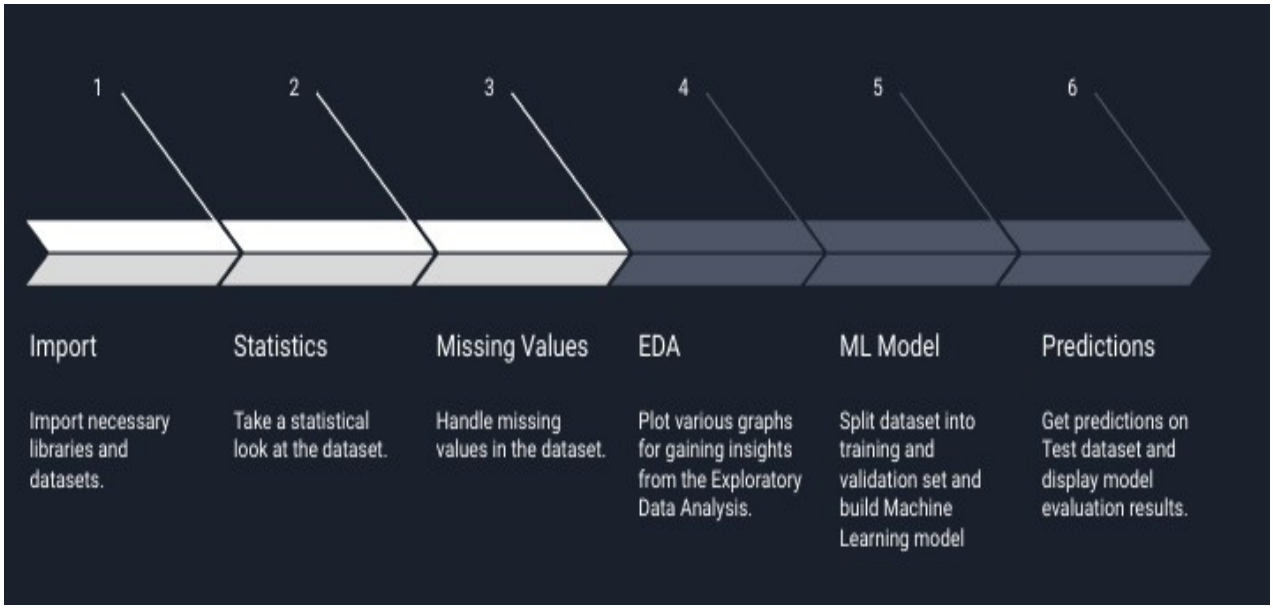
   Dataset Link: https://www.kaggle.com/competitions/titanic/data

3. Course Outcomes:

   1. To build a model for classification.
   2. To analyze its performance on Titanic Dataset.
   3. To use different ML and Feature Selection concepts to optimize the model's perfo rman ce.

4. Actual Methodology Followed:

Algorithm Implementation:



5. Actual Resources used:

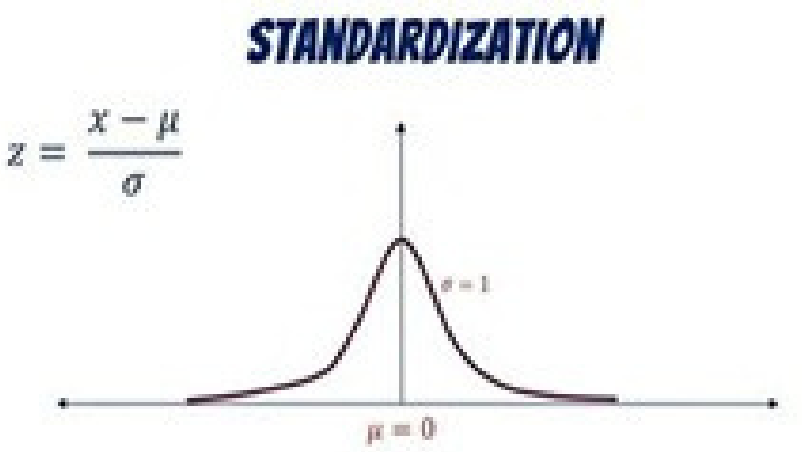
Sr. No	Name of resources/ Materials	Specifications	Qty.	Remark

1.	Computer System	10th generation, i5, RAM 8 GB, ROM- ITB	1	__
2.	Software	Jupyter Notebook	1	__
3.	Text editor	Microsoft word	1	__
4.	Information Source	https://www.geeksforgeeks.org/rabin-karp-algorithm-for-pattern-searching/ https://www.geeksforgeeks.org/naive-algorithm-for-pattern-searching/	1	__

Theory:

Binary Classification

Binary classification is a supervised learning algorithm that categorizes new observations into one of two classes:



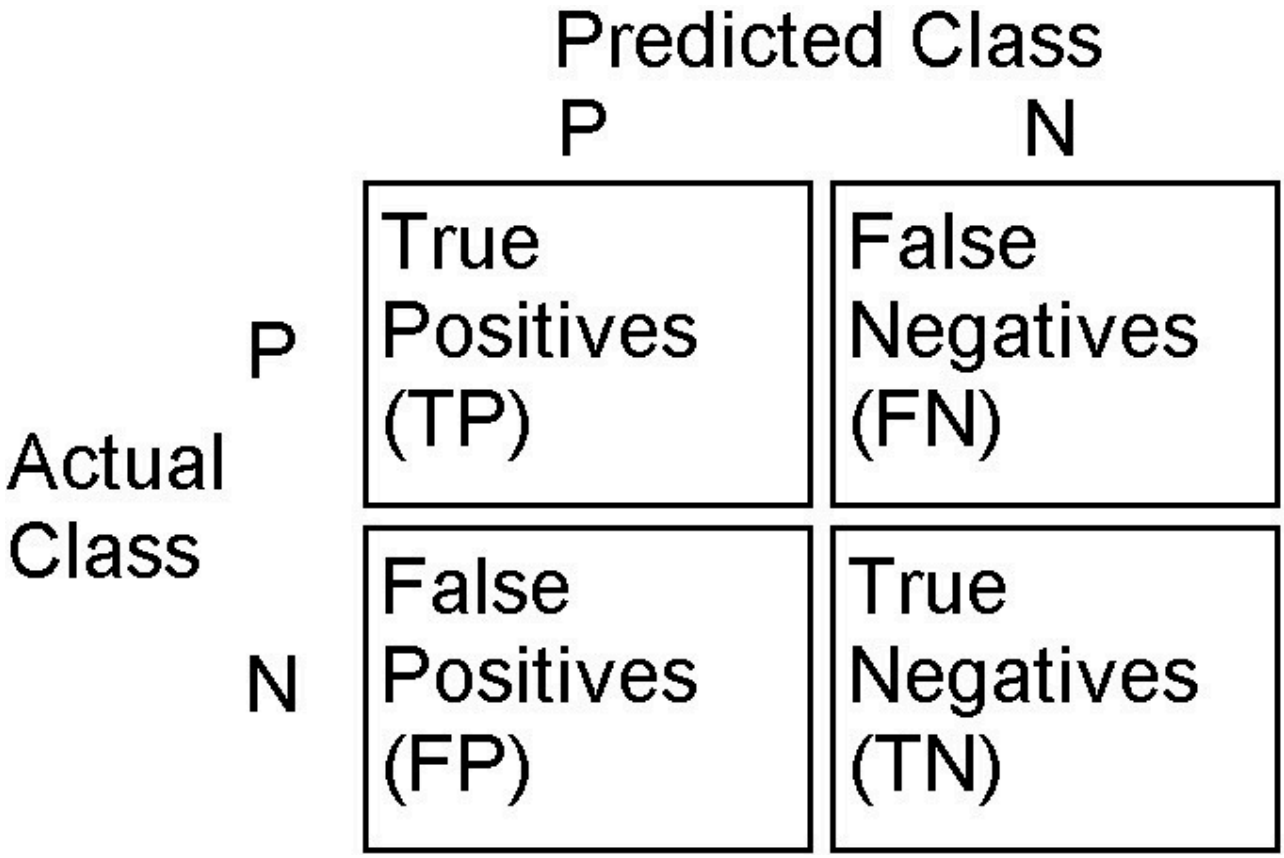
Standard Scaling

Standardization is a scaling technique where the value are manipulated such that it becomes centered around the mean with a unit standard deviation.

Confusion Matrix:

Confusion matrix is a very popular measure used while solving classification problems. It can be applied to binary classification as well as for multiclass classification problems. Confusion matrices represent counts from predicted and

actual values. The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.



The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are: 1.TP (True Positive): TP represents the number of patients who have been properly classified to have malignant nodes, meaning they have the disease. 2.TN (True Negative): TN represents the number of correctly classified patients who are healthy. 3.FP (False Positive): FP represents the number of misclassified patients with the disease but actually they are healthy. FP is also known as a Type I error. 4.FN (False Negative): FN represents the number of patients misclassified as healthy but actually they are suffering from the disease. FN is also known as a Type II error.

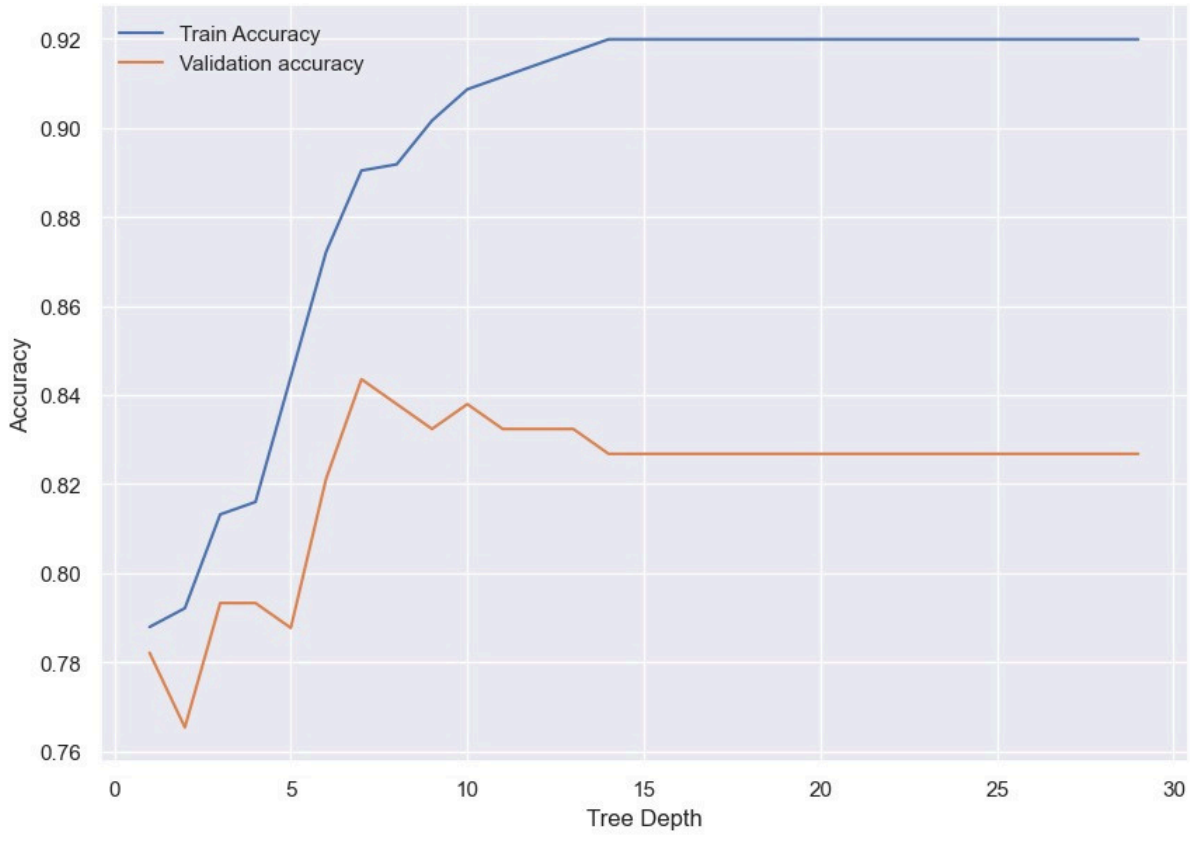
RandomForestClassifier:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. In random forests (see RandomForestClassifier and RandomForestRegressor classes), each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features or a random subset of size max_features. (See the parameter tuning guidelines for more details).

MODULES:

Numpy :For Line Algebra and Maths in the Notebook
Pandas:. For Data Processing in the Notebook.
Matplotlib :For Plotting charts and graphs for better visualization.
Seaborn :For lightweight, powerful visualization of data.
Scikit Learn :For model evaluation, preprocessing, Data Splitting etc.

7. Output:



	precision	recall	f1-score	support
0	0.84	0.88	0.86	105
1	0.81	0.76	0.78	74
accuracy			0.83	179
macro avg	0.82	0.82	0.82	179
weighted avg	0.83	0.83	0.83	179

```
array([0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0,
       1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1,
       1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1,
       0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0,
       1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1,
       1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
       0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0,
       1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0,
       1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0,
       0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0,
       1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1,
       0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0],
      dtype=int64)
```

8. Conclusion:

Hence, a machine learning model using Random Forest Classifier has been build, that predicts the type of people who survived the Titanic shipwreck using the passenger data.

9. CODE: