PROJECT REPORT:

TITANIC DATASET

1. Introduction

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.

2. Project Objective

(i) To build a predictive model that predict whether the passenger was survived or not.

(ii) Answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e, Age, Gender, passenger class, boarding place, cabin or not etc).

3. Data Preprocessing

(i) Finding and treating Null values.

(ii) Convert data types.

(iii) Feature Engineering/Extra New Features.

(iv) Finding and treating Duplicates values.

4. Model Building

(a) Logistic Regression

Logistic regression is a type of statistical model used for binary classification, which means it's used when you want to predict whether something is true or false, yes or no, 0 or 1. It's called "logistic" because it's based on the logistic function, which is an S-shaped curve that maps any real-valued number to a value between 0 and 1.

(b) Random Forest Classifier

A Random Forest classifier is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. Overall, Random Forest is a powerful and versatile machine learning algorithm that is widely used for classification and regression tasks due to its robustness and high performance.

(c) Decision Tree Classifier

A Decision Tree classifier is a machine learning model that uses a tree-like structure of decisions and their possible consequences to classify input data. It breaks down a dataset into smaller subsets based on different conditions and recursively constructs a tree to make predictions. Decision Tree classifiers are valuable for their simplicity, interpretability, and ability to handle a variety of data types, making them a popular choice for many machine learning tasks.

(d) Support Vector Machine

A Support Vector Classifier (SVC), also known as Support Vector Machine (SVM), is a powerful supervised learning algorithm used for classification tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space. SVC is a versatile and effective algorithm for classification tasks, especially in scenarios with high-dimensional data and complex decision boundaries.

5. Validation Technique/Evaluation Technique

(i) Confusion Matrix

(ii) Accuracy

(iii) Classification Report/F-1 Score

(iv) Receiver Operating Characteristic Curve

6. Final Model

Random Forest Classifier

7. Rationale behind Model Choice

(i) Random Forest and Decision Tree Classifier both have highest accuracy from the other models i.e, 85%.

(ii) But, ROC Curve of Random Forest Classifier becomes best from Decision Tree Classifier with low FPR, high TPR and area = 85.

(iii) And F-1 Score of the Random Forest Classifier is also higher than the Decision Tree Classifier model.

Hence, we concluded that Random Forest Classifier is overall best model for the problem with high accuracy, high F-1 score and good ROC Curve.

8. Insights from the project.

(i) Female Passengers had higher chance to survive than male passengers.

(ii) Class 1st Passengers had higher chance to survive rather than 2nd or 3rd class.

(iii) Passengers having 1 SibSp (No. of Sibling or Spouse) had higher chance to survive.

(iv) Passengers having 3 Parch (No. of Parents or Children) had higher chance to survive.

(v) Passengers who had a cabin had a higher chance of survival.

(vi) Passengers who had departure from Cherbourg had higher chance to survive.

(vii) Those Passengers whose age lied in between 1-18 had high chance to survive.

(viii) Those Passengers who paid high amount as fare, had higher chance to survive.