Titanic

Problem Description:

Build a predictive model for obtaining the analytical result from titanic dataset using Machine Learning Algorithms.

The sinking of the Titanic in 1912 remains one of the most infamous maritime disasters in history, resulting in significant loss of life. This tragic event has since sparked a multitude of inquiries and analyses, particularly within the realm of machine learning, aimed at understanding the dynamics of survival aboard the Titanic. The objective is to develop predictive models utilizing machine learning techniques to determine the likelihood of survival for passengers aboard the Titanic based on various attributes such as age, gender, ticket class, and embarkation port. We are analysis the dataset, worked on the feature engineering for obtaining the better accuracy.

Data:

Dataset: This Data was originally taken from Titanic: Machine Learning from Disaster. But its better refined and cleaned & some features have been self-engineered typically for Support Vector Machine and Random Forest.

Here's an overview of the dataset and its key attributes:

Attributes:

- PassengerId: A unique identifier for each passenger.
- Survived: Indicates whether the passenger survived (0 = No, 1 = Yes).
- Pclass: Ticket class (1 = First, 2 = Second, 3 = Third).
- Name: Passenger's name.
- Sex: Passenger's gender (male or female).
- Age: Passenger's age.
- SibSp: Number of siblings/spouses aboard.
- Parch: Number of parents/children aboard.
- Ticket: Ticket number.
- Fare: Passenger fare.
- Cabin: Cabin number.
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

This dataset has 1309 sample and 15 columns.

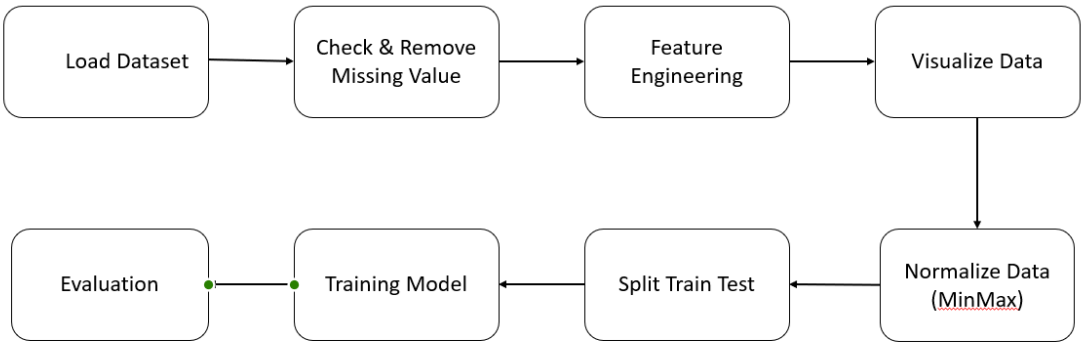
Solution:

Pattern recognition plays a crucial role in solving the Titanic project using machine learning techniques. The goal of pattern recognition in this context is to identify and leverage meaningful patterns or relationships within the dataset that can help predict whether a passenger survived or not based on various features. Here's a detailed description of how pattern recognition is applied in this project:

Understanding Patterns in Titanic Dataset

1. Feature Analysis:
    - Identifying Relevant Features: Pattern recognition involves analyzing the dataset to identify which features (such as age, gender, ticket class, etc.) are likely to influence survival outcomes.
2. Feature Engineering:
    - Creating New Features: Pattern recognition often involves creating new features from existing ones to capture hidden patterns or improve predictive power. For example, extracting titles from names or deriving a family size feature from the number of siblings/spouses and parents/children aboard.

Project Stages with Approachs:



Model evaluation and discussion:

I have used two machine learning algorithms e.x. Support Vector Machine & Random Forest. From both of these algorithms random forest (RF) work better in our dataset. I got 82.44% accuracy for random forest and 80.92% accuracy for support vector machine (SVM) on our test dataset.

Recall (Sensitivity): Recall measures the ability of a model to identify all relevant instances (true positives) within a dataset. It is calculated as the ratio of true positives (TP) to the sum of true positives and false negatives (FN).
A high recall indicates that the model is good at finding all positive instances, minimizing false negatives. For example, in a medical diagnosis scenario, recall would measure the proportion of actual positive cases (like disease presence) that the model correctly identifies.

Precision: Precision quantifies the accuracy of the positive predictions made by the model. It is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP).
Precision is a measure of exactness and tells us what proportion of positive identifications made by the model was actually correct. For instance, in a spam email detection task, precision would measure the proportion of emails flagged as spam that are genuinely spam.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a combined measure of both precision and recall into a single metric. The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.
The F1 score is particularly useful when you want to seek a balance between precision and recall and there is an uneven class distribution (e.g., many more negatives than positives or vice versa). This metric is widely used in binary classification tasks, such as fraud detection or disease diagnosis.

Confusion matrix, Precision, Recall and F1-Score of SVM:

	precision	recall	f1-score	support
0	0.83	0.85	0.84	155
1	0.77	0.76	0.76	107
accuracy			0.81	262
macro avg	0.80	0.80	0.80	262
weighted avg	0.81	0.81	0.81	262

Fig1: SVM

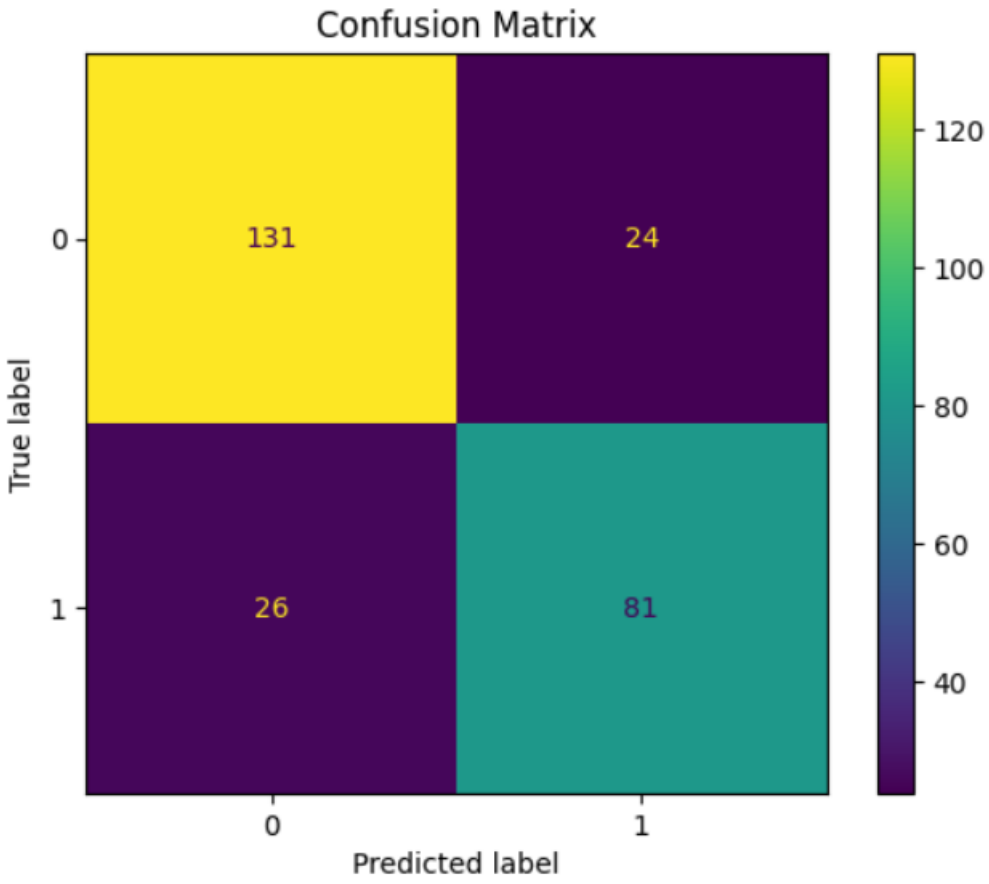


Fig2: SVM Confusion matrix

Confusion matrix, Precision, Recall and F1-Score of RF:

	precision	recall	f1-score	support
0	0.82	0.90	0.86	155
1	0.84	0.71	0.77	107
accuracy			0.82	262
macro avg	0.83	0.81	0.81	262
weighted avg	0.83	0.82	0.82	262

Fig3: RF

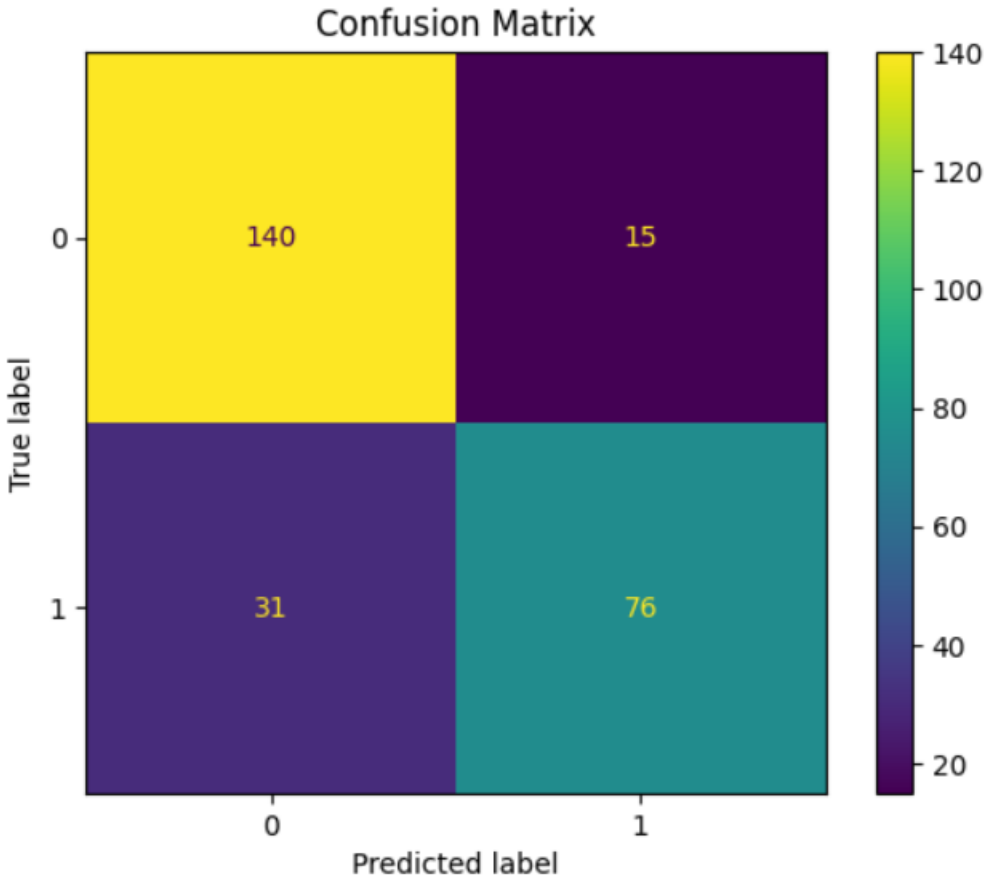


Fig4: RF confusion matrix

From these two models, we see that we get better performance on the random forest model based on precision, Recall, and F1-Scores.

Summary:

The Titanic project involved applying machine learning techniques to predict the survival outcomes of passengers on the Titanic based on various features such as age, gender, ticket class, and fare. Initially, data preprocessing was conducted, which included handling missing values and encoding categorical variables. Exploratory data analysis was performed to understand the distributions and correlations within the dataset. Feature engineering was crucial, where new features like family size were derived to enhance model performance. The dataset was split into training and testing sets for model development and evaluation. Several machine learning algorithms such as random forest, and support vector machines were trained and optimized using techniques like hyperparameter tuning. The models were evaluated based on metrics like accuracy, precision, recall, and F1-score. Ultimately, the random forest algorithm emerged as the best performer with an accuracy of 82.44% on the test set. The project highlighted the importance of feature engineering and model selection in improving prediction

accuracy. Future work could involve more advanced techniques such as neural networks or exploring additional data sources to further enhance predictive capabilities.