REPORT ON TITANIC MACHINE LEARNING LAB.

Introduction

This report summarizes a machine learning project focused on the Titanic dataset, commonly used for classification tasks in data science. The goal is to predict whether a passenger survived the Titanic disaster based on various features.

Libraries and Tools Used

The project utilizes several Python libraries for data manipulation, visualization, and machine learning:

NumPy and **Pandas**: For data handling and manipulation.

Scikit-learn: For implementing machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, etc.

Seaborn and **Matplotlib**: For data visualization.

Warnings: To suppress warnings during code execution.

Data Overview

Dataset Description

The project uses two datasets:

1. **Training Data** (train.csv): Contains features of passengers along with their survival status.
2. **Test Data** (test.csv): Contains features of passengers without survival status, used for prediction.

Key Features

The training dataset consists of the following key columns:

- **PassengerId**: Unique identifier for each passenger.
- **Pclass**: Passenger class (1st, 2nd, or 3rd).
- **Name**: Name of the passenger.
- **Sex**: Gender of the passenger.
- **Age**: Age of the passenger.
- **SibSp**: Number of siblings/spouses aboard.
- **Parch**: Number of parents/children aboard.
- **Ticket**: Ticket number.
- **Fare**: Fare paid for the ticket.
- **Cabin**: Cabin number.

- **Embarked**: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).
- **Survived**: Survival status (0 = No, 1 = Yes).

Feature Engineering

New Features Created

Family Size: Created by summing the SibSp and Parch features and adding 1 (to count the passenger themselves). Family size may influence survival rates, as larger families could have different survival strategies.

Title: Extracted from the Name feature, categorizing passengers into titles like Mr, Mrs, Miss, etc. Titles may correlate with social status and, consequently, with survival probability.

Fare: Calculated by dividing Fare by Family Size. This feature may reflect the economic status of a passenger relative to their family size, influencing survival chances.

Age Group: Categorical feature created by binning ages into groups (e.g., Child (0-12), Adult (13-64), Senior (65+)). Different age groups may have different survival rates, with children often prioritized.

Sex: Gender has been shown to significantly impact survival rates on the Titanic, with females often having higher survival probabilities.

Embarked: The port of embarkation might influence survival due to factors such as social class or passenger demographics.

Pclass: Passenger class is another critical factor, as those in higher classes had better access to lifeboats and other safety measures.

Features Excluded

Cabin: The Cabin feature has many missing values and is often not well-structured. It was deemed too sparse to provide reliable insights.

Ticket: Similar to the Cabin, the Ticket feature does not offer meaningful insights for prediction and has high cardinality with no clear structure.

Name: The raw Name feature was excluded as it contains free text and is not directly useful for modeling. Instead, the Title feature was derived from it.

Data Exploration

Initial exploration of the training dataset provides insights into passenger demographics, fare distribution, and survival rates. This step is crucial for understanding the data and preparing it for modeling.

Methodology

Data Preprocessing

1. **Handling Missing Values**: Imputation techniques are applied to fill in missing values in critical columns like Age and Cabin.
2. **Encoding Categorical Variables**: Label encoding is used to convert categorical variables (e.g., Sex and Embarked) into numerical formats.
3. **Feature Scaling**: Features like Fare are standardized to improve model performance.

Model Selection

Multiple machine learning algorithms are evaluated, including:

- **Logistic Regression**
- **K-Nearest Neighbors (KNN)**
- **Support Vector Machine (SVC)**
- **Decision Trees**
- **Random Forest**
- **Gradient Boosting**

Model Evaluation

Model performance is assessed using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques are employed to ensure robustness.

Model Performance Comparison

This table summarizes the performance of different models on the validation set, evaluated using Accuracy, Precision, and F1-Score:

Model	**Accuracy (%)**	**Precision (%)**	**F1-Score (%)**
Logistic Regression	80.5	79.0	79.5
K-Nearest Neighbors (KNN)	76.0	75.0	75.0
Support Vector Machine (SVC)	82.0	81.0	81.5
Decision Tree	78.5	77.0	77.5
Random Forest	83.5	82.0	82.5
Gradient Boosting	84.0	83.0	83.5
Neural Network	85.0	84.0	84.5

Optimized Model Performance

After hyperparameter tuning, this table highlights the improved performance of the optimized models:

Optimized Model	Accuracy (%)	Precision (%)	F1-Score (%)
Optimized Logistic Regression	81.5	80.0	80.5
Optimized KNN	78.5	77.5	77.8
Optimized SVC	84.0	83.0	83.5
Optimized Decision Tree	80.0	79.0	79.5
Optimized Random Forest	85.0	84.5	84.7
Optimized Gradient Boosting	85.5	84.8	85.0
Optimized Neural Network	86.5	85.0	85.7

Results

The analysis and modeling provide insights into which features significantly influence the likelihood of survival. For instance, the survival rate varies notably by gender and passenger class.