Project Title: Titanic Survival Prediction

Name:Shakil Hossain

Cource:Python Programming in Data Science

Batch:B1

Introduction

This project aims to predict the survival of passengers aboard the Titanic using machine learning models. Various factors such as age, gender, passenger class, and fare amount are used as input features to predict whether a passenger survived or not. The goal is to compare the performance of different machine learning algorithms and select the best-performing model.

Data Feature

The dataset has been collected from Kaggle(Titanic - Machine Learning from Disaster | Kaggle)

The dataset used for this project is train.csv, which contains features related to the passengers. The key features are as follows:

- **PassengerId**: Unique identifier for each passenger (dropped during analysis).
- **Survived**: Target variable indicating survival (1 = survived, 0 = did not survive).
- **Pclass**: Passenger's class (1st, 2nd, or 3rd class).
- **Name**: Name of the passenger (dropped during analysis).
- **Sex**: Gender of the passenger.
- **Age**: Age of the passenger (177 missing values).
- **SibSp**: Number of siblings or spouses aboard the Titanic.
- **Parch**: Number of parents or children aboard.
- **Ticket**: Ticket number (dropped during analysis).
- **Fare**: The fare paid by the passenger.
- **Cabin**: Cabin number (687 missing values, dropped during analysis).

- **Embarked**: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton, 2 missing values).

Models and Algorithms

The following machine learning models have been used to predict the survival of passengers:

1. **Logistic Regression**

   Algorithm: Logistic Regression

2. **K-Nearest Neighbors (KNN)**

   Algorithm: K-Nearest Neighbors

3. **Gradient Boosting Classifier**

   Algorithm: Gradient Boosting Classifier

Evaluation Metrics

The models were evaluated using the following metrics:

- **Accuracy**: The proportion of correct predictions out of all predictions.
- **Precision**: The ratio of correctly predicted positive observations (survival) to the total predicted positives.
- **Recall**: The ratio of correctly predicted positives to all actual positive observations.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.

Results Comparison

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Logistic Regression	81	79	72	75

Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
K-Nearest Neighbours (KNN)	65	63	35	45
Gradient Boosting Classifier	80	84	64	72

From the results, **Logistic Regression** is the best-performing model overall, balancing all metrics effectively, while **Gradient Boosting Classifier** offers slightly higher precision but lower recall.

Conclusion

In this project, three machine learning models—**Logistic Regression**, **K-Nearest Neighbors**, and **Gradient Boosting Classifier**—were applied to the Titanic survival dataset. After comparing the models, **Logistic Regression** was found to be the best performer with an accuracy of 81% and an F1-Score of 75%. Gradient Boosting also performed well, particularly with high precision, but had a slightly lower recall.