Titanic Classification Project Report

Objective

The goal of this project is to build a predictive model to determine the likelihood of survival for passengers on the Titanic using data science techniques in Python.

Data Overview

The dataset used in this project is the Titanic dataset, which contains information about passengers on the Titanic. Key features include:

- `Pclass`: Passenger class
- `Sex`: Gender of the passenger
- `Age`: Age of the passenger
- `SibSp`: Number of siblings/spouses aboard
- `Parch`: Number of parents/children aboard
- `Fare`: Fare paid by the passenger
- `Embarked`: Port of embarkation
- `Survived`: Target variable indicating whether the passenger survived (1) or not (0)

Data Preprocessing

Handling Missing Values

- Filled missing values in `Age` with the median age.
- Filled missing values in `Embarked` with the mode (most frequent value).

Feature Engineering

- **Family Size**: Created a new feature `FamilySize` by summing `SibSp` and `Parch`.
- **Title**: Extracted `Title` from `Name` to capture social status.

Feature Selection

Selected features for the model:

- `Pclass`

- `Sex`
- `Age`
- `FamilySize`
- `Fare`
- `Embarked`

Dropped features:

- `Name`
- `Ticket`
- `Cabin`

Modeling

Models Evaluated

1. **Logistic Regression**
2. **Random Forest**
3. **SVM**
4. **Gradient Boosting**
5. **Naive Bayes**
6. **KNN**

Evaluation Metrics

- **Accuracy**: proportion of correctly predicted instances.
- **Precision:** proportion of true positives among the predicted positives.
- **Recall:** proportion of actual positives correctly identified.
- **F1 Score:** harmonic mean of precision and recall.

Model Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7989	0.7794	0.7162	0.7465
Random Forest	0.8212	0.7917	0.7703	0.7808
SVM	0.8156	0.8154	0.7162	0.7626
Gradient Boosting	0.8156	0.8154	0.7162	0.7626
Naive Bayes	0.7765	0.7125	0.7703	0.7403

KNN	0.8101	0.7857	0.7432	0.7639

Optimized Random Forest

- **Parameter Tuning**: Used GridSearchCV to find optimal parameters.
- **Best Parameters**:
  - `n_estimators`: 300
  - `max_depth`: 30
  - `min_samples_split`: 10
- **Performance**:
  - **Accuracy**: 0.8492
  - **Precision**: 0.8615
  - **Recall**: 0.7568
  - **F1 Score**: 0.8058

Confusion Matrix

The confusion matrix for the optimized Random Forest model is shown below, which visualizes the performance of the classification model:

Prediction Function

A function was created to predict survival based on user input. Here is an example input:

- **Pclass**: 3
- **Sex**: male
- **Age**: 22
- **Family Size**: 1
- **Fare**: 7.25
- **Embarked**: S

Prediction: Did not survive

Conclusion

The Random Forest model, after hyperparameter tuning, achieved strong performance with an accuracy of 84.92%, precision of 86.15%, recall of 75.68%, and an F1 score of 80.58%. The model demonstrated effective prediction capabilities for the Titanic survival problem.