### **Internship Task 2: Predictive Analysis Using Machine Learning**

#### **Introduction**

This report documents the process of building a predictive machine learning model using the Titanic dataset. The objective was to predict whether a passenger survived the Titanic disaster based on features such as age, sex, class, fare, and port of embarkation.

#### **Dataset Description**

The Titanic dataset contains detailed information on 891 passengers. The features include:

* **Pclass**: Ticket class (1st, 2nd, 3rd)
* **Sex**: Gender of the passenger
* **Age**: Age in years
* **SibSp**: Number of siblings/spouses aboard
* **Parch**: Number of parents/children aboard
* **Fare**: Ticket fare
* **Embarked**: Port of embarkation (C, Q, S)
* **Survived**: Target variable (0 = No, 1 = Yes)

Unnecessary columns like PassengerId, Name, Ticket, and Cabin were removed to reduce noise.

#### **Feature Selection and Preprocessing**

* Missing values in the **Age** column were filled with the mean age.
* Missing values in **Embarked** were filled with the most frequent value (mode).
* Categorical columns like **Sex** and **Embarked** were label-encoded.
* The features used for prediction included: Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked.

#### **Model Building**

A **Logistic Regression** model was used due to its simplicity and interpretability in binary classification tasks. The dataset was split into:

* **80% Training Data**
* **20% Testing Data**

The model was trained on the training set and evaluated on the test set.

#### **Model Evaluation**

The model was evaluated using the following metrics:

* **Accuracy**: 0.81
* **Classification Report**: Provided detailed performance including precision, recall, and F1-score for both survival and non-survival classes.
* **Confusion Matrix**: Visualized the number of true positives, true negatives, false positives, and false negatives.

The confusion matrix showed good performance, with the model correctly predicting a high number of both survivors and non-survivors.

#### **Conclusion**

The logistic regression model provided a strong starting point for predictive analysis using machine learning. With minimal preprocessing, we were able to reach an accuracy of over 80%. This task demonstrated the complete ML workflow:

* Data loading
* Cleaning & encoding
* Feature selection
* Model training
* Evaluation and interpretation

Further improvements can be made by:

* Using advanced models (Random Forest, XGBoost)
* Hyperparameter tuning
* Cross-validation
* Feature engineering