



**ML and Big Data PBL 2023**

**TITANIC PREDICTION**

Name	Enrollment	Batch
Aditi Mahabole	20103023	B1
Molshree Sharma	20103060	B2
Arushi Sharma	20103010	B1
Stuti Sharma	20103043	B2

**Submitted to: Tarun Agarwal Sir.**

## ***Problem Statement:***

The Titanic Survival Prediction involves predicting the survival outcomes of passengers aboard the Titanic based on various factors. The dataset provides details about passengers, such as age, gender, class, and more. The objective is to develop a machine learning model that accurately predicts whether a passenger survived or not during the tragic incident.

## ***Objective:***

The primary goal is to create a predictive model that generalizes well to unseen data, offering insights into the determinants of survival on the Titanic. This model contributes to the ongoing analysis of the Titanic dataset and enhances our understanding of the factors influencing survival during the disaster.

## **Tech Stack:**

### **Libraries**

- Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn

### **Modeling Algorithms**

- Logistic Regression
- Random Forest
- Support Vector Machines

### **Data Processing**

- Handling missing values
- Data visualization using Seaborn and Matplotlib
- Feature engineering and scaling using Scikit-learn

### **Evaluation Metrics**

- Accuracy Score

## Data Processing:

- Explored the dataset using Pandas and Seaborn.
- Identified missing values and implemented smart strategies for handling them.
- Visualized the correlation matrix using Seaborn to understand relationships between variables.
- Converted categorical variables into numerical format.
- Split the dataset into training and testing sets.

## Modelling:

- Implemented three machine learning algorithms: Logistic Regression, Random Forest, and Support Vector Machines.
- Standardized the features using Scikit-learn's StandardScaler.
- Trained each model using the training dataset.
- Evaluated model performance using the accuracy score.

## Results:

- Logistic Regression: Achieved an accuracy of **0.7988826815642458** on the test set.
- Random Forest: Achieved an accuracy of **0.8212290502793296** on the test set.
- Support Vector Machines: Achieved an accuracy of **0.8044692737430168** on the test set.

## Model Comparison:

- Logistic Regression, Random Forest, and Support Vector Machines were evaluated.
- The Random Forest model demonstrated the highest accuracy among the three algorithms.

## Conclusion:

The Random Forest model outperformed Logistic Regression and Support Vector Machines in predicting survival outcomes for passengers on the Titanic. This analysis contributes to the broader understanding of the factors influencing survival during historical disasters. Further refinement and optimization of the model could enhance its predictive capabilities.