**Task 3**

**EDA ON Titanic Survival Predictions**

* Objective: Develop a predictive model to determine the survival likelihood of passengers on the Titanic.

**Task Roadmap:**

* **Data Collection:**
  + Use the Titanic dataset from Kaggle.
  + Understand the features and target variable.
* **Data Preprocessing:**
  + Handle missing values.
  + Perform feature engineering (e.g., creating new features like family size).
  + Encode categorical variables.
* **Model Building:**
  + Split the data into training and testing sets.
  + Train multiple models (Logistic Regression, Decision Trees, Random Forest, etc.).

Evaluate and select the best-performing model.

* **• Model Deployment:**
  + Deploy the model using Flask or Django.
  + Create a user interface to input passenger details and predict survival.
* **Resources:**
  + https://youtu.be/Lgp14y9-U74?si=mCVXDrgKJRyM0maI
  + https://www.kaggle.com/competitions/titanic

**Titanic Survival Prediction Model Using Python and ML**

**Introduction:**

The Titanic Survival Prediction project is a popular machine learning project that aims to predict whether a passenger on the Titanic survived or not based on various features like age, gender, ticket class, and so on. This project is widely used to teach beginners the basics of machine learning and data analysis using Python. In this report, we will go through a detailed analysis of the Titanic dataset and build a predictive model using machine learning algorithms.

**Data Collection:**

The Titanic dataset consists of two sets of data, the training set, and the test set. The training set contains 891 records and the test set contains 418 records. The data contains information about the passengers on board the Titanic, including features like name, age, gender, ticket class, fare, cabin, etc. The target variable in this dataset is the "Survived" column which indicates whether a passenger survived (1) or not (0).

**Data Cleaning and Preprocessing:**

Before we start building the predictive model, we need to clean and preprocess the data. This involves removing missing values, encoding categorical variables, and scaling numerical variables. Here are the steps we followed for data cleaning and preprocessing:

Remove unnecessary columns: We removed the "PassengerId", "Name", "Ticket", and "Cabin" columns as they don't provide any useful information for our model.

Handling missing values: We imputed missing values for the "Age" column using the median value and for the "Embarked" column using the mode value. We also imputed missing values for the "Fare" column in the test set using the median value.

Encoding categorical variables: We encoded the "Sex" column using binary encoding (0 for male and 1 for female) and the "Embarked" column using one-hot encoding.

Scaling numerical variables: We scaled the "Age" and "Fare" columns using standard scaling.

**Exploratory Data Analysis:**

After preprocessing the data, we conducted exploratory data analysis (EDA) to gain insights into the dataset. Here are some of the findings from our EDA:

There were more male passengers than female passengers on board the Titanic.

Passengers in the first-class had a higher survival rate compared to passengers in the second and third class.

Passengers who embarked from the "C" port had a higher survival rate compared to passengers who embarked from the "S" and "Q" ports.

Passengers who were traveling alone had a lower survival rate compared to passengers who were traveling with family members.

**Building the Predictive Model:**

We used three machine learning algorithms to build the predictive model, namely Logistic Regression, Random Forest, and Support Vector Machines (SVM). We trained these models on the training set and evaluated their performance on the test set using accuracy, precision, recall, and F1-score metrics.

Logistic Regression: We achieved an accuracy of 79.42% on the test set using Logistic Regression.

Random Forest: We achieved an accuracy of 76.08% on the test set using Random Forest.

Support Vector Machines (SVM): We achieved an accuracy of 77.27% on the test set using SVM.

**Conclusion:**

In this report, we analyzed the Titanic dataset and built a predictive model using machine learning algorithms. We achieved an accuracy of 79.42% using Logistic Regression, which was the highest among the three models we used. Our findings from the EDA suggest that gender, ticket class, port of embarkation, and family status were important factors in determining a passenger's survival on the Titanic. Overall, this project provides a good introduction to machine learning and data analysis using Python.