Logistic Regression

**1. Data Exploration:**

a. Load the dataset and perform exploratory data analysis (EDA).

b. Examine the features, their types, and summary statistics.

c. Create visualizations such as histograms, box plots, or pair plots to visualize the distributions and relationships between features.

Analyze any patterns or correlations observed in the data.

**2. Data Preprocessing:**

a. Handle missing values (e.g., imputation).

b. Encode categorical variables.

**3. Model Building:**

a. Build a logistic regression model using appropriate libraries (e.g., scikit-learn).

b. Train the model using the training data.

**4. Model Evaluation:**

a. Evaluate the performance of the model on the testing data using accuracy, precision, recall, F1-score, and ROC-AUC score.

Visualize the ROC curve.

**5. Interpretation:**

a. Interpret the coefficients of the logistic regression model.

b. Discuss the significance of features in predicting the target variable (survival probability in this case).

**6. Deployment with Streamlit:**

In this task, you will deploy your logistic regression model using Streamlit. The deployment can be done locally or online via Streamlit Share. Your task includes creating a Streamlit app in Python that involves loading your trained model and setting up user inputs for predictions.

(optional)For online deployment, use Streamlit Community Cloud, which supports deployment from GitHub repositories.

Detailed deployment instructions are available in the Streamlit Documentation.

<https://docs.streamlit.io/streamlit-community-cloud/deploy-your-app>

**Interview Questions:**

1. What is the difference between precision and recall?

2. What is cross-validation, and why is it important in binary classification?

**Answer:**

<https://colab.research.google.com/drive/1d_Mhj9E-hBlT9OWKQByQqTlQKMEVnmw6?usp=sharing>

**Ans 1**) In logistic regression, precision and recall are both evaluation metrics used to assess the performance of a binary classification model.  
  
Precision measures the proportion of true positive predictions among all positive predictions made by the model. Mathematically, it's defined as the ratio of true positives to the sum of true positives and false positives.  
Precision =True Positives / (True Positives +False Positives)  
Precision=True Positives / (False Positives +True Positives)  
​   
Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances in the dataset. Mathematically, it's defined as the ratio of true positives to the sum of true positives and false negatives.  
Recall =True Positives / (True Positives +False Negatives)

**Ans 2)** Cross-validation is a technique used in machine learning to assess how well a predictive model generalizes to an independent dataset. In binary classification, it involves splitting the dataset into multiple subsets or folds, training the model on a subset of the data, and evaluating its performance on the remaining data.

Cross-validation is crucial in binary classification for performance estimation, model selection, and preventing overfitting. It provides a reliable estimate of a model's performance by utilizing multiple data splits, reducing variance and indicating model generalization. It also helps identify the best model on average by systematically evaluating each model across multiple data folds.