Logistic Regression

1. **Data Exploration**

**a. Loading the Dataset and Performing EDA**

Exploratory Data Analysis (EDA) is the initial step in understanding the structure, patterns, and anomalies within a dataset. It helps identify important variables, data types, outliers, and missing values.

**b. Examining Features, Types, and Summary Statistics**

* **Data Types**: Categorical, Numerical (Continuous or Discrete).
* **Summary Statistics**: Mean, Median, Standard Deviation, Min, Max, Quartiles.
* df.info() and df.describe() functions help inspect data structure and numerical summaries.

**c. Visualizations**

* **Histograms**: Show the distribution of numerical features.
* **Box Plots**: Identify outliers and compare distributions.
* **Pair Plots**: Examine relationships between features, especially how they vary by the target variable.
* **Correlation Heatmap**: Shows strength and direction of relationships between numerical variables.

**2. Data Preprocessing**

**a. Handling Missing Values**

* **Numerical Features**: Impute using mean, median, or predictive models.
* **Categorical Features**: Impute with mode or create an "Unknown" category.
* Use SimpleImputer or fillna() for basic imputation.

**b. Encoding Categorical Variables**

* **One-Hot Encoding**: Convert categorical variables into binary indicators.
* **Label Encoding**: Assign numerical codes to categories (use carefully).

**3. Model Building**

**a. Building the Logistic Regression Model**

Logistic Regression is a supervised learning algorithm used for binary classification. It predicts the probability of a binary outcome using the logistic (sigmoid) function.

**b. Training the Model**

Using libraries like scikit-learn, the model is trained on a portion of the data (X\_train, y\_train) and tested on unseen data (X\_test, y\_test) to evaluate performance.

**4. Model Evaluation**

**Evaluation Metrics**

* **Accuracy**: Overall percentage of correct predictions.
* **Precision**: Proportion of predicted positives that are actual positives.
* **Recall (Sensitivity)**: Proportion of actual positives that are correctly predicted.
* **F1-Score**: Harmonic mean of precision and recall.
* **ROC-AUC Score**: Measures the area under the ROC curve; higher is better.

**ROC Curve**

The ROC curve plots the **True Positive Rate** against the **False Positive Rate**. A model with good predictive power will have a curve closer to the top-left corner.

**5. Interpretation**

**a. Coefficients Interpretation**

In logistic regression, each feature has a coefficient:

* **Positive Coefficient**: Increases the log-odds of the target being 1.
* **Negative Coefficient**: Decreases the log-odds.
* Coefficients can be exponentiated to get the **odds ratio**.

**b. Feature Significance**

The magnitude and sign of the coefficients indicate feature importance:

* Larger absolute values → more influence on the target.
* Statistically significant coefficients (low p-values) imply strong evidence that the feature affects the target.

**6. Deployment with Streamlit**

**Local Deployment Steps:**

1. Save the model using pickle.
2. Build a UI using Streamlit to accept user input.
3. Load the model and scaler.
4. Display prediction results.

**Online Deployment (Optional):**

* Host the project on GitHub.
* Deploy on Streamlit Community Cloud.
* Follow official documentation for setup.