**Report**

1. **Key Preprocessing steps taken:**

* **Categorical Feature Encoding:**

Since many columns in the dataset were categorical these were converted into numerical representation using Label Encoding. This transformation was applied to age, medical\_speciality, diag\_1, diag\_2, diag\_3, change.

Label Encoding assigns unique numerical values to each category, making these features usable by machine learning algorithms.

* **Splitting the Dataset**: The dataset was divided into a trainingset (70%) and a testingset (30%) using the train\_test\_split function from sklearn. This split ensures that the model is trained on one portion of the data and evaluated on a separate, unseen portion to assess performance.
* **Mapping the target variable**:

Since we use Supervised learning classification, we must map the target variable. For example, Yes was replaced by 1, and No was replaced by 0.

1. **Model Choice and Rationale**

**Logistic Regression** was selected for this task because:

1. It is ideal for binaryclassification, like predicting hospital readmission.
2. It is efficient and handles large datasets well.
3. It serves as a strong baselinemodel to compare with more complex models.

This model was chosen for its balance of simplicity, performance, and insight into the impact of key features.

1. **Performance Metrics of the Model**

The Logistic Regression model was evaluated using the following performance metrics:

1. **Accuracy**: Measures the percentage of correct predictions over the total number of predictions.
2. **Precision**: Indicates the proportion of true positive predictions out of all positive predictions made by the model.
3. **Recall (Sensitivity)**: Reflects the model's ability to correctly identify actual positives from the dataset.
4. **F1-Score**: The harmonic mean of precision and recall, providing a balanced metric when both are important.
5. **ROC-AUC Score**: Evaluates the model’s ability to distinguish between the positive and negative classes, where a higher score indicates better performance.
6. **Theoretical Explanation of Logistic Regression**

Logistic Regression is a statistical model used for binary classification, predicting the probability of an outcome belonging to one of two classes. The model is based on the logistic (sigmoid) function:

* **Training**: It uses maximum likelihood estimation (MLE) to optimize parameter weights and bias by minimizing the log loss function. This is typically done through gradient descent.
* **Interpretation**: The coefficient weight indicates the influence of each feature on the likelihood of the positive class.

1. **Suggested improvements to your model and why they might work.**

In this case, I used the XGBoost Classifier algorithm to improve the model accuracy, as XGBoost is a Boosting algorithm, the output of the previous weak model is fed to the input of the next weak learner. So there is improvement with each step.