**Name: Tan Nguyen**

**CS 422**

**INTRODUCTION TO MACHINE LEAERNING**

**FALL 2023**

**ASSIGNMENT 3**

**Machine Learning Report: Heart Disease Prediction Using Logistic Regression and Naïve Bayes Models**

# Dataset

**Note to TA: Dr. Fatma Nasoz agreed for me using different datasets to assignment 3 and doing one-hot encoding**

The dataset under consideration is sourced from the University of California Irvine's Machine Learning Repository, specifically from the Heart Disease dataset, accessible at [UCI's Heart Disease Dataset](https://archive.ics.uci.edu/ml/datasets/Heart+Disease).

It comprises 918 records, each with 11 features capturing a range of patient demographics and clinical measurements. The features include age, sex (denoted by 0 or 1 after one-hot encoding), types of chest pain, resting blood pressure (RestingBP), serum cholesterol levels (Cholesterol), fasting blood sugar levels (FastingBS), resting electrocardiographic results (RestingECG), maximum heart rate achieved (MaxHR), exercise-induced angina (ExerciseAngina), ST depression induced by exercise relative to rest (Oldpeak), and the slope of the peak exercise ST segment (ST\_Slope).

The dataset also contains a label column named 'HeartDisease,' where a value of 0 signifies a negative diagnosis (Health Control or HC), and a value of 1 indicates a positive diagnosis (referred to as Heart Disease or HD).

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Column Name** | **Description** | **Dtype** |
| 0 | Age | Age of the patient in years | int64 |
| 1 | Sex | Sex of the patient, where 1 represents male and 0 female after encoding | float64 |
| 2 | ChestPainType | Type of chest pain experienced by the patient encoded as values | float64 |
| 3 | RestingBP | Resting blood pressure in mm Hg | int64 |
| 4 | Cholesterol | Serum cholesterol in mg | int64 |
| 5 | FastingBS | Fasting blood sugar > 120 mg | int64 |
| 6 | RestingECG | Resting electrocardiographic results, values encoded | float64 |
| 7 | MaxHR | Maximum heart rate achieved during thallium test | int64 |
| 8 | ExerciseAngina | Exercise-induced angina, 1 if yes, 0 if no | float64 |
| 9 | Oldpeak | ST depression induced by exercise relative to rest | float64 |
| 10 | ST\_Slope | The slope of the peak exercise ST segment | float64 |
| 11 | HeartDisease | Presence of heart disease, 1 for positive, 0 for negative | int64 |

# Research hypothesis

**Hypothesis #1**: Heart disease affect each gender group differently.

* Q1.1: What is the proportion of each gender group that has HD?
* Q1.2: HD positive of each gender group experience different type of chest pain?

**Hypothesis #2**: Certain small groups of vitals would be significantly indicative of HD.

* Q2.1: What is the correlation of each vital with chance of HD.

**Hypothesis #3**: Logistical Regression (LR) model would perform better at HD prediction than GaussianNB (NB) due to its assumption (conditional independence of input features).

* Q3.1: What are the hyperparameters significantly affect the LR using SGDClassifier() for this dataset
* Q3.2: What is the performance gap between 2 models in terms of Accuracy, Precision, Recall, Selectivity, F1 and log loss

# Methodology

**Data Processing**

* Transform data set into pd.DataFrame
* One-hot encode all string-type features using `OneHotEncoder()`:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Column** | **Non-Null Count** | **Dtype** |
| 0 | Age | 918 non-null | int64 |
| 1 | Sex | 918 non-null | float64 |
| 2 | ChestPainType | 918 non-null | float64 |
| 3 | RestingBP | 918 non-null | int64 |
| 4 | Cholesterol | 918 non-null | int64 |
| 5 | FastingBS | 918 non-null | int64 |
| 6 | RestingECG | 918 non-null | float64 |
| 7 | MaxHR | 918 non-null | int64 |
| 8 | ExerciseAngina | 918 non-null | float64 |
| 9 | Oldpeak | 918 non-null | float64 |
| 10 | ST\_Slope | 918 non-null | float64 |
| 11 | HeartDisease | 918 non-null | int64 |

* Split into an 80/20 ratio for training and testing, respectively.

**Tunning model:**

**A computer code on a black background

Description automatically generated**

**Parameter Vector:**

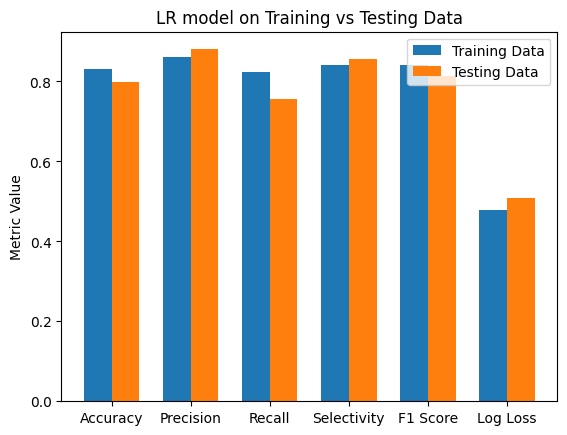
A computer screen with white text

Description automatically generated

W = [ **0.06** 0.13 -0.16 0.27 0.06 -0.14 0.13 -0.01 -0.2 -0.26 0.22 0.05]

# Result

**Train vs Test data**:

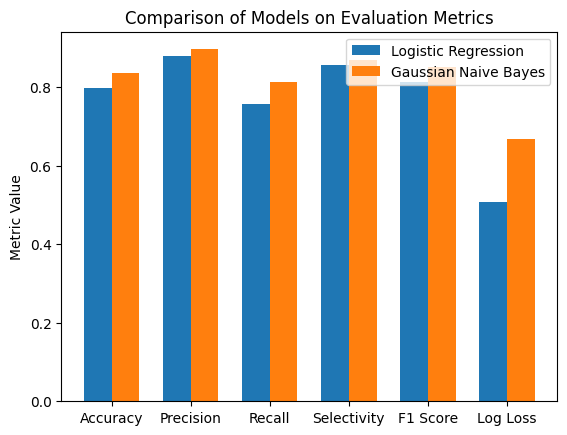


A chart of data testing

Description automatically generated with medium confidence

The Logistic Regression model displays consistent accuracy, precision, recall, selectivity, and F1 score between training and testing data. However, the model's higher log loss on testing data points to overconfidence in its probability predictions. Confusion matrix analysis reveals a balanced true positive and negative prediction rate across both datasets, though with slightly more false predictions during training.

**Logistic Regression vs Naïve Bayes Gaussian:**

****

**A comparison of a number of colored squares

Description automatically generated with medium confidence**

Logistic regression slightly outperformed by GaussianNB in most of the metrics, although it has it has lower log loss. The potential explanation would be the underlying assumptions of each model. Logistic regression might be better suited for this dataset as it does not assume features to be conditionally independent within one class as GaussianNB does.