**Project Report**

**1.Introduction**

Dataset: The dataset contains information about individuals and their heart health parameters. It includes 303 rows, each representing a different individual, and 17 columns covering various features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, and the target variable indicating the presence of heart disease.

The objective is predicting the presence of heart disease based on the provided health parameters. This is a classification problem, where the target variable is binary (0 for no heart disease, 1 for heart disease).

**2.Method**

Variables information

1. age -Age in year
2. sex -Sex 1, 0
3. chest pain type -chest pain type 1,2,3,4
4. resting blood pressure- resting bp s in mm Hg
5. serum cholesterol -cholesterol in mg/dl
6. fasting blood sugar -fasting blood sugar 1,0 > 120 mg/dl
7. resting electrocardiogram results- resting ecg 0,1,2
8. maximum heart rate achieved- max heart rate 71–202
9. exercise induced angina exercise angina 0,1
10. old peak =ST -old peak depression
11. the slope of the peak exercise ST segment -ST slope 0,1,2 Nominal
12. class target 0,1

<https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final>

*problem definition*

1.The objective of the analysis is to predict the likelihood of a heart condition based on various health indicators. The goal is to build a logistic regression model that can accurately predict the likelihood of a heart condition given these health indicators.

This problem is relevant in the healthcare sector, as it can help medical professionals and researchers better understand the factors associated with heart disease and develop more effective prevention and treatment strategies.

*Model used*

Logistic regression -Logistic regression is a statistical model used for binary classification tasks, making it suitable for predicting the presence or absence of heart disease based on the given health parameters.

*Steps in coding*

1.importing dataset, checking of missing values, data description.

2.Assessing multicollinearity and dropping of features with higher Variance inflation factor.

3.Tuning of Hyper parameters

4. Confusion matrix and interpretation of results.

**3.Evaluation of model**

The best score achieved by the logistic regression model on the test set is approximately 0.7888509231193165, which represents an accuracy of around 78.9% in predicting the presence of heart disease.

**Cross-Validation Scores:** The cross-validation scores, which provide a more robust estimate of the model's performance, are: [0.80672269, 0.81092437, 0.78991597, 0.72689076, 0.73109244] The mean cross-validation score is approximately 0.773109243697479, which is slightly lower than the test set score.

**Model Performance:** The combination of the test set score and the cross-validation scores suggests that the logistic regression model with the specified hyper parameters (C=0.1, max\_iter=100, penalty='l2') performs reasonably well in predicting the presence of heart disease based on the data.

**Generalization Ability:** The relatively high cross-validation scores, with a mean of around 0.773, indicate that the model has a good generalization ability and is likely to perform well on unseen data, not just the test set.

*Confusion matrix*

**Accuracy (0.80):** The accuracy of 80% indicates that the model correctly predicted 80% of the instances in the data.

**Precision (0.84):** The precision of 84% suggests that when the model predicts an individual has heart disease, it is correct 84% of the time. It is a measure of the model's ability to avoid false positives.

**Recall (0.79):** The recall of 79% indicates that the model correctly identifies 79% of the individuals who actually have heart disease. It is a measure of the model's ability to capture all positive instances.

**F1-score (0.81):** The F1-score of 81% is the harmonic mean of precision and recall. It provides a balance between precision and recall, giving a single metric to evaluate the model's performance.

*Conclusion*

The logistic regression model with the specified hyperparameters (C=0.1, max\_iter=100, penalty='l2') performs reasonably well in predicting the presence of heart disease based on the data.

The model demonstrates a good balance between accuracy, precision, recall, and generalization, making it a reliable tool for predicting heart disease.

**References**

Scikit-learn Documentation: <https://scikit-learn.org/stable/documentation.html>

StatQuest YouTube Channel: <https://www.youtube.com/user/joshstarmer>

Towards Data Science on Medium: <https://towardsdatascience.com/>

Kaggle: <https://www.kaggle.com/>