**Prediction for Getting Heart Disease Using Logistic Regression Model**

**Introduction**

Heart diseases are increasing all over the world specially for the past few years. We found out that heart diseases are number one disease that can cause of death. There are many factors that can affect the heart and may cause these kinds of diseases such as age, sex, blood pressure, Cholesterol level, the level of sugar in the blood.

Since heart disease can be hard to detect in early stage, we need to find a way to predict this hard disease earlier. I found out that machine learning as usual can help us in many things so it will help us to detect if this person will get heart disease or not.

**The problem**

Using Logistic Regression Model to predict heart disease that can assist medical professionals in predicting heart disease status based on the clinical data of patients.

**Dataset features**

The dataset I will be using hosted on Kaggle and it was from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Heart+Disease).

|  |  |
| --- | --- |
| **age** | Patient age in years |
| **sex** | Patient sex |
| **cp** | Chest pain type |
| **trestbps** | Resting blood pressure in millimeters of mercury (mm Hg) |
| **chol** | Cholesterol level in mg/dl |
| **fbs** | Whether the level of sugar in the blood is higher than 120 mg/dl or not |
| **restecg** | Results of the electrocardiogram on rest |
| **thalach** | Maxium heart rate during the stress test |
| **exang** | Whether the patient had angina during exercise |
| **oldpeak** | Decrease of the ST segment during exercise according to the same one on rest |
| **slope** | Slope of the ST segment during the most demanding part of the exercise |
| **ca** | Number of main blood vessels coloured by the radioactive dye |
| **Thal** | Results of the blood flow observed via the radioactive dye. |
| **target** | Target variable: whether the patient has a heart disease or not |

Table 1

**Work and Visualization**

First of all, I explored the data set, and it’s features then visualize the data by using Plot graph to understand and explain each feature and their values. After that I checked if the dataset has any non-value. Heatmap was used to visualize if there is any non-value or not. After getting zero non-values I trained and tested the data by importing train and test split and using 0.2 of the data set to test using the logistic regression model. After that I import the **Confusion Matrix** to represent false positive, false negative, True positive and True Negative then printing the Accuracy Score, Precision Score, Recall Score, F1 Score and Confusion Matrix. After printing the accuracy **GridSearchCV** was imported to improve the accuracy to get the final and the best result of the accuracy score.

**Results**

The result was a clear EDA of the dataset and after the training and testing using Logistic Regression Model, we found out about:

* **Accuracy Score :** 0.8360655737704918
* **Precision Score :** 0.8709677419354839
* **Recall Score :** 0.8181818181818182
* **F1 Score :** 0.84375
* **Confusion Matrix :** [[24 4][ 6 27]]

At the end **GridSearchCV** was imported to get better result and to improve the accuracy score

**Conclusion**

This project will provide better diagnosis the heart patients. I used here Logistic Regression Model to predict heart disease that can assist medical professionals in predicting heart disease status based on the clinical data of patients. First, we select 14 important clinical features from the dataset then using the Model to predict the disease depending on the target feature. Finally we got the accuracy score and the Confusion Matrix