Heart Disease Diagnostic Analysis

- Project Report

by Sanil Jain

GitHub Repo

CONTENTS

INTRODUCTION

PROBLEM STATEMENT

You have just been hired as a Data Scientist at a Hospital with an alarming number of patients coming in reporting various cardiac symptoms. A cardiologist measures vitals & hands you this data to perform Data Analysis and predict whether certain patients have Heart Disease. We would like to make a Machine Learning algorithm where we can train our AI to learn & improve from experience. Thus, we would want to classify patients as either positive or negative for Heart Disease.

ABOUT DATASET

Our dataset contains 14 columns & 1025 rows.

The dataset has 3 types of data :

1. **Continuous : which is quantitative data that can be measured.**
2. **Ordinal Data: categorical data that has a order to it**

   (0,1,2,3, etc).

3. **Binary Data: data whose unit can take on only two possible states (0 &1).**

The last column, named "target," is our dependent variable. This variable is what we aim to predict or explain through our data analysis and modeling efforts. The remaining 13 columns are our independent variables. These variables are the predictors or features that provide the input information to our models.

```
data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

PROJECT METHODOLOGY

TECHNOLOGIES USED

- **Programming Language -> Python**
- **Data Manipulation -> NumPy & Pandas**
- **Data Visualization -> Matplotlib & Seaborn**
- **Machine Learning -> Scikit-Learn**
- **API -> Flask & HTML**
- **Version Control System -> Git**
- **Repository Hosting -> GitHub**

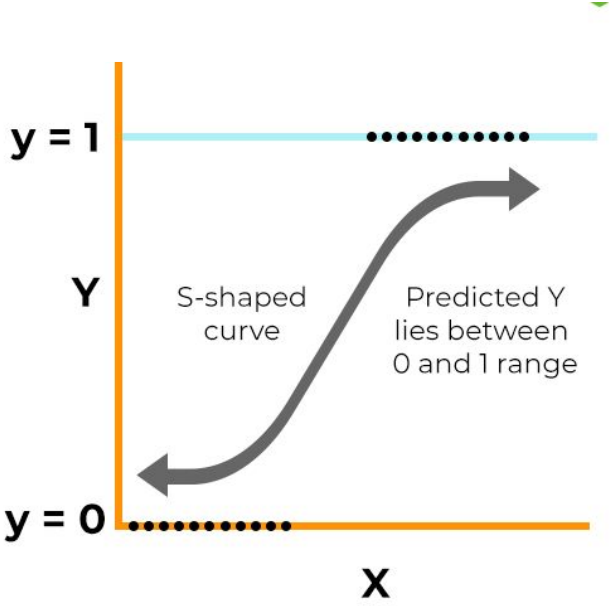
MACHINE LEARNING ALGORITHM

In my data science project, I chose to use logistic regression due to several compelling reasons. Firstly, the nature of the problem was inherently a classification task, where the goal was to predict categorical outcomes based on a set of input features. Logistic regression is particularly well-suited for binary classification problems, making it an ideal choice for this project. Additionally, the data I worked with was labeled, meaning that each instance in the dataset had a known class label. This supervised learning scenario is where logistic regression excels, as it can effectively learn the relationship between the input features and the binary target variable. Moreover, logistic regression provides a clear probabilistic interpretation of the predictions, allowing us to understand not just the predicted class but also the confidence in those predictions. Its simplicity, ease of implementation, and robustness in handling linear relationships between the dependent and independent variables further reinforced the decision to use logistic regression in this project.

The *standard* logistic function $\sigma : \mathbb{R} \to (0, 1)$ is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Graph Of The Function



FINAL OUTPUT

We have this data with output. Let's check if our machine learning model can predict the output correctly by using these same values.

Age : 41

Sex : Female

Chest Pain Type : Atypical Angina

Resting Blood Pressure:130

Serum Cholsetoral (mg/dL) : 204

Fasting Blood Sugar : No

Resting ECG : Normal

Maximum Heart Rate Achieved : 172

Exercise Induced Angina : No

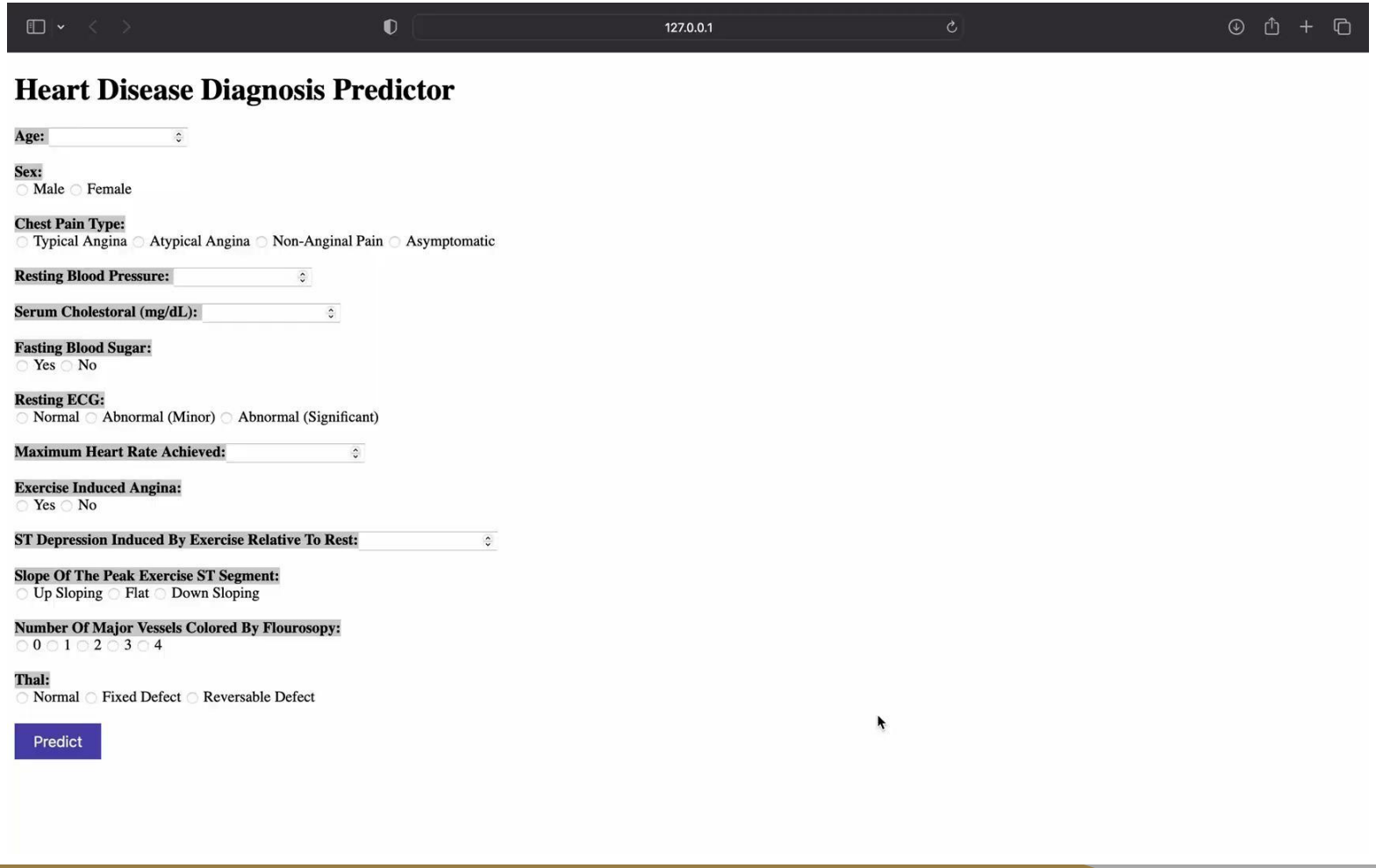
ST Depression Induced By Exercise Relative To Rest : 2

Slope Of The Peak Exercise ST Segment : Down Sloping

Number Of Major Vessels Colored By Flourosopy : 0

Thal : Reversable Defect

Prediction : Patient Diagnosed


127.0.0.1
Heart Disease Diagnosis Predictor
Age:
Sex:
Male
Female
Chest Pain Type:
Typical Angina
Atypical Angina
Non-Anginal Pain
Asymptomatic
Resting Blood Pressure:
Serum Cholestoral (mg/dL):
Fasting Blood Sugar:
Yes
No
Resting ECG:
Normal
Abnormal (Minor)
Abnormal (Significant)
Maximum Heart Rate Achieved:
Exercise Induced Angina:
Yes
No
ST Depression Induced By Exercise Relative To Rest:
Slope Of The Peak Exercise ST Segment:
Up Sloping
Flat
Down Sloping
Number Of Major Vessels Colored By Flourosopy:
0
1
2
3
4
Thal:
Normal
Fixed Defect
Reversable Defect
Predict