
CBD – GROUP PROJECT

Heart Disease Prediction Machine Learning Model

MARCH 17, 2021

Aarushi Pandey, Karthikeyan Mohan, Robin Chhabra, Yugandhar Kumar Savalam

Contents

A. Project Background:.....	3
B. Problem Statement:.....	3
C. Tools used:	3
D. Model Design Approach:.....	4
E. Dataset explanation:.....	5
F. Data Visualization:	6
G. Data Cleaning:	8
H. Feature Engineering:.....	8
I. Model Building and evaluation:	9
J. Model Deployment:	9
K. Future Enhancements:.....	10
L. Conclusion:.....	10

A. Project Background:

The main aim of this project is to help patients to detect heart disease based on some attributes. The need for this project can be understood when we look at the report issued by WHO, according to which 17.9 million lives are lost each year. It would be unchallenging for patients to decipher if they are suffering from cardiovascular disease with our project. Early detection of such illness could benefit patients to start their treatment and get cured in time.

B. Problem Statement:

In this project, we have been provided with the Heart disease data of the Cleveland patients. We need to use this data to build a machine learning model that predicts whether the patients are diagnosed with heart disease. The key prerequisites in this project are a dataset containing heart disease-related attributes, python libraries, various machine learning algorithms, visualization tools, and deployment in the cloud.

C. Tools used:

1. **Python:** To do coding



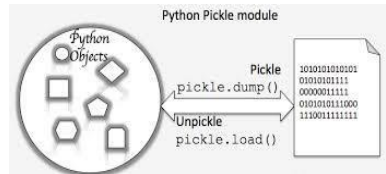
2. **Python Libraries:** To perform cleanup and modeling.



3. **Data Visualization tool:** To Visualize the data and find the relations between them.



4. **Deployment:** To create a web application and host it in the cloud.



D. Model Design Approach :

Before starting working on the project, we brainstormed and listed all the steps required to get the required end product. We followed the below design approach and assigned it to each other.

- Understood each attribute in the dataset and found their datatypes and the values present in each. Identified the target variables and the input attributes.
- Used the Power BI tool to visualize the data and found the pattern and relations between each attribute. Made an interactive dashboard with the dataset.
- Based on the power bi analysis, performed the data cleansing and made the dataset ready for the model.
- Assigned the input and output into the separate variables for the model input.
- Split the dataset for test and train.
- Identified the algorithms, build the model, and evaluated the model.
- Provided the user's sample input to the model and checked whether it predicts the correct value as expected.
- Built the pickle file out of the final python code and used it in the web application to call our model to predict the output. Pickle file is nothing but the dump of the entire code, and the model we created will be called while creating a web page.

- Created an interactive web page using HTML with all the 13 inputs required as the model's input.
- Created a web application using Flask in python to assign the 13 inputs received from the HTML page with the pickle file's model function. This gave us the complete web application.
- Provided the user's inputs in the web application and ran it locally to check whether the model predicts the expected output.
- Hosted it in the AWS EC2 server and ensured that it is running from different machines.

E. Dataset explanation:

The dataset used is a sample of patients in a particular age group to screen for the [heart disease dataset](#). This dataset consists of 303 records and 76 characteristics, but we have used below 14 of them.

- **Age:** This feature defines the patient's age in years.
- **Sex:** This indicates the patient's gender; if the patient is male, the output is 1; if the patient is female, the output is 0 (1=male, 0=female).
- **Chest pain type (cp):** Defines the type of chest pain patient is suffering. Categorized in four parts:- Typical angina = 1, Atypical angina = 2, non-anginal pain = 3, asymptomatic = 4.
- **Resting Blood Pressure(trestbps):** It refers to the patient's normal blood pressure, which should be less than 120 over 80 (120/80) to avoid the risk of danger, measured in mm Hg.
- **Serum cholesterol in mg/dl(chol):** A serum cholesterol level depicts the amount of high-density lipoprotein cholesterol (HDL) and low-density lipoprotein cholesterol (LDL) in a person's blood. The average serum cholesterol level should be less than 200 mg/dl.

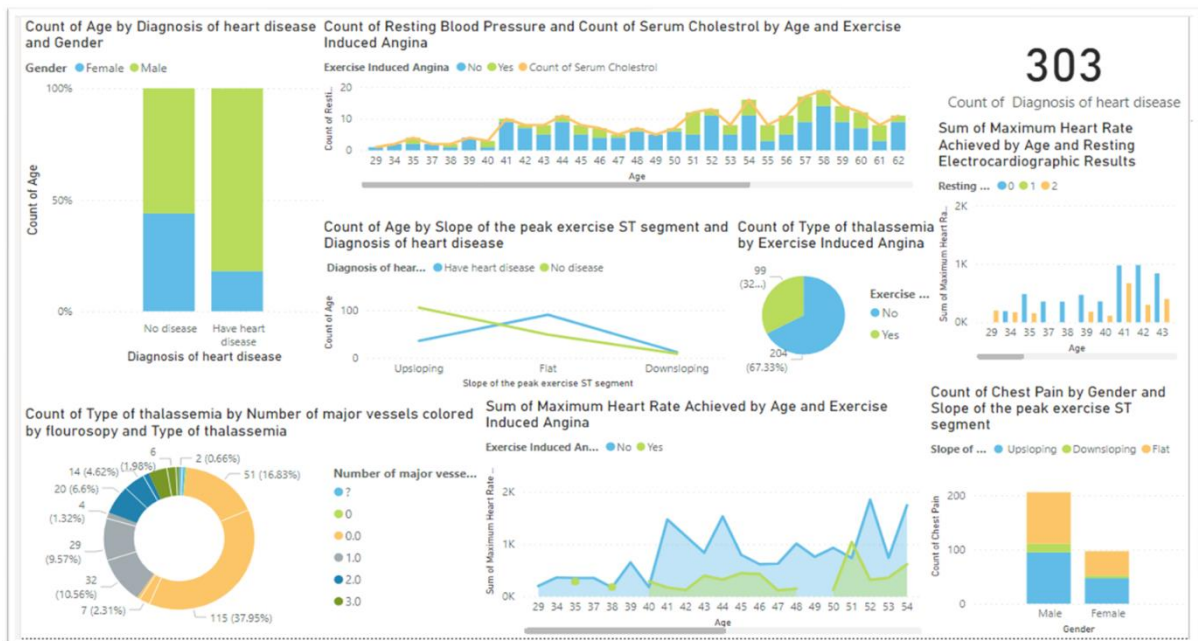
- **Fasting blood sugar(fbs):** This indicates whether the blood sugar level after overnight fasting of a patient is greater than 120 mg/dl or not. The values are:- yes =1 and no =0.
- **Maximum Heart Rate Achieved(thalach):** This characteristic explains the maximum heartbeat recorded of the patient.
- **Exercise-Induced Angina(exang):** Angina is a form of chest pain caused by reducing heart blood flow. It is used to indicate whether the patient feels angina during physical activity. Values induced are: (yes = 1, no = 0).
- **ST depression value(old peak):** ST Depression occurs when the J point is placed below the baseline, and the value measured at which the J curve occurs is the ST depression value.
- **Slope of peak in ST segment(slope):** This refers the type of slope occurred on ECG : Value 1 = UPSLOPING, Value 2 = HORIZONTAL, Value 3 = DOWNSLOPING .
- **Major Vessel colored by fluoroscopy(ca):** Fluoroscopy is the procedure of examining the disease and injuries by generating images of body parts. This feature specifies the number of vessels colored in the fluoroscopy range (0-3).
- **Type of Thalassemia(thal):** Thalassemia is a genetic blood condition in which the body produces an irregular type or insufficient hemoglobin. It is categorized in 3 forms: 3 = Normal, 6 = fixed defect, 7= reversible defect.
- **Diagnosis of Heart Disease(num):** This attribute defines the targeted outcome which we are predicting from this machine learning model, presence of heart disease (value 1,2,3,4) or absence of heart disease (value 0).

F. Data Visualization :

We used Power BI as the visualization tool to visualize the data efficiently to understand each attribute and gained below insights :

- The first visualization is the Count of age vs. Diagnosis of heart disease stacked column chart which depicts as compared to males, females are at less risk of heart disease.

- Another visualization line and stacked column chart represent the relation between the Count of resting blood pressure and the Count of serum cholesterol by Age and Exercise-induced angina. With an increase in age the Count of serum cholesterol increases and exercise induce angina is not showing any particular trend with an increase in age.



Power BI Dashboard created by us

- The pie chart of the Count of type of thalassemia vs Exercise-induced angina shows that people with complaints of exercise-induced angina show less Count of a type of thalassemia in comparison to people having no complaints of exercise-induced angina.
- Line chart between Count of age by Slope of peak exercise ST-segment and Diagnosis of heart disease depicts slope of peak exercise segment having heart disease shows an upward and then a sharp downward trend while the slope of peak exercise segment with no heart disease shows a continuous downward trend.
- With an increase in age, the sum of maximum heart rate achieved increases, and complaints of exercise-induced angina as well. This generalization can be drawn through a stacked area chart of Sum of maximum heart rate achieved by Age and Exercise-induced angina.

- Visualization of Count of chest pain by Gender and Slope of the peak exercise ST segment is shown through a 100% stacked column chart that concludes that chest pain complaints are more in males than females. Also, the upsloping and flat trend of peak exercise ST segment are more common than downsloping in both males and females.
- The donut chart represents the Count of type of thalassemia by the number of major vessels colored by fluoroscopy and type of thalassemia. The majority of patients with 0 major vessels colored by fluoroscopy have normal thalassemia, almost equal number of patients with 1 major vessel colored by fluoroscopy suffers from either normal or reversible defect.
- The Sum of maximum heart rate achieved by the number of major vessels colored by fluoroscopy and Resting Electrocardiographic results is visualized through clustered column chart. Patients with 0 value of resting electrocardiographic can be observed at age of 52. An almost negligible number of patients are there with value 1 of resting electrocardiographic.

G. Data Cleaning :

- Data cleaning is the process of replacing or deleting inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset.
- we have not found any null values in any dataset but we have got a special character (?) in the 'ca' and 'thal' fields. Since it's a categorical value, we have used the impute method to fill '?' with the Mode value.

H. Feature Engineering :

Based on our analysis from visualisation, we identified that all the attributes are related are required for modelling. Hence we are going with all the attributes. We have updated the target dataset from (0,1,2,3,4) to just two categorical values(0,1) to fit into the model. We have split the dataset into test and training with the 80:20 ratio.

I. Model Building and evaluation:

Since it's a discrete output attribute, we have chosen a classification algorithm. We modeled with following three classification algorithms and choose the best one based on accuracy:

When dealing with classification problems, several metrics can be used to gain insights into how the model performs.

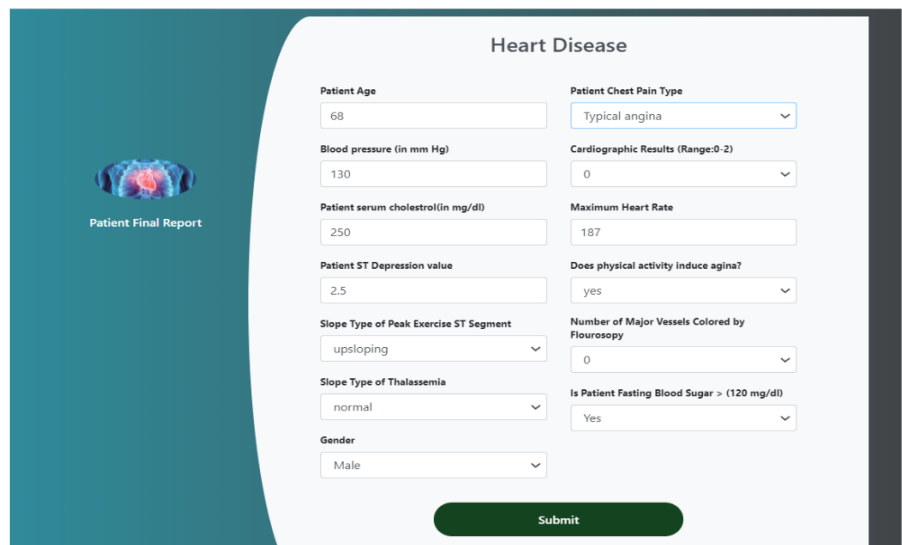
- **Accuracy:** The metric measures the ratio of correct predictions over the total number of instances evaluated.
- **Precision:** It is used to measure the positive pattern that is correctly predicted from the total predicted patterns in a positive class.

Algorithm	Accuracy
Logistic Regression	0.80328
Naive Bias	0.78689
Support Vector Machine	0.57377

We preferred the Logistic regression algorithm to do Patient Heart Disease data predictions because it has the highest test dataset accuracy.

J. Model Deployment :

We created a below web application out of our model :



The screenshot shows a web application interface for heart disease prediction. On the left, there is a teal sidebar with a heart icon and the text "Patient Final Report". The main area is white and titled "Heart Disease". It contains a form with various input fields and dropdown menus for patient data. The fields are arranged in two columns. At the bottom right, there is a green "Submit" button.

Field	Value
Patient Age	68
Patient Chest Pain Type	Typical angina
Blood pressure (in mm Hg)	130
Cardiographic Results (Range:0-2)	0
Patient serum cholesterol(in mg/dl)	250
Maximum Heart Rate	187
Patient ST Depression value	2.5
Does physical activity induce agina?	yes
Slope Type of Peak Exercise ST Segment	upsloping
Number of Major Vessels Colored by Flourosopy	0
Slope Type of Thalassemia	normal
Is Patient Fasting Blood Sugar > (120 mg/dl)	Yes
Gender	Male

We tested the model with sample random data and predicted the output. we created the web application using Flash.

We had created an instance in EC server and deployed our code in it. Then hosted our website in that server. We will start the start the server during the demo and present it.

<http://ec2-18-223-21-172.us-east-2.compute.amazonaws.com:8080/>

K. Future Enhancements:

- More accuracy in the prediction.
- Input Validations in the patient web form fields to avoid wrong values.
- Unique Identity Number for the Patient to find the reports quickly.

L. Conclusion:

This project helps us to involve more practical work as a team. We tried out of the box and implemented new tools and techniques to do innovative work. At this point, we have trained one complete model that will classify patient heart disease. It motivates us to learn and read new concepts and work towards building more machine learning projects.

M. Reference :

- <https://www.twilio.com/blog/deploy-flask-python-app-aws>
- <https://www.linkedin.com/learning/power-bi-essential-training-3/create-rich-interactive-reports-with-power-bi?u=56968457>
- <https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af>
- <https://www.slideshare.net/CharlesVestur/building-a-performing-machine-learning-model-from-a-to-z>