

# **CARDIO CARE**

## **(HEART DISEASE RISK PREDICTOR)**

Amar Parameswaran

Date: 21<sup>st</sup> June, 2022

### *Abstract:*

The health care institutes collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Risk Predictor (HDRP) is developed using machine learning for predicting the risk level of heart disease. The system uses 11 medical parameters such as age, sex, blood pressure, cholesterol, and blood sugar for prediction. The HDRP predicts the likelihood of patients getting heart disease. It enables significant knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established.

## 1.0 Introduction

Among various life-threatening diseases, heart disease has garnered a great deal of attention in medical research. The diagnosis of heart disease is a challenging task, which can offer automated prediction about the heart condition of patient so that further treatment can be made effective. The diagnosis of heart disease is usually based on signs, symptoms, and physical examination of the patient. There are several factors that increase the risk of heart disease, such as blood sugar, Oldpeak, body cholesterol level, family history of heart disease, high blood pressure, and lack of physical exercise.

A major challenge faced by health care organizations, such as hospitals and medical centers, is the provision of quality services at affordable costs.<sup>1</sup> The quality service implies diagnosing patients properly and administering effective treatments. The available heart disease database consists of both numerical and categorical data. Before further processing, cleaning, and filtering are applied on these records in order to filter the irrelevant data from the database. The proposed system can determine an exact hidden knowledge, i.e., patterns and relationships associated with heart disease from a historical heart disease database. It can also answer the complex queries for diagnosing heart disease; therefore, it can be helpful to health care practitioners to make intelligent clinical decisions. Results showed that the proposed system has its unique potency in realizing the objectives of the defined mining goals.

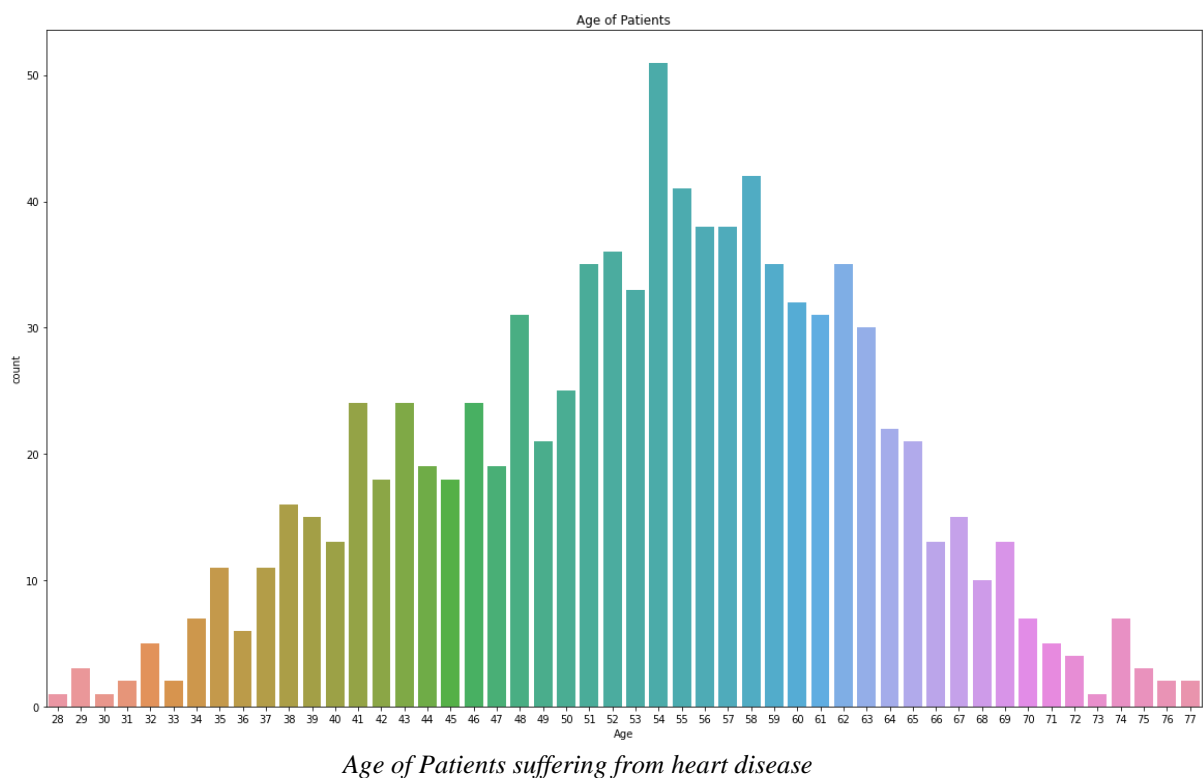


## 2.0 Customer Needs Assessment

According to WHO data, heart disease is the leading cause of mortality globally, resulting in 17.9 million deaths annually. A heart attack occurs when the heart's blood circulation is obstructed by arteries plaque build-up. A thrombus in an artery causes a stroke by impeding blood flow to the brain. The symptoms are common to other illnesses and might be confused with indicators of ageing, making diagnosis difficult for practitioners. Precision prediction and timely identification of cardiac disease are essential for improving patient survival rate. Because of the increased collection of medical data, practitioners now have a great opportunity to promote healthcare diagnosis.

Imagine clinics and other small to medium scale diagnostic centers being capable of giving early and accurate decisions regarding the heart condition of a particular patient. Timely and proper suggestions could be given which may prove to be crucial in certain circumstances.

In this article we will explore how AI or rather Machine Learning could be used on medical data collected by various medical institutes to predict whether an individual is likely to suffer from a heart disease.



### 3.0 Target Specifications

Targets to be achieved by CardioCare (a web-application) are -

- To improve the traditional diagnostic process by predicting the heart condition of a patient and providing accurate remedies.
- Reducing frustration and death of patients due to misinformation and delay in such diagnosis.
- To make heart diagnosis more feasible and available to all sections of the society.

*Consider a scenario where a patient might potentially suffer from a heart disease. He/she goes to a clinic for their usual checkup. Tests are conducted based on their blood sugar levels and heart rate. Without this software available to the diagnostic centers certain standard precautionary measures might be suggested to the patient whose accuracy depends entirely upon the qualifications and judgement of a doctor. In most cases patients consider themselves to be the judge of their own medical condition.*

*By the inclusion of this software, an accurate and informed suggestion could be made if the heart of that patient is at risk. Furthermore, if this software is recognized by doctors and practitioners, it could be made a standard issue and would overall improve the lifestyle of the patients.*

Some key characteristics to extracted from the above use case are –

- An intuitive simple software such as a web application
- To be available to all small to medium scale diagnostic centers
- Take traditional inputs like blood sugar and heart rate of a person
- Simple questions regarding exercise and occasional chest pains.
- Predict the risk of getting a heart disease and portray parameters that are posing to be an issue.

### 4.0 External Searches

#### 4.1 Prediction of hospitalization due to heart diseases by supervised learning methods

Background:

In 2008, the United States spent \$2.2 trillion for healthcare, which was 15.5% of its GDP. 31% of this expenditure is attributed to hospital care. Evidently, even modest reductions in hospital care costs matter. A 2009 study showed that nearly \$30.8 billion in hospital care cost during 2006 was potentially preventable, with heart diseases being responsible for about 31% of that amount.

#### Methods:

Our goal is to accurately and efficiently predict heart-related hospitalizations based on the available patient-specific medical history. To the best of our knowledge, the approaches we introduce are novel for this problem. The prediction of hospitalization is formulated as a supervised classification problem. We use de-identified *Electronic Health Record (EHR)* data from a large urban hospital in Boston to identify patients with heart diseases. Patients are labeled and randomly partitioned into a training and a test set. We apply five machine learning algorithms, namely Support Vector Machines (SVM), AdaBoost using trees as the weak learner, Logistic Regression, a naïve Bayes event classifier, and a variation of a Likelihood Ratio Test adapted to the specific problem. Each model is trained on the training set and then tested on the test set.

### 4.2 A novel approach for heart disease prediction using strength scores with significant predictors

#### Background:

Cardiovascular disease is the leading cause of death in many countries. Physicians often diagnose cardiovascular disease based on current clinical tests and previous experience of diagnosing patients with similar symptoms. Patients who suffer from heart disease require quick diagnosis, early treatment and constant observations. To address their needs, many data mining approaches have been used in the past in diagnosing and predicting heart diseases. Previous research was also focused on identifying the significant contributing features to heart disease prediction; however, less importance was given to identifying the strength of these features.

#### Method:

This paper is motivated by the gap in the literature, thus proposes an algorithm that measures the strength of the significant features that contribute to heart disease prediction. The study is aimed at predicting heart disease based on the scores of significant features using Weighted Associative Rule Mining.

### 4.3 Heart Failure Prediction Dataset - Kaggle

#### Context

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5 CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

#### Attribute Information

- 1) Age: age of the patient [years]
- 2) Sex: sex of the patient [M: Male, F: Female]
- 3) ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- 4) RestingBP: resting blood pressure [mm Hg]
- 5) Cholesterol: serum cholesterol [mm/dl]
- 6) FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- 7) RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- 8) MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- 9) ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- 10) Oldpeak: oldpeak = ST [Numeric value measured in depression]
- 11) ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
- 12) HeartDisease: output class [1: heart disease, 0: Normal]

#### Source

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations  
Duplicated: 272 observations  
Final dataset: 918 observations

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

## 5.0 Benchmarking alternate products

Ways to diagnose one's heart:

- **Electrocardiogram**
  - An electrocardiogram (also called EKG or ECG) is a test that records the electrical activity of your heart through small electrode patches attached to the skin of your chest, arms, and legs.
- **Chest X-Ray**
  - Like the term suggests, it uses a small amount of radiation to produce an image of your heart, lungs, and blood vessels.
- **CT Heart Scan**
  - Computed tomography (CT scan) of the heart can visualize your heart's anatomy. Calcium-score heart scan and coronary CT angiography are just a few types used to diagnose heart disease.
- **Heart MRI**
  - One test that looks for heart disease is called an MRI (Magnetic Resonance Imaging). It uses large magnets and radio waves to make pictures of your body's internal organs. You're not exposed to X-rays. This can also make images of your heart's pumping cycle.
- **Stress Test**
  - A stress test can estimate your risk of having a heart disease. A doctor or trained technician performs the test. They'll learn how much your heart can manage before an abnormal rhythm starts or blood flow to your heart muscle drops.

- **Electrophysiology Test**

- An electrophysiology (EP) study is a test that records the electrical activity and the electrical pathways of your heart.
- It can help find what's causing your irregular heartbeat. It also helps figure out the best treatment for you.

- **Myocardial Biopsy**

- A heart biopsy, also called myocardial biopsy or cardiac biopsy, is an invasive procedure to detect heart disease. It entails using a biptome (a small catheter with a grasping device on the end) to obtain a small piece of heart muscle tissue that is sent to a laboratory for analysis.

❖ *All these methods can only see the current condition of the heart. These are great methods that could be utilized after an informed prediction by the software. This would save time as well as ensure the evasion of future irregularities that may go unnoticed through traditional methods.*

## **6.0 Applicable Patents & Regulations**

- Patent on the web application developed
- Patent on the modified modules and algorithms
- Laws related to privacy for collecting data from users
- Protection as well as ownership regulations
- Creating an e-mail service to mail the report to the patient and doctor
- Being responsible for the design
- Review of existing work authority regulations

## **7.0 Applicable Constraints (need for space, budget, expertise)**

Expertise:

- Requires a lot of research to obtain respective regional dataset of heart patients in-order to provide more informed and accurate results.
- Establishing an e-mail service in the product which can send the report to the registered patient after the outcome is predicted by the deployed model.
- Confidential health data to be obtained to further train the model.
- Data Analysts and Machine Learning Engineers required.



Space:

- A reliable database integration in the service to collect and store the patient's data for future purposes.

Budget:

- Maintenance of the machine learning pipeline deployed in the service and the user-interface.
- Maintenance of the database and updating the model based on the data collected.
- Security protocols and regulations to safeguard the information and service.

## **8.0 Business Opportunity**

The correct prediction of heart diseases could prevent life threats. Also, its regarded as one of the most important topics in clinical data analysis. Accurate and consistent predictions of heart diseases in patients made available easily at every diagnostician's palette would indeed be very handy. Some irregularities when predicted could then be confirmed by traditional methodologies and taken care of at an early stage itself.

## **9.0 Concept Generation**

Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports. Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc.

Nowadays lot of research data and patients records of hospitals are available. There are many open sources for accessing the patient's records and research can be conducted so that various computer technologies could be used for doing the correct diagnosis of the patients and detect this disease to stop it from becoming fatal. Nowadays it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. We can use different machine learning and deep learning models to diagnose the disease and classify or predict the results.

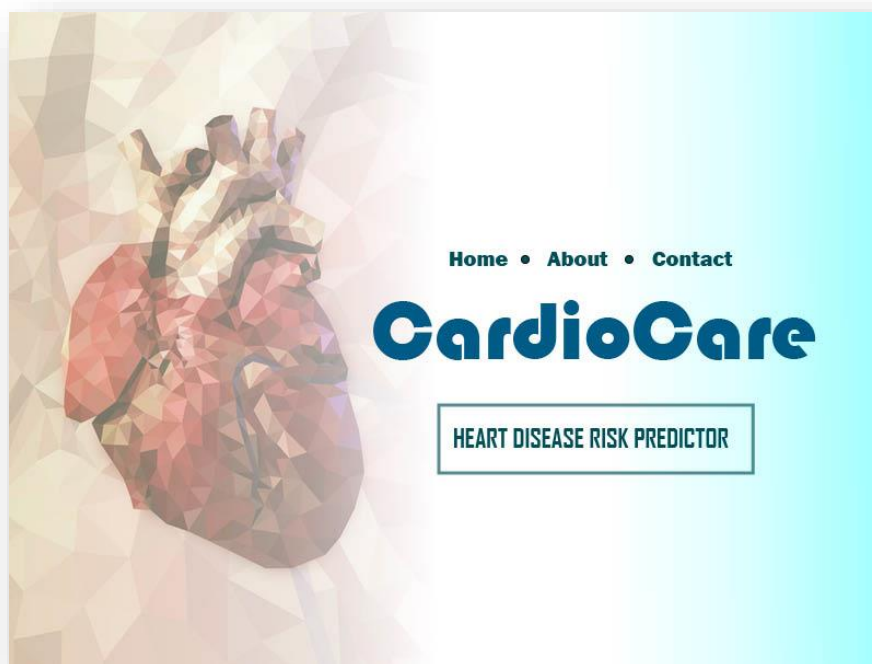
Many studies have been performed and various machine learning models are used for doing the classification and prediction for the diagnosis of heart disease.


*Thus, the idea of creating and deploying a feasible and reliable service for the populace based on the above data is generated.*

## 10.0 Concept Development

- Cardio Care web application would be available to all diagnosticians
- They will have to proceed to the Heart Disease Predictor Page
- Here, patient details will be entered according to the tests
- The submit button will then run the Classification Algorithm through the deployed Machine Learning Pipeline.
- This Classification Algorithm is decided upon, based on a predetermined dataset and previous test results stored in the database.
- The outcome is portrayed on a separate web page displaying the heart disease risk a patient might have and whether its heavy/mild/none.
- An email service would then send the report to the patients email id for reference

## 11.0 Final Product Prototype (abstract) with Schematic Diagram





**Age:**

**Sex:**

**Chest Pain Type:**

**Resting BP:**

**Cholestrol:**

**Fasting Blood Sugar:**

**Resting ECG:**


**Maximum HR:**

**Exercise Angina:**

**Oldpeak:**

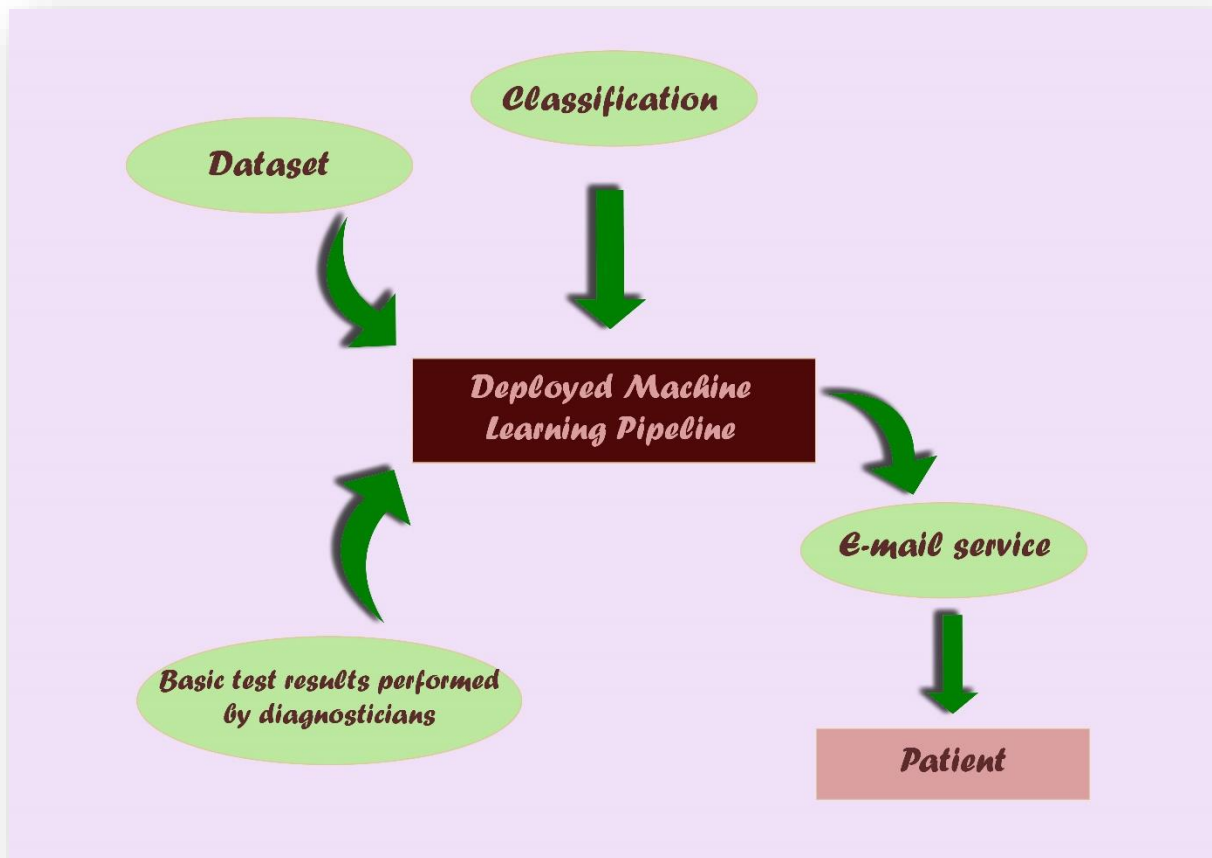
**ST Slope:**

**Submit**



**HEART DISEASE=**  %

**REMARKS: HIGH/MEDIUM/LOW/NO RISK**



*Schematic Diagram*

## 12.0 Product Details

How does it work?

- Deployed web application takes inputs from the user
- The inputs are then fed to a pre-trained machine learning model
- This then predicts the outcome with high accuracy
- The report is then sent to the patient through an email-service

What is the source of your data?

- The model is pre-trained based on a [Kaggle Dataset](#)
- This model will be further optimized on the results of patient data stored in the database

What are the different classification algorithms used?

- Logistic Regression
- Random Forest Classifier
- Support Vector Classifier
- K Nearest Neighbors Classifier
- Decision Tree Classifier
- Ada Boost Classifier
- Also applied hyper-parameter tuning to the top 3 performing models based on the Kaggle dataset

What are the frameworks used?

- The CardioCare web application would be deployed on Heroku Cloud service using Flask as a framework

Team required to develop and cost?

- A group of data scientists or machine learning engineers required to maintain the scalability and accuracy of the deployed machine learning pipeline
- A group of frontend developers to maintain the User-interface or improve it as and when necessary
- Maintaining a database and email-service would add to the cost. As it needs to be feasible the production of CardioCare shall be kept very economical. Hence automation of the machine learning pipeline is essential.

## 13.0 Code Implementation/ Validation on small scale

**Dataset Link** = <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>

**Github Link** = [https://github.com/Amar1709/Heart\\_Disease\\_Risk\\_Predictor.git](https://github.com/Amar1709/Heart_Disease_Risk_Predictor.git)

**Some Code Implementation Snapshots:**

Importing the Libraries

```
1 #Importing the necessary libraries
2
3 import pandas as pd
4 import numpy as np
5 import matplotlib.pyplot as plt
6 import seaborn as sns
7
8 %matplotlib inline
9
10 import warnings
11 warnings.filterwarnings('ignore')
✓ 3.9s
```

## Checking for Null values

```
1 df.isnull().sum() #No missing values
✓ 0.6s
```

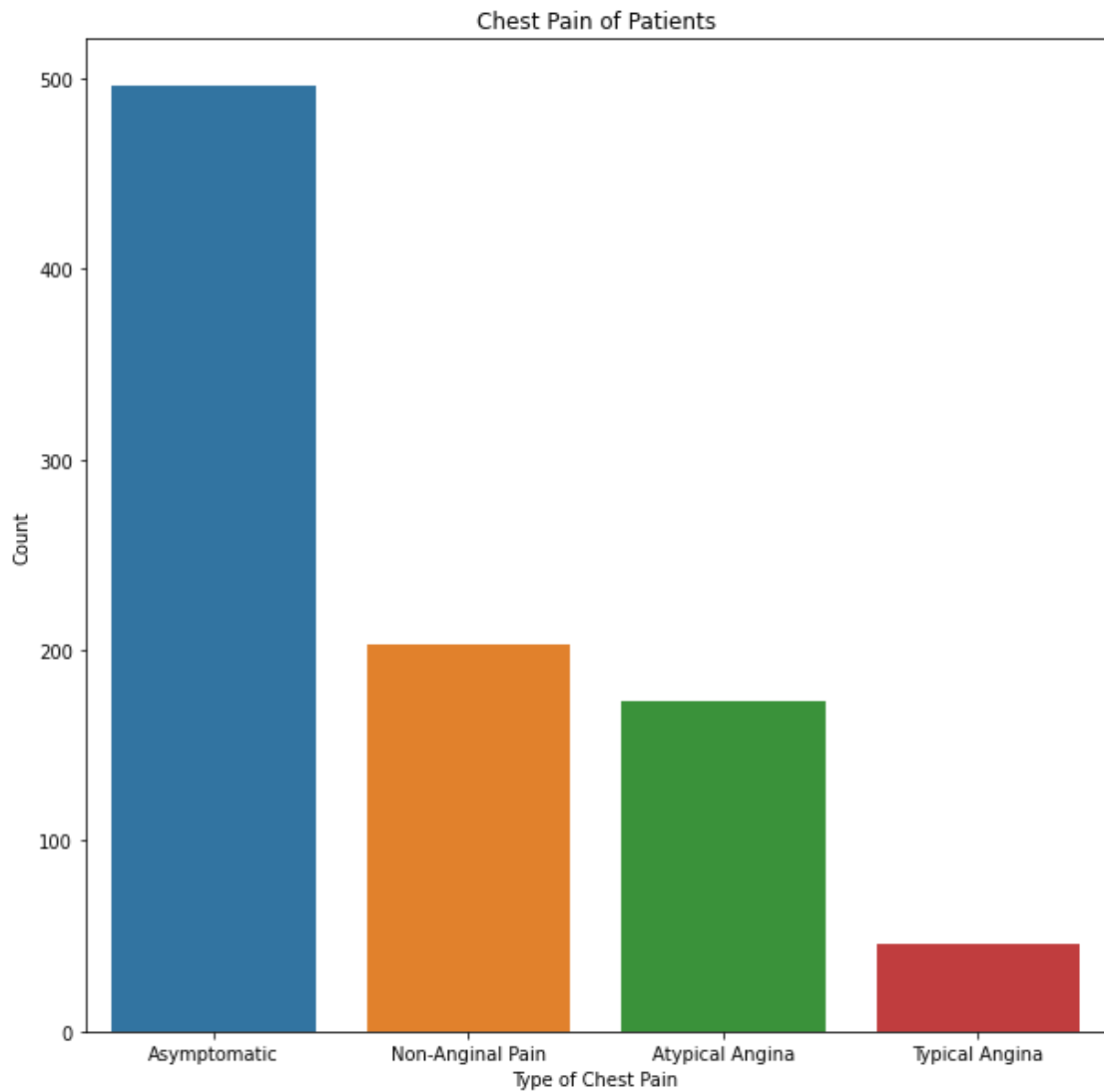
Age	0
Sex	0
ChestPainType	0
RestingBP	0
Cholesterol	0
FastingBS	0
RestingECG	0
MaxHR	0
ExerciseAngina	0
Oldpeak	0
ST_Slope	0
HeartDisease	0
dtype:	int64

## Bar plot for Chest Pain in Patients

```
1 # Chest Pain Attribute
2 cp_data = df['ChestPainType'].value_counts().reset_index()
3 cp_data['index'][3] = 'Typical Angina'
4 cp_data['index'][2] = 'Atypical Angina'
5 cp_data['index'][1] = 'Non-Anginal Pain'
6 cp_data['index'][0] = 'Asymptomatic'
7
8 cp_data
✓ 0.4s
```

	index	ChestPainType
0	Asymptomatic	496
1	Non-Anginal Pain	203
2	Atypical Angina	173
3	Typical Angina	46

```
1 # bar plot to view the types of ChestPain
2
3 plt.figure(figsize=(10,10))
4 plt.title('Chest Pain of Patients')
5 sns.barplot(x='index', y='ChestPainType', data=cp_data)
6 plt.xlabel('Type of Chest Pain')
7 plt.ylabel('Count')
8 plt.show()
✓ 0.2s
```



Label Encoder for categorical data – Converting them to numeric

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	0.479270	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	0.298507	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	0.469320	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	0.354892	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	0.323383	0	Normal	122	N	0.0	Up	0

```

1 from sklearn.preprocessing import LabelEncoder # Label Encoder for categorical data
2
3 le = LabelEncoder()
4
5 df['ChestPainType'] = le.fit_transform(df['ChestPainType'])
6
7 df['RestingECG'] = le.fit_transform(df['RestingECG'])
8
9 df['ExerciseAngina'] = le.fit_transform(df['ExerciseAngina'])
10
11 df['Sex'] = le.fit_transform(df['Sex'])
12
13 df['ST_Slope'] = le.fit_transform(df['ST_Slope'])
✓ 0.5s

```

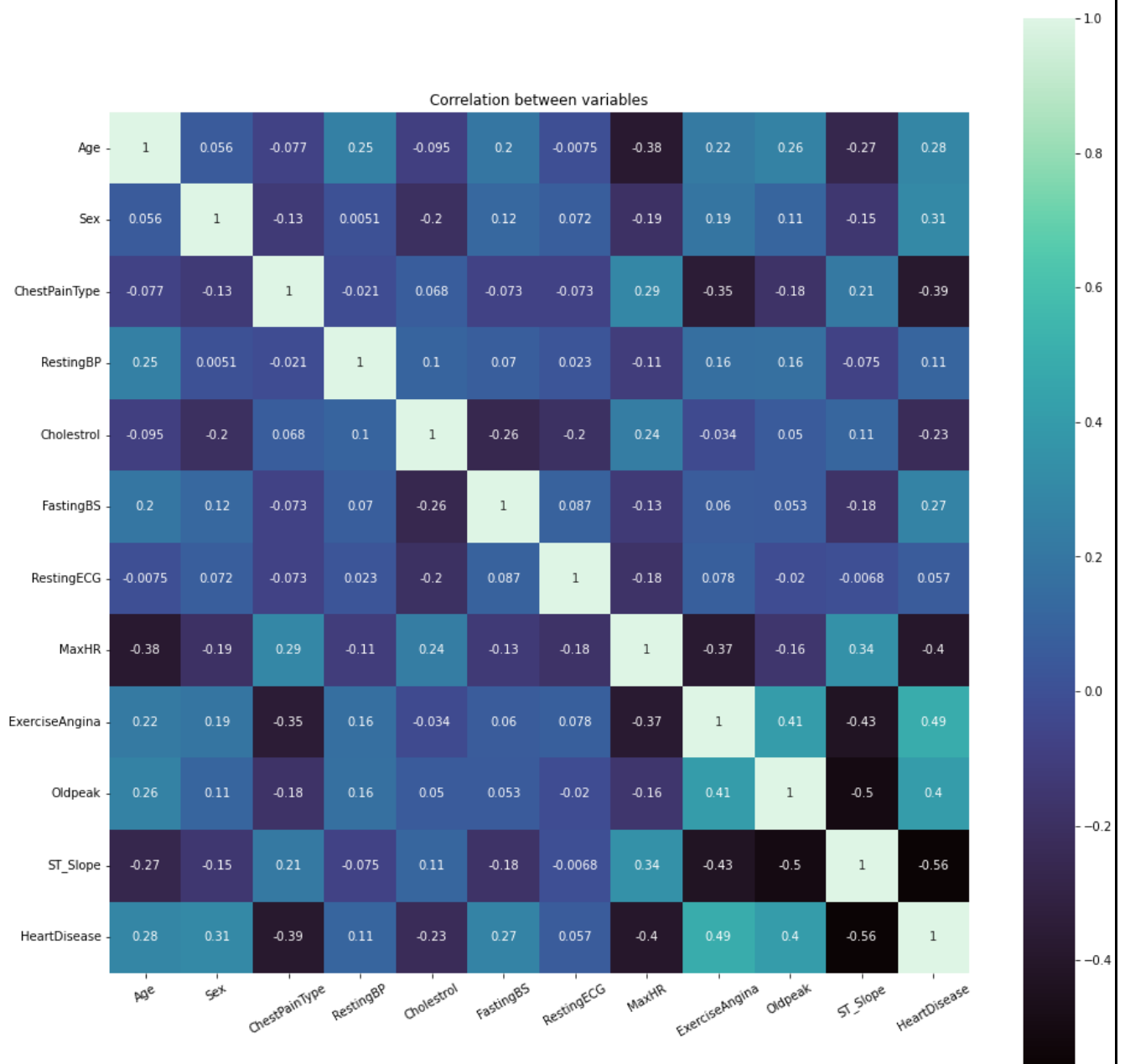
## Feature Scaling

```
1 # Standard Scaler to scale the data (mean = 0 & std dev = 1) for model building
2
3 from sklearn.preprocessing import StandardScaler
4 scale = StandardScaler()
5 scale.fit(df)
```

✓ 0.4s

StandardScaler()

## Correlation between attributes - Heatmap





## Train test split and Model Building – Logistic Regression

```
1 from sklearn.model_selection import train_test_split
2
3 # Train = 0.7 and Test = 0.3
4 X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=1)
```

✓ 0.1s

### Logistic Regression

⌵ ⏪ ⏩ ☐ ... 🗑

```
1 from sklearn.linear_model import LogisticRegression
2 encoded_y = le.fit_transform(y_train)
3 model1 = LogisticRegression()
```

✓ 0.1s

```
1 model1.fit(X_train, encoded_y)
```

✓ 0.6s

LogisticRegression()

```
1 y_pred1 = model1.predict(X_test)
2 encoded_y_test = le.fit_transform(y_test)
```

✓ 0.4s

```
1 from sklearn.metrics import confusion_matrix
2 from sklearn.metrics import accuracy_score
```

✓ 0.6s

```
1 lr_conf_matrix = confusion_matrix(encoded_y_test, y_pred1)
2 lr_accuracy_score = accuracy_score(encoded_y_test, y_pred1)
```

✓ 0.4s

```
1 lr_conf_matrix
```

✓ 0.3s

```
array([[ 96,  13],
       [ 24, 143]], dtype=int64)
```

```
1 print(f"{lr_accuracy_score*100}%") # Accuracy of the model
```

✓ 0.4s

```
86.59420289855072%
```

## Using Grid Search CV for HyperParameter tuning (top 3 performing algorithms)

```
1 from sklearn.model_selection import GridSearchCV
2 from sklearn.metrics import classification_report
3 models_acc
```

✓ 0.6s

	Model	Accuracy
4	SVM	88.405797
2	Random Forest	86.956522
0	Logistic Regression	86.594203
3	KNN	86.231884
1	Decision Tree	76.086957

## For Support Vector Classifier (SVC)

```
1 # defining parameter range
2 param_grid = {'C': [0.1, 1, 10, 100, 1000],
3               'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
4               'kernel': ['rbf']}
5
6 grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
7
8 # fitting the model for grid search
9 grid.fit(X_train, encoded_y)
10
11 # print best parameter after tuning
12 print(grid.best_params_)
13
14 # print how our model looks after hyper-parameter tuning
15 print(grid.best_estimator_)
```

✓ 3.4s

Fitting 5 folds for each of 25 candidates, totalling 125 fits

```
[CV 1/5] END .....C=0.1, gamma=1, kernel=rbf; score=0.690 total time= 0.0s
[CV 2/5] END .....C=0.1, gamma=1, kernel=rbf; score=0.721 total time= 0.0s
[CV 3/5] END .....C=0.1, gamma=1, kernel=rbf; score=0.703 total time= 0.0s
[CV 4/5] END .....C=0.1, gamma=1, kernel=rbf; score=0.641 total time= 0.0s
```

## Final Prediction Accuracies

	Model	Accuracy
0	SVM	88.405797
7	Random Forest_hyper	88.043478
1	Random Forest	86.956522
6	SVM_hyper	86.956522
2	Logistic Regression	86.594203
3	KNN	86.231884
5	Logistic Regression_hyper	84.057971
4	Decision Tree	76.086957

## 14.0 Conclusion

Currently, ML models are still in the testing and experimentation phase for heart disease predictions. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models.

Launching a service that can consistently and accurately predict heart disease in patients is something that the world will witness soon. Machine Learning not only can automate this process but also it can make it available to all sections of the society.

Its interesting to see how lives of certain individuals can be altered drastically as a use case of a technology that has witnessed massive growths in recent years.

What's more interesting is to be a part of it.

## 15.0 References

- [Artificial Intelligence Tool May Help Predict Heart Attacks](#)
- [Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning](#)
- [Heart Disease Prediction](#)
- [Heart Disease Prediction using Exploratory Data Analysis](#)
- [Machine learning prediction in cardiovascular diseases: a meta-analysis](#)