

Heart Failure Diagnosis

using

K-Nearest Neighbor and Logistic Regression

*By*

*Dikshak Khanna*

*Revanth Posina*

*Saurabh Dete*

**Indiana University Bloomington**

# Abstract

Heart Failure being a chronic disease remains one of the costliest conditions to manage worldwide due to its frequency and occurrence. Significant occurrence of Heart Failure results in hospital admission, re-admissions where chronic disease management programs are not running efficiently and resulting in a burden on the Healthcare Industry. As of the current situation in the society we live in, Heart Failure Prediction is one of the most compounded tasks in Medical Field. Predictive analysis of cardiac conditions can provide a patient-specific window into the progression of Heart Failure and guide treatment planning accordingly.

The heart is our body's most actively functioning life-supporting organ. The average human heart beats 100,000 times every day, from day to night, every single day. Engineering has always aided medical sciences in their developments and detections, and medical diagnosis plays a crucial role in medicine. A key element in treating such conditions is early identification. Prediction and diagnosis are crucial since cardiovascular diseases account for 12 million fatalities yearly, a significant problem in modern society. The World Health Organization estimates that in India, heart failure causes 125 fatalities per 100,000 individuals.

In this project Machine learning models are used to enable a program to analyze data, understand correlations and make use of insights to solve the problem by predicting the chances of having a cardiac arrest. In this paper we are comparing 2 major algorithms or predictive analysis techniques called K-Nearest Neighbor Algorithm and Logistic Regression to test the dataset which we have acquired from multiple sources. The Healthcare Industry is suffering from a huge amount of data which is not used effectively. Use of Datamining techniques gives us a leverage over such huge amount of data and removes redundancy.

Technology breakthroughs have led to doctors incorporating contemporary scientific discoveries for cardiovascular disease detection and diagnosis. Doctors may diagnose a patient's heart disease incorrectly due to human error; therefore, prognostic analysis using Machine Learning Algorithms helps in such circumstances to obtain reliable results. The real-world application for this is a website of the predictive analysis of cardiovascular diseases, such as heart failure, utilizing machine learning algorithms. Finally, a conclusion about the level of accuracy and effectiveness of the Machine Learning algorithm is reached.

# *Contents*

## ***Chapter 1:*** Introduction to CVD

- 1.1 Introduction
- 1.2 International Statistical Review
- 1.3 Types of CVDs
- 1.4 Common Symptoms of Major types of CSDs

## ***Chapter 2:*** Proposed Framework

- 2.1 Dataset Description
- 2.2 Explanation of Features of Dataset
- 2.3 Exploratory Data Analysis (EDA)
- 2.4 Dimensionality Reduction using PCA

## ***Chapter 3:*** Algorithms and Analysis

- 3.1 K-Nearest Neighbor Algorithms
- 3.2 Logistic Regression

## ***Chapter 4:*** Proposed Real World Application

## ***Chapter 5:*** Discussion

- 5.1 Discussion of Results

# Chapter 1

## Introduction to CVD

### 1.1 Introduction

Heart and/or blood vessel diseases are collectively referred to as cardiovascular disease (CVD) in this context (vascular). Many cardiovascular diseases lead to long-lasting, chronic conditions that develop or endure. However, it can also lead to sudden, acute occurrences like heart attacks and strokes, which happen when a blood vessel supplying the brain, or the heart becomes blocked. Smokers, those with high blood pressure, high cholesterol, those who are overweight, do not exercise, and/or those with diabetes are more likely to develop CVD.

More people die from CVDs worldwide than from any other cause, over 17.9 million every year, according to the World Health Organization. Of these deaths, 80% are due to coronary heart disease (e.g.: Heart attack) and cerebrovascular diseases (e.g.: strokes) and mostly affect low and middle-income countries.

**Keywords:** Cardiovascular Disease, Exploratory Data Analysis, Dimensionality Reduction, K-nearest neighbors, Cosine Distance, Minkowski Distance Logistic Regression, Sigmoid Activation Function.

### 1.2 International Statistical Review

In the year 2000, 76,426 people died from CVD, with 34% of male fatalities and 36% of female deaths. While CHD and AMI mortality rates are still on the decline, there has been little change in stroke fatality rates over the previous ten years. This has led to an increase in the number of stroke and CHD-related deaths. The next 15 years are anticipated to see this trend continue.

Compared to the year 2000, in 2030, the number of years of productive life lost to CVD will have increased by 20% in the United States and 30% in Portugal. Brazil has a figure of 64%, China has a figure of 57%, and India has a figure of 95%. South Africa's increase is 28%, which is higher than the US and comparable to Portugal. Only in Russia does the number of years lost lag, because death rates are already at such high levels and the population at risk shrinking.

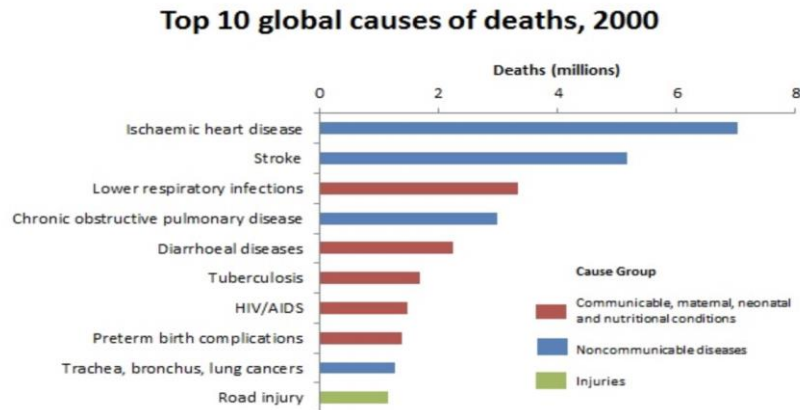


Fig. 1: Mortality Rate

Fig. 1 shows heart disease as one of the most death causing diseases around the world as per the year 2000

### 1.3 Types of CVDs:

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels, and it includes:

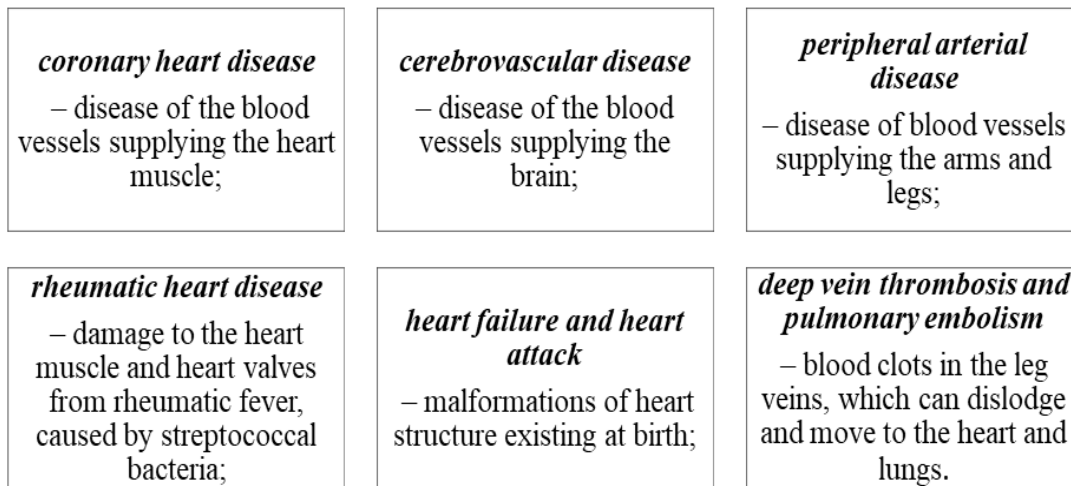


Fig 2: Types of CVDs

Fig. 2 shows a block diagram of different known types of CVDs

## 1.4 Common Symptoms of Major Types of CVDs

### Coronary Heart Disease:

- **Pain areas** in the chest.
- **Gastrointestinal:** indigestion or nausea.
- **Whole body** light-headedness or sweating.
- **fast heart rate** or shortness of breath.

### Cerebrovascular disease:

- **Muscular:** Difficulty walking, instability, paralysis with weak muscles, problems with coordination, stiff muscles, overactive reflexes, or paralysis of one side of the body.
- **Visual:** Blurred vision, double vision, sudden visual loss, or temporary loss of vision in one eye.
- **Speech:** Difficulty speaking, slurred speech, or speech loss.
- **Whole body:** Fatigue, light-headedness, or vertigo.
- **Limbs:** Numbness or weakness.
- **Sensory:** Pins and needles or reduced sensation of touch.
- **Facial:** Muscle weakness or numbness.

### Peripheral arterial disease:

- **Pain areas:** In the buttocks.
- **Pain circumstances:** Can occur in the leg and improved with rest or in the leg while exercising.
- **Skin:** Cool skin, loss of hair on the legs, thinning of skin on the legs, or ulcers.

### Heart failure and heart attack:

The most common symptom of a stroke is sudden weakness of the face, arm, or leg, most often on one side of the body. Other symptoms include sudden onset of:

- Numbness of the face, arm, or leg, especially on one side of the body.
- Confusion, difficulty speaking or understanding speech.
- Difficulty seeing with one or both eyes.
- Difficulty walking, dizziness, loss of balance or coordination.
- Severe headache with no known cause; and
- Fainting or unconsciousness.

## Chapter 2

### Proposed Framework

Our suggested approach makes use of Cardiovascular dataset to predict the chances of a patient getting heart attack based on the symptoms they are having. Fig 2 depicts a high-level view of the project framework's primary components. During the data pretreatment stage, all databases are cleaned to remove any potential noise and unusable outliers, then converted and integrated depending on attributes character. The feature extractor extracts and generates handmade attributes that are intended to describe the character and carry the value of that respective part of the data. The goal of adaptive filtering modules is to improve accurate channels by assessing the important attributes and removing any unnecessary characteristics that does not add anymore meaning or value to the present information. Finally, we use the selected criteria to build our predictive model, to predict the patient's result based on the symptoms he's having.

### 2.1 Dataset Description

- **Age:** Integer value to take the age of the individual under consideration
- **Sex:** Below represents the gender of the individual using the following format,  
1 = male  
0 = female
- **Chest-pain type(cp):** Below is the representation of the type of chest-pain experienced by the individual using the following format,  
1 = typical angina  
2 = atypical angina  
3 = non-anginal pain  
4 = asymptotic
- **Resting Blood Pressure(trestbp):** Integer value is taken for the resting blood pressure value of an individual in mmHg.
- **Serum Cholesterol(chol):** Integer value is taken for the serum cholesterol in mg/dl for an individual.
- **Fasting Blood Sugar(fbs):** Below is the representation for the fasting blood sugar value of an individual with 120mg/dl.
  - ❖ If fasting blood sugar > 120mg/dl, then: 1 (true)
  - ❖ Else: 0 (false)

- **Resting ECG (restecg):** Below is the representation for resting electrocardiographic results  
0 = normal  
1 = having ST-T wave abnormality  
2 = left ventricular hypertrophy
- **Max heart rate achieved(thalach):** displays the max heart rate achieved by an individual.
- **Exercise induced angina(exang):** Below is the representation,  
1 = yes  
0 = no
- **ST depression induced by exercise relative to rest (oldpeak):** Displays the value which is an integer or float.
- **Thalassemia(thal):**  
3 = normal  
6 = fixed defect  
7 = reversible defect
- **Diagnosis of heart disease:** Represents whether the individual is suffering from heart disease or not,  
0 = absence  
1 = present.

## 2.2 Explanation of the features of the Dataset

In the actual dataset, we had 76 features but for our study, we chose only the above 12 features that will help use predict whether the patient will suffer from a heart attack or not. We selected these 14 parameters after consulting with a family doctor and for the below mentioned reasons as well,

- **Age:** The most important risk factor in developing cardiovascular or heart diseases is age, with the risk roughly tripling with each decade of life. According to estimates, 82 percent of people who die from coronary heart disease are 65 or older.
- **Sex:** Men are more likely than premenopausal women to develop heart disease. After menopause, it has been argued that a woman's risk is comparable to a man, though more recent data from the WHO and UN challenges this. A female with diabetes is much more likely than a male with diabetes to develop heart disease.



- **Angina (Chest Pain):** Angina is a type of chest pain or discomfort caused by a lack of oxygen-rich blood to your heart muscle. In your chest, you may feel pressure or squeezing. You may also experience pain in your shoulders, arms, neck, jaw, or back. Angina pain can even mimic indigestion.
- **Resting Blood Pressure:** High blood pressure can damage the arteries that supply your heart over time. High blood pressure combined with other conditions, such as obesity, high cholesterol, or diabetes, raises the risk of CVD even further.
- **Serum Cholesterol:** High level of low-density lipoprotein (LDL) cholesterol increases the chance of artery narrowing. A high level of triglycerides, a type of blood fat related to your diet, increases your risk of having a heart attack.
- **Fasting Blood Sugar:** When your pancreas does not produce enough insulin or your body does not respond properly to insulin, your blood sugar levels rise, increasing your risk of a heart attack.
- **Resting ECG:** The USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits for people at low risk of cardiovascular disease. Current evidence is insufficient to assess the balance of benefits and harms of screening for people at intermediate to high risk.
- **Max heart rate achieved:** The increase in cardiovascular risk associated with increased heart rate was comparable to the increase in risk associated with high blood pressure. It has been demonstrated that an increase in heart rate of 10 beats per minute is associated with a 20% increase in the risk of cardiac death.
- **Exercise induced angina:** Angina pain or discomfort is typically tight, gripping, or squeezing, and can range from mild to severe. Angina is typically felt in the center of your chest, but it can spread to one or both shoulders, as well as your back, neck, jaw, or arm. It's even palpable in your hands.

## 2.3 Exploratory Data Analysis (EDA)

In this section we will go through the analysis of the dataset and how we have cleaned the dataset and visualized the cleaned dataset. We have a total dataset of 6283 patients out of which we have taken a training test dataset split of 80% Training Set and 20% Test Set.

- ✓ **Checking for null values:** At first, we are taking the dataset and checking if it has any null values in it. We found that the feature column “Chol” which represents cholesterol and the other feature “Thalach” which represents maximum heart rate has

11 and 12 null values respectively. We replaced them with the respective mean of the columns. Please refer to the below screenshot for the number of null values which we found.

```
Number of Null values:
age: 0
sex: 0
cp: 0
trestbps: 0
chol: 11
fbs: 0
restecg: 0
thalach: 12
exang: 0
oldpeak: 0
thal: 0
target: 0
```

Fig 3: #Null Values present

*Fig. 3 is an output recorded after checking for null values in the dataset and it shows a total of 22 null values*

- ✓ **Checking for Missing values:** Moving forward we checked for missing values in the dataset and found out that there is no missing value present in any of the features in the dataset. Please refer to the below attached screenshot.

```
Number of Null values:
age: 0
sex: 0
cp: 0
trestbps: 0
chol: 0
fbs: 0
restecg: 0
thalach: 0
exang: 0
oldpeak: 0
thal: 0
target: 0
```

Fig 4: #Missing Values present

*Fig. 4 shows the number of missing values in the dataset*

- ✓ **Visualization of Dataset:** Please refer to the below screenshot which shows a histogram plot for each feature and the representation of their values respectively.

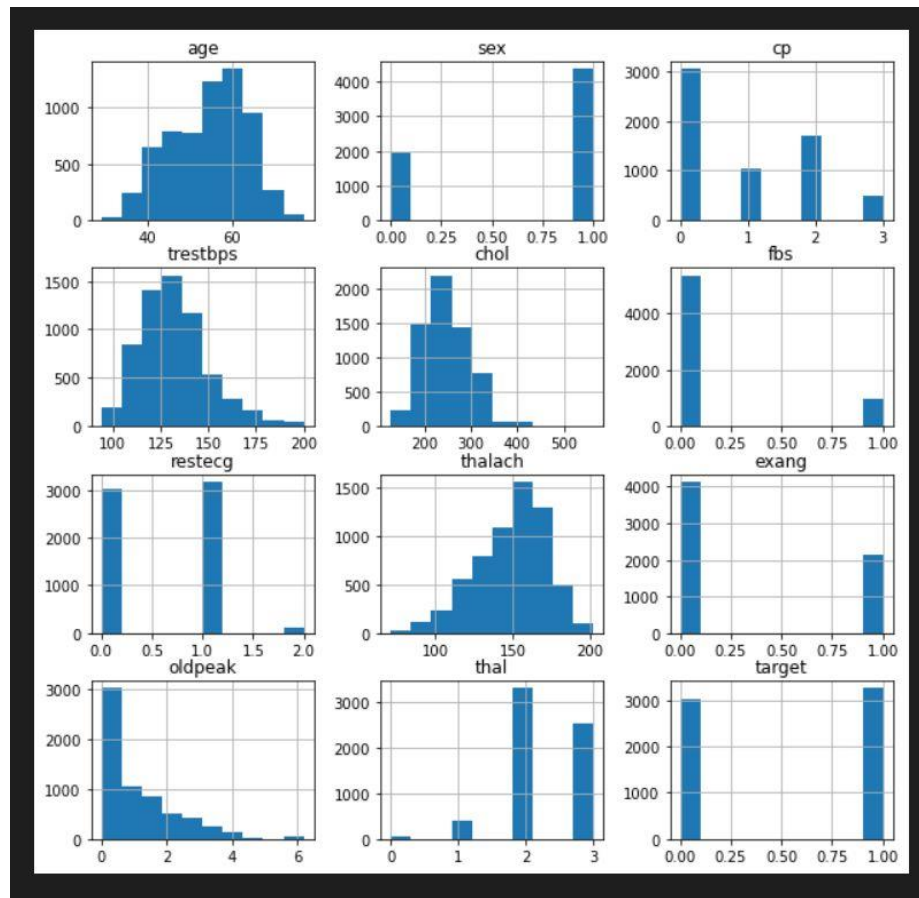


Fig 5: Visualization of Features in the Dataset

*Fig. 5 shows a histogram of all the features from the dataset, it represents a bar graph of the values varying from feature to feature and helps to understand more about the attributes*

- ✓ **Detection of Outliers:** Next we are plotting a box plot for each feature to get an idea how much the data is skewed because of the outliers present in the dataset for each feature separately. Fig 6 shows the box plot of the dataset below. We can infer from the box plot that for the features “trestbps” which represents Resting Blood Sugar Level and “chol” which represents cholesterol level. To get an idea of the number of outliers present we first got the z score for the dataset, and we defined a range of “ $-3 < z < 3$ ” and found that there are a total of 124 outliers present in the dataset. We decided not to remove the outliers or replace them with the mean since we did not want to overfit our model.

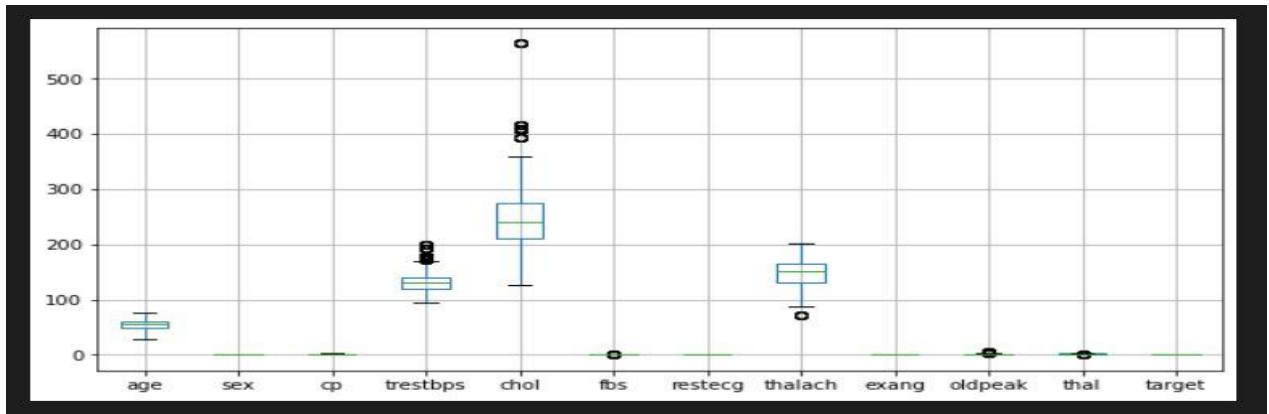


Fig 6: Box plot of the dataset

*Fig. 6 shows boxplot plotted to study the outliers*

- ✓ **Removing Duplicates:** Lastly, we updated the data by removing the duplicate values in it. There was a total of 958 duplicate values. Finally, we got a cleaned dataset for 5,325 patients or individuals.

## 2.4 Dimensionality Reduction using PCA

After cleaning the dataset, we moved forward with Standard Scalarization of the dataset. This is done because all the features in the dataset have different range and they signify different risk factors beyond a particular level. Therefore, to standardize the data we use Standard Scalar function for it.

Moving forward we proceed to dimensionality reduction using Principal Component Analysis technique. This was done to discover a new collection of dimensions (attributes) that better captures the data's variability. The first dimension is chosen intentionally to capture as much variety as possible. The second dimension is orthogonal to the first and captures as much of the remaining variability as possible within that limitation, and so on.

## Chapter 3

### Algorithms and Analysis

#### 3.1 K-Nearest Neighbor Algorithm

K-nearest neighbors (KNN) is a supervised learning algorithm that can be used for regression and classification. By calculating the distance between the test data and all the training points the algorithm chooses the K number of points that are closest to the test data and then attempts to predict the correct class for the test data. The KNN algorithm calculates the likelihood of test data belonging to the classes of 'K' training data, and the class with the highest probability is chosen.

##### **Advantages:**

1. **Training Period:** KNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training.
2. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.
3. KNN is very easy to implement. There are only two parameters required to implement KNN i.e., the value of K and the distance function (e.g., Minkowski or Cosine etc.)

##### **Disadvantages:**

1. In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.
2. The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

Assume there are two categories A (person getting a cardiac arrest) and B (person getting a cardiac arrest, and we have a new data point  $x_1$ . To determine which of these categories this data point belongs to, K-NN algorithm comes in handy.

We can easily identify the category or class of a particular dataset using K-NN. There are several distance metrics we can use to compute KNN. Metrics like Euclidean, Manhattan, Minkowski, Cosine are often used.

The **Minkowski distance** between two variables X and Y is defined as

$$(\sum_{i=1}^n |X_i - Y_i|^p)^{1/p}$$

The case where  $p = 1$  is equivalent to the Manhattan distance and the case where  $p = 2$  is equivalent to the Euclidean distance.

### **Cosine Distance:**

The cosine similarity is described mathematically as the division between the dot product of vectors and the product of the Euclidean norms or magnitude of each vector.

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

### **Working of KNN Algorithm:**

- Selecting K value to define the neighbors.
- Calculate the distance (based on metric you choose) of K number of neighbors to the given data point
- Take the K nearest neighbors as per the calculated distance.
- Among these k neighbors, count the number of the data points in each category.
- Now assign the new data points to that category for which the number of the neighbors are maximum and that completes classification.

#### **3.1.1 Deploying KNN:**

We are using the predefined library sklearn to deploy KNN on our dataset. We first initialized the value of  $k=3$  and ran the algorithm and got an accuracy of 48.63%.

We first decided to deploy the KNN algorithm for unscaled data and compared the results for both Minkowski and Cosine distance which was used. We defined a range of (2,70) for the value of “k” and plotted the error rate against the value of “k” in order to get the most efficient value of it.

We found that for unscaled data the accuracy of the model and value of efficient “k” is as below:

1. Minkowski Distance: 50.14%, k=48
2. Cosine Distance: 47.88%, k=8

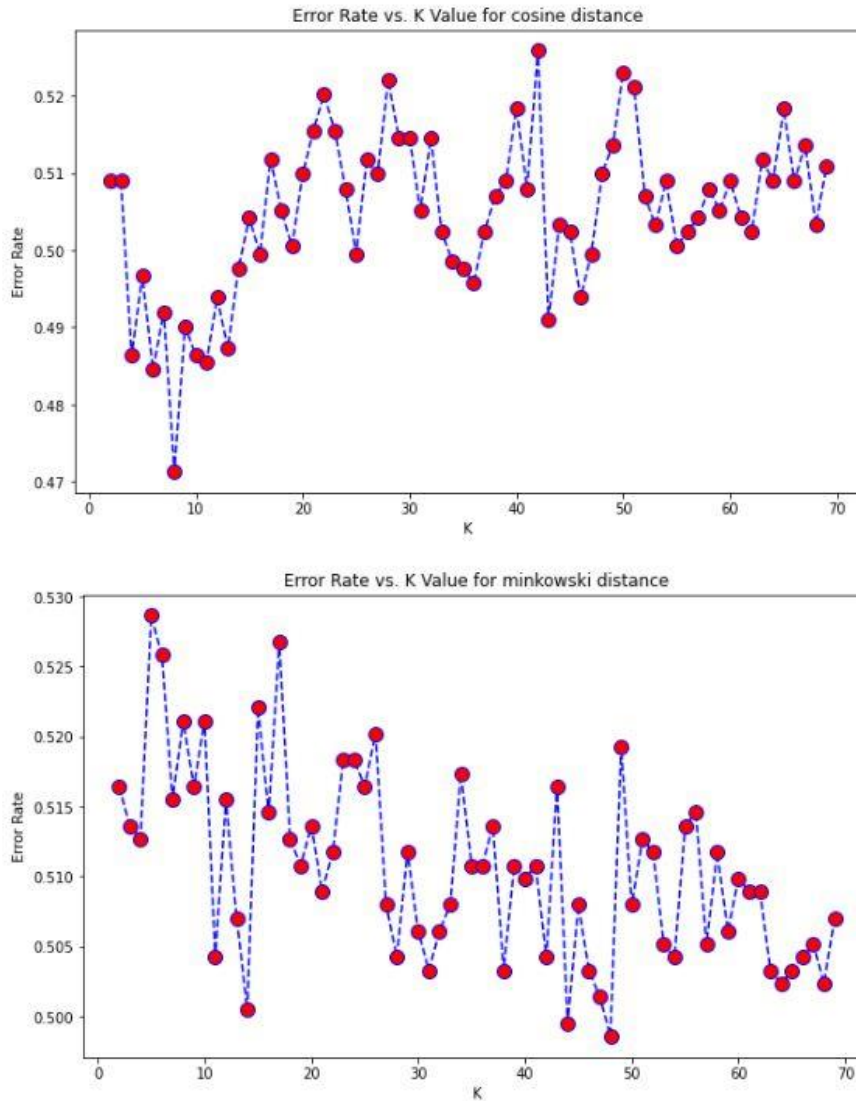


Fig 7: Represents the Error Rate vs K value for Unscaled Data

*Fig. 7 shows a comparison of graphs between Cosine and Minkowski distances for Error Rate vs K value, the effective K value taken based on this graph analysis is mentioned further in the document (Unscaled Data)*

Similarly, we moved forward with scaled data and found out the below result,

1. Minkowski Distance: 53.99%,  $k=38$
2. Cosine Distance: 48.45%,  $k=9$

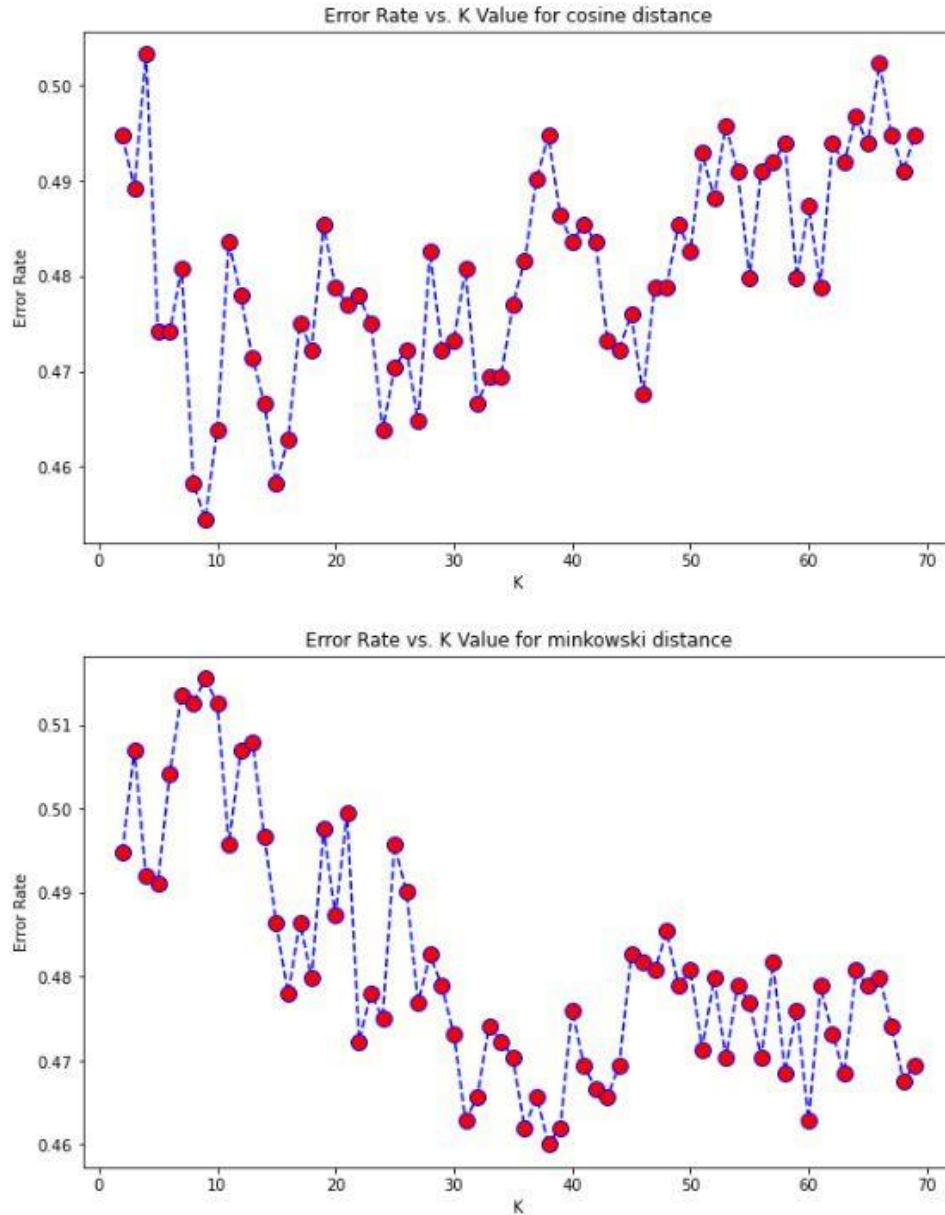


Fig 8: Represents the Error Rate vs K value for Scaled Data

*Fig. 8 shows a comparison of graphs between Cosine and Minkowski distances for Error Rate vs K value, the effective K value taken based on this graph analysis is mentioned further in the document (Scaled Data)*



## 3.2 Logistic Regression

Logistic Regression statistical model frequently used for classification and predictive analytics. Based on a set of independent variables, logistic regression calculates the likelihood of occurring of an event. As the outcome is a probability, the dependent variable has a range of 0 to 1. A probability less than 0.5 predicts 0 in binary classification, while a probability greater than 0.5 predicts 1. In logistic regression probability of success divided by the probability of failure. This is also known as the log odds or the natural logarithm of odds, and it is represented by the following formulas.

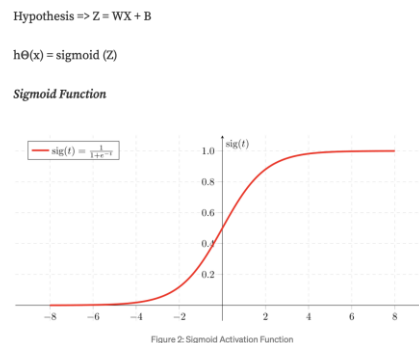
$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Where  $\mu$  is a location parameter (the midpoint of the curve, where  $p(\mu) = 1/2$  and  $s$  is a scale parameter. This expression may be rewritten as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Where  $\beta_0 = -\mu/s$  and is known as the intercept (it is the y-intercept of the line  $y = \beta_0 + \beta_1 x$  and  $\beta_1 = 1/s$  (inverse scale parameter) these are the y-intercept and slope of the log-odds as a function of  $x$ . Conversely,  $\mu = -\beta_0 / \beta_1$  and  $s = 1 / \beta_1$ .

In this model, the beta parameter, or coefficient, is commonly estimated using maximum likelihood estimation (MLE). This method iteratively tests different beta values to find the best fit of log odds. The log likelihood function is produced by all these iterations, and logistic regression seeks to maximize this function to find the best parameter estimate.



**Fig 9: Sigmoid Activation Function**

*Fig. 9 shows Sigmoid Activation function, a function used to map any real value into another value between 0 and 1*

If 'Z' attains infinity, Y(predicted) becomes 1, and if 'Z' attains negative infinity, Y(predicted) will become 0. The data is fit into a linear regression model, which is then used by a logistic function to predict the target categorical dependent variable.

We are using the predefined library sklearn to deploy Logistic Regression on our dataset. We first decided to deploy the Logistic Regression algorithm for unscaled data and then for scaled data.

We found the accuracy of the model as below:

1. Unscaled Data: 52.20%
2. Scaled Data: 50.61%

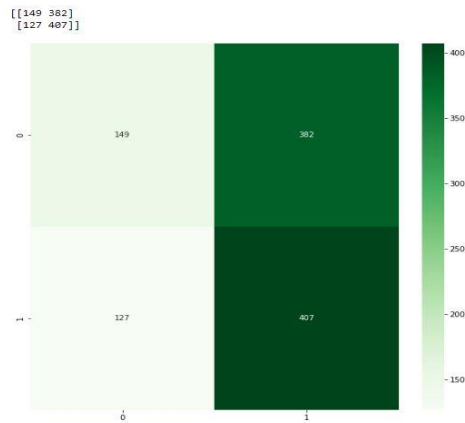


Fig 10: Confusion Matrix for Unscaled Data

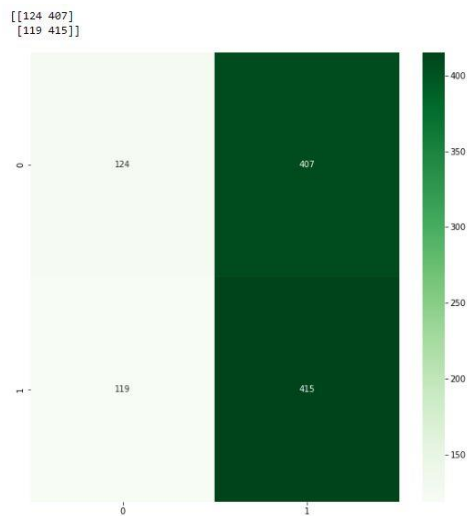


Fig 11: Confusion Matrix for Scaled Data

Fig. 10 and Fig. 11 shows Confusion Matrix plotted for Unscaled and Scaled data respectively

# Chapter 4

## Proposed Real World Application

In this paper, we took a dataset of 6283 patients and deployed K-NN and Logistic Regression on it. We recognized error rate as well as accuracy in prediction of Heart Failure for both the algorithms. Here we aim to provide the real-world application by deploying a webpage interface so that anyone anywhere can get access and be prepared beforehand. The following figures are that of the proposed webpage.

### A. Login Page:

This is the first page where the basic details of the patients are requested so that a database can be maintained.

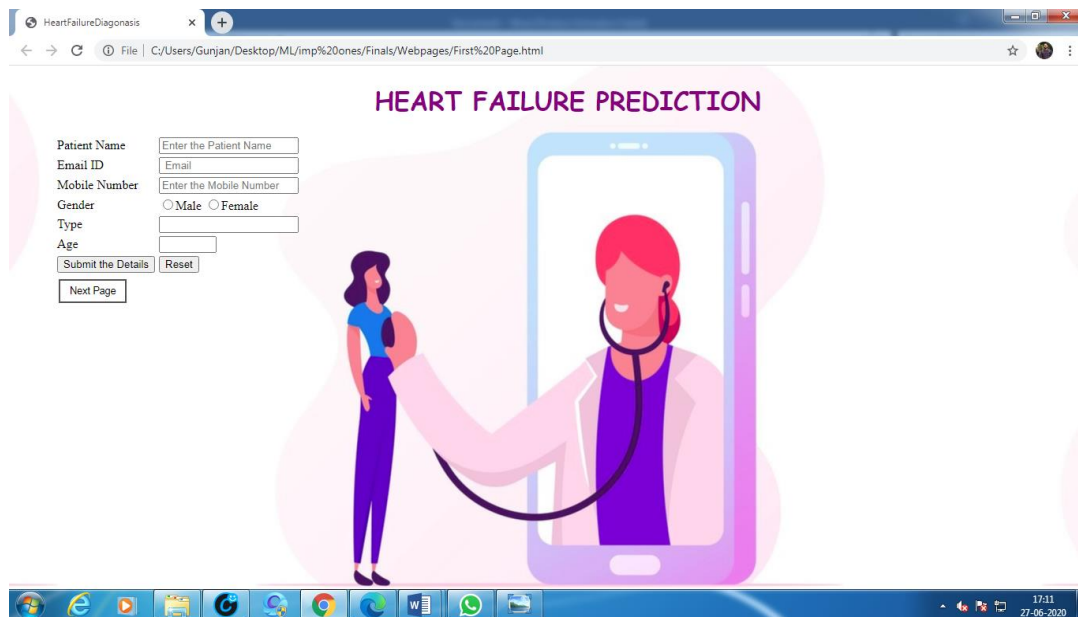
The image shows a web browser window displaying a login page titled "HEART FAILURE PREDICTION". The page features a form on the left with fields for Patient Name, Email ID, Mobile Number, Gender (Male/Female), Type, and Age. Below these fields are "Submit the Details" and "Reset" buttons, followed by a "Next Page" button. The background of the page has a large illustration of a doctor in a white coat using a stethoscope to examine a patient's chest. The doctor is shown from the waist up, and the patient is shown from the side. The entire scene is set against a light blue and white background with a subtle grid pattern. The browser's address bar shows the file path "C:/Users/Gunjan/Desktop/ML/imp%20ones/Finals/Webpages/First%20Page.html". The Windows taskbar at the bottom shows various application icons and the system clock indicating 12:11 on 27-06-2020.

Fig 12: Login Page

*Fig. 12 shows our Login page of our user-interactive website, this page allows user to provide his details and start the process*

### B. Clinical Symptoms:

Both the figures represent the webpage where the patient can select the symptoms he has observed and then proceed to the next page. This helps to determine the type of Heart Failure.

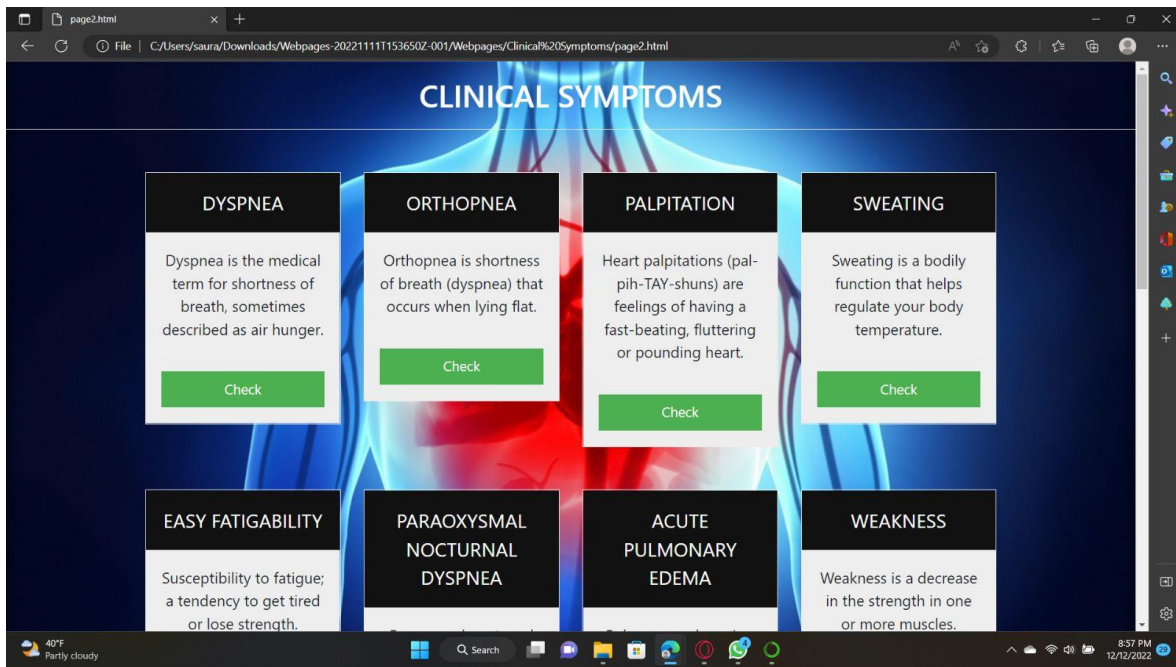


Fig 13a: Clinical symptoms

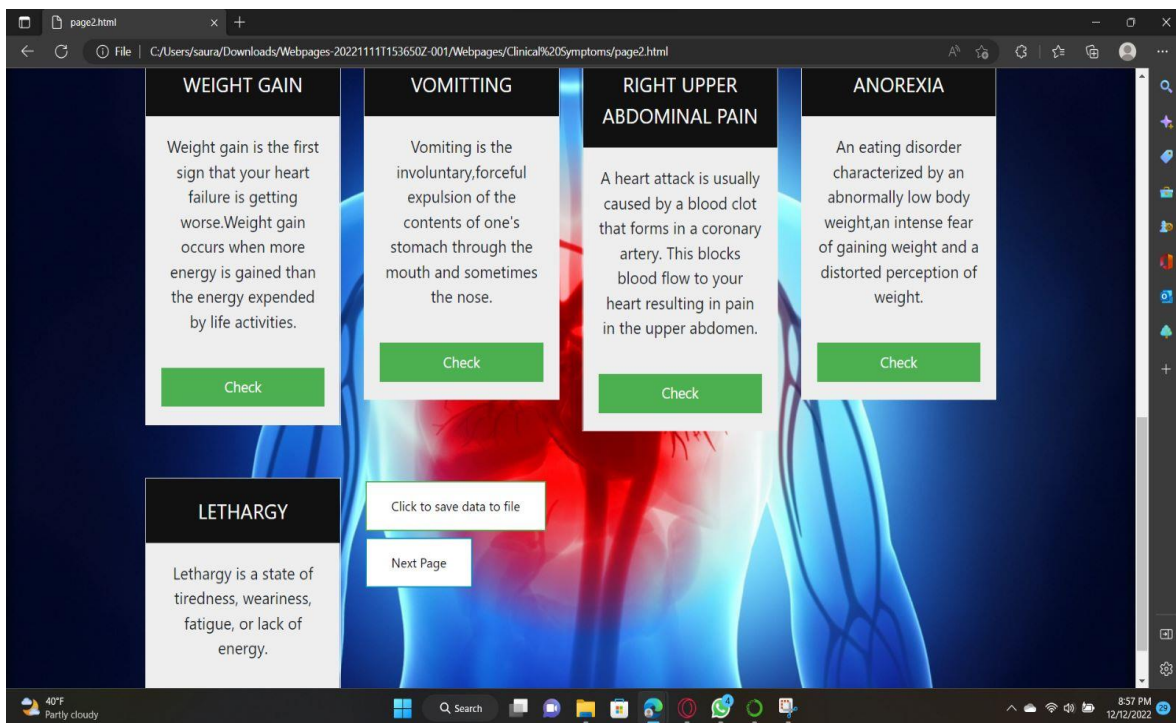


Fig 13b: Clinical Symptoms

*Fig. 13a and Fig. 13b shows Clinical Symptoms page, this is where user will be able to select the symptoms accordingly*

## C. Collection of Parameter Value:

Here the patients enter the value of the 14 major attributes which will determine the output and proceeds to the next page. The following figures display the webpage for the above-mentioned criteria.

The screenshot shows a web browser window with the title 'page3.html'. The URL is 'C:/Users/saura/Downloads/Webpages-20221111153650Z-001/Webpages/Data%20Input/page3.html'. The page has a blue background with a cartoon doctor character. The title 'ENTER NECESSARY DATA' is in yellow. There are eight input boxes arranged in two rows. Each box has a title, instructions, a text input field, and a green 'Submit' button. The first row contains: 'Chest Pain Type' (instructions: 1=Typical Angina, 2=Atypical Angina, 3=Non-Anginal Pain, 4=Asymptotic), 'Resting Blood Pressure' (instructions: Enter the resting Blood Pressure:), 'Serum Cholesterol' (instructions: Enter the Serum Cholesterol value:), and 'Fasting Blood Sugar' (instructions: Enter Fasting Blood Sugar Level: 1=If Blood sugar>120mg/dl, 0=Otherwise). The second row contains: 'Resting ECG' (instructions: Enter the resting), 'Maximum Heart Rate' (instructions: Enter the value:), 'Exercise induced Angina' (instructions: Enter the value:), and 'ST depression induced by exercise relative to rest' (instructions: Enter the value:). The Windows taskbar at the bottom shows the date and time as 8:58 PM 12/12/2022.

Fig 14a: Entering of necessary data.

The screenshot shows the same web browser window as Fig 14a, but with different input boxes visible. The first row contains: '1=Having ST-T wave abnormality 2=Left Ventricular Hypertrophy' (instructions: Enter the value:), 'Number of Major Vessels colored by Fluoroscopy' (instructions: Enter the value:), and 'Thalassemia' (instructions: Enter from the following: 0=normal, 1=Fixed Defect, 2=Reversible Defect). The second row contains: 'Peak exercise ST segment' (instructions: Choose from the following: 1=Upsloping, 2=Flat, 3=Downsloping), 'Number of Major Vessels colored by Fluoroscopy' (instructions: Enter the value:), and 'Thalassemia' (instructions: Enter from the following: 0=normal, 1=Fixed Defect, 2=Reversible Defect). There are also buttons for 'Click to save data to file' and 'Next Page'. The Windows taskbar at the bottom shows the date and time as 8:58 PM 12/12/2022.

Fig 14b Entering of necessary data.

*Fig. 14a and Fig. 14b allows the user to enter some necessary data to evaluate the condition*

## D. Output Page:

Finally, the output for the entered values of the parameters is entered in the learning algorithm obtained and a final prediction is displayed in the format below.

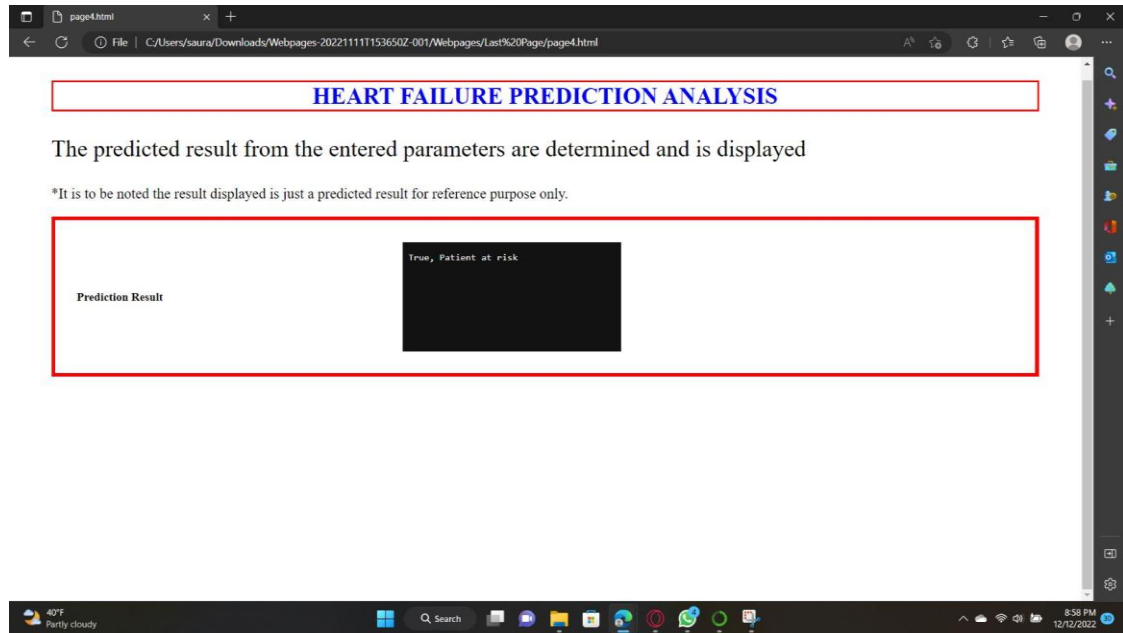


Fig 15: Heart failure prediction analysis

*Fig. 15 fetches the output from the model and shows it on the website, this interface lets the user know the final result.*

## **Chapter 5**

### **Discussion**

#### **5.1 Discussion of result**

In this paper, an experiment to find the predictive performance of K-Nearest Neighbour algorithm and Logistic Regression was carried with the help of a dataset of 6283 patients. We also concluded that as the value of  $k$  didn't vary much when it came to the performance measure or the error rate, but it was found that for scaled data we observed that the performance or the accuracy was better compared to unscaled data. It was the same for Logistic Regression as well where the scaled data performed better than the unscaled data. Therefore, an optimum value of  $k$  i.e.,  $k = 38$  was observed when we used Minkowski Distance for scaled data compared to Cosine distance if used which was  $k = 9$ . Therefore, even if the value of  $K$  is less it might not be the case that the model will be accurate, hence we can say that  $k$  value can be more or even less, and it depends on dataset to dataset.

Furthermore, we took a linear regression approach and found out that if we compare K-NN and Logistic Regression then the better accuracy was found for K-NN algorithm for scaled data.

We have proposed a system of webpage combined with the learning algorithm to provide the masses access to such Machine Learning algorithms and help lifting of the burden from the healthcare industry and putting loads of data to a more organized use. We believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

## **References:**

1. Introduction to Data Mining – Pang, Kumar, Steinbach 12<sup>th</sup> Edition
2. Data Mining Concepts and Techniques – Han and Kambel 2<sup>nd</sup> Edition
3. Kaggle for dataset
4. A V.V Ramalingam, “heart disease prediction using machine learning,” from Science pubco. (2018)
5. A Sanjay Kumar Sen, “Predicting and Diagnosing of Heart Disease Using Machine Learning,” International Journal of Engineering and Computer Science. (2017)
6. A Purushottama, Prof. (Dr.) Kanak Saxenab, Richa Sharma, “Efficient Heart Disease Prediction System”, science direct. (2016)
7. A Jagdeep Singh, Amit Kamra, Harbhag Singh, “Prediction of Heart Diseases Using Associative Classification”, IEEE. (2016)