

CENTRE FOR INNOVATIVE STUDIES

BCA – 694

Heart Disease Prediction Using Machine Learning Algorithm

BY

BCA_PROJ_2023_<Group_no4>

Name of Students
1. Ayan Bhattacharjee
2. Bikiran Majumdar
3. Baibhab Bagchi
4. Manmohan Rauth
5. Rupal Mondal

1. ABSTRACT

The main objective of this research is to develop an Intelligent System using **data mining** modelling technique, Naive Bayes. It retrieves hidden data from stored databases and compares the user values with a trained data set. It can answer complex queries for diagnosing **heart disease** and thus assist healthcare workers to make an intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. Heart is the most essential or crucial organ of our body. Heart is used to maintain and conjugate blood in our body. There are a lot of cases in the world related to heart diseases. People are dying due to heart disease. Various symptoms are mentioned. The health care industries found a large amount of data. This paper gives the idea of predicting heart disease using machine learning algorithms. Here, we will use several machine learning algorithms like **KNN , Naive Bayes , Linear Regression , SVM** & after that we will use **Ensemble Methods** & see the accuracy rate of each Algorithm. The algorithms are used on the basis of features and for predicting heart disease.

Keyword: Data Mining, Naive Bayes , KNN , heart disease, prediction , Linear Regression , SVM , Ensemble Method.

2. OBJECTIVES :

Most hospitals today employ sort of hospital information systems to manage their healthcare. These systems typically generate huge amounts of data. There is a wealth of hidden information in these data that is largely untapped. How data is turned into useful information that can enable healthcare practitioners to make intelligent clinical decisions. The main objective of this research is to develop a Decision Support in Heart Disease Prediction System (DHDPS) using one data mining modelling technique, Naïve Bayes. DSHDPS is implemented as web based questionnaire application. Based on user answers, it can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. We provide the report of the patient in two ways using chart and pdf which indicates whether that particular patient having the heart disease or not. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions. The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Thus the effort to utilize knowledge and experience of numerous specialists and clinical screening data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Providing precious services at affordable costs is a major constraint encountered by the healthcare organizations (hospitals, medical centers). Valuable quality service denotes the accurate diagnosis of patients and providing efficient treatment. Poor clinical decisions may lead to disasters and hence are seldom entertained. Besides, it is essential that the hospitals decrease the cost of clinical test. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Naive Bayes or Bayes' Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables.

3. SCOPE OF THE PROJECT :

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. The application is fed with varied details and therefore the cardiovascular disease related to those details. The application permits user to share their heart connected problems. It then processes user specific details to ascertain for varied illness that might be related to it. Here we tend to use some intelligent data mining techniques to guess the foremost correct illness that might be related to patient's details. Based on result, system automatically shows the result specific doctors for more treatment. The system permits user to look at doctor's details. The system can be use in case of emergency.

4. PROBLEM DEFINITION

- Now a Days A major challenge facing in Health organizations(hospitals , medical centers) is to provide Quality of services at affordable costs. Quality of Service is mainly depends on correct prediction of doctor .
- Poor clinical decisions can lead to disastrous consequences which are unacceptable.
- So to Reduce the cost of clinical tests , organizations need computer based information for prediction of best decision.

5. DATA SET EXPLANATION

- **Gender**

Male : 1

Female : 0

- **Age**

<25 : Low

>25 & <50 : Medium

> 50 : High

- **Current Smoker**

Yes : 1

No : 0

- **BPMeds**

BPMeds : Blood Pressure Medicine

Yes : 1

No : 0

- **Stroke**

Yes : 1

No : 0

- **Hypertension**

Yes : 1

No : 0

- **Diabetes**

Yes : 1

No : 0

- **Cholesterol**

< 200 : Low
>= 200 AND <= 240: Medium
> 240 : High

- **Systolic**

> 120 : High
< 120 : Low

- **Diastolic**

> 80 : High
<80 : Low

- **HeartRate**

<60 : Low
>=60 & <=100 : Medium
>100 = High

- **Heart Problem**

Yes : 1
No : 0

6. FEATURE EXTRACTION

In our dataset few features are not necessary , so those features are removed. The removed features are :

- BirthDay
- Education


These features removed because these are not related to Heart Disease.


7. CATEGORIZATION

Example : Cholesterol

```
totCholNormal = []
for i in totChol:
    if(i>240):
        x=2;
    elif(i>200 and i<240):
        x=1;
    elif(i<200):
        x=0;
    else:
        x=-1;
    totCholNormal.append(x)
```

This is how we categorize the dataset records like High , Medium , Low to the numerical value.

 PR1.shape

 (4240, 13)

So here we see, we are working with the dataset that has 4240 Records. Though this is not large dataset but not a small sample size either.

This is a sample of our dataset

Gender	Age	Smoker	BPMed	Stroke	Hypertensi	Diabetes	Cholestrol	Systolic	Diastolic	BP	Heartrate	HeartProblem
1	1	0	0	0	0	0	0	0	0	0	1	1
0	1	0	0	0	0	0	2	2	2	2	1	0
1	1	1	0	0	0	0	2	2	2	2	1	1
0	2	1	0	0	1	0	1	1	2	2	1	1
0	1	1	0	0	0	0	2	2	2	2	1	0
0	1	0	0	0	1	0	1	1	2	2	1	0
0	2	0	0	0	0	0	1	1	0	2	1	1
0	1	1	0	0	0	0	2	2	0	0	1	1
1	2	0	0	0	1	0	2	2	2	2	1	0
1	1	1	0	0	1	0	1	1	2	2	1	0

8. KNN (K-Nearest Neighbor Algorithm)

i. What is KNN ?

- K-Nearest Neighbor is very simple Machine Learning algorithm based on Supervised Learning technique.
- KNN can be used for Regression & as well as for Classifications.
- KNN involves pre-processing the dataset , training the model & testing the model.
- KNN is a non parametric algorithm , it means it does not make any assumption on underlying data
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset & at the time of classification it performs an action on the dataset.

ii. How KNN works ?

The KNN working can be explained on the basis of the below algorithm :

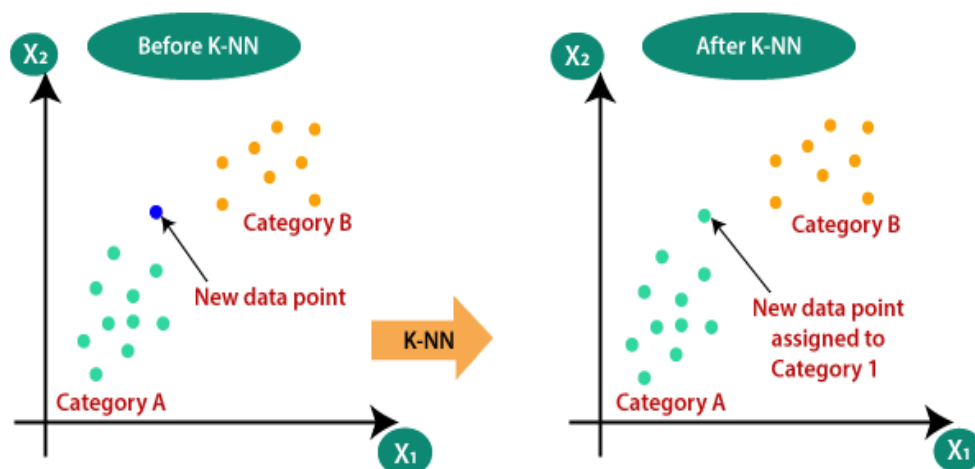
- **Step – I :** Select the number K of the neighbors
- **Step – II :** Calculate the Euclidean distance of K number of neighbors
- **Step – III :** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step – IV :** Among these k neighbors, count the number of data points in each category
- **Step – V :** Assign the new data points to that category for which the number of the neighbor is maximum
- **Step – VI :** Now our model is ready.

iii. What is value of K ?

- To avoid any confusion , it is always preferable to take the odd number as value of K
- But K value can't be too small. If the K value is very small (e.g 1) it can lead us to an outlier
- The high K value is also not desirable because it may not give the perfect result always. That's why K should be a single digit number.

Due to the above 3 rules we have taken the value of K is 5.

iv. Graphical Representation of K-NN



v. KNN (Using Libraries)

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import mean_absolute_error
```

We have used 70% of the dataset for training purpose & 30% of the dataset for testing purpose.

```
print(model1.score(test_X, test_Y))
```

```
0.9819182389937107
```

```
val_mae = mean_absolute_error(test_Y, pred1)
print("Mean Absolute Error is : ", val_mae)
```

```
Mean Absolute Error is : 0.018081761006289308
```

So , here we see our model is giving more than 98% accurate result. This is really Great accuracy.

vi. Mean Absolute Error :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$