ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

APPLICATION OF CLASSIFICATION TECHNIQUES TO DETECT HYPERTENSIVE HEART DISEASE

Tulasimala B. N*[1], Elakkiya S[2] & Keerthana N[3]
[*1]Assistant Professor, Department of MCA, Mount Carmel College, Autonomous, India
[2]Lecturer, Dept. of Computer Science, Mount Carmel College, Autonomous, India
[3]Student, BCA, Mount Carmel College, Autonomous, India

ABSTRACT

Heart diseases remain the main cause of death worldwide. Doctors generate data withhidden information present and it is quite a complicated process to effectively predict to make conclusions. Hypertension is the high blood pressure which is widely considered to be one of the most important risk factors for heart disease. Classification algorithm is computational based method used to verify data and predict whether a person is normal with healthy or unhealthy or with an average rate of heart functioning process. This paper describes the results obtained using computational tool used for predicting heart disease. The main intention is to acknowledge the cause for hypertensive heart disease from the results obtained through the computational tools used. Relevant primary data is collected. This data is normalized and classified using K Nearest Neighbor and Naive Bayes algorithms. Further results are compared, verified and concluded.

Keywords- *Classification algorithm, Hypertensive, K Nearest Neighbor, Naive Bayesian, Data normalization*

I. INTRODUCTION

Data Mining is a subject used widely to discoverpatterns and hence deduce knowledge from large amount of data. The applications of data mining are limitless. Application of data mining in the field of medical research is described here. Every hospital will have volumes of data related to patient obtained through diagnosis and other means. Presently huge medical data is available for analysis in the field of computational theory, which is fascinating the computer professionals to understand life better by analyzing the hidden information through computational techniques.

Data Mining is describes the method of prediction and descriptions of complex data. Prediction involves attributes or variables in the data set to find unknown or future state values of other attributes. Description emphasize on discovering patterns that explains the data to be interpreted by humans. Here proficient classification techniques is used to discover structure inside unstructured data through filtering the noise from the data set and discover patterns in apparently random data anduse all this information to better understand trends, patterns, correlations and ultimately predict rate of heart disease in a person. The information thus extracted will help anyone to take precautionary measures to maintain normal health.

The paper intends to predict the probability of getting heart disease given patient data set. The purpose of predictions helps discover trends in patient data in order to improve one’s health. Data generated by medical practitioners will have hidden information and it could be used in a better way for predictions with utmost certainty. For this purpose, the raw data is converted in to a dataset for modeling using classification techniques.

Hypertension is widely considered to be one of the most important risk factors for Heart disease. Hypertensive heart disease includes heart failure, thickening of the heart muscle, coronary artery disease and other conditions. Blood pressure should be checked at regular intervals. High blood pressure is often called the “silent killer” because it has no symptoms and can go undetected for years.

Here, for predicting and classifying from the hidden data with the given data set two algorithms are used

1. K Nearest Neighbor algorithm
2. Naive Bayesian algorithm

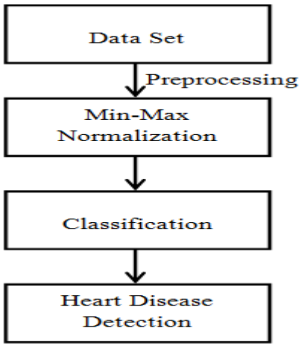


Figure 1: Overall Process of the system

II. CLASSIFICATION PROCESS

Primary data is described by the clinical pathologist for the threshold data which is collected and is as follows

i. **Age:**

Any Group

ii. **Sex:**

Male – 1
Female – 2

iii. **BP Level:**

Systolic Pressure - The top number recorded in a blood pressure reading. This amounts to functioning of arteries during contraction of your heart muscle

Diastolic Pressure - The bottom number recorded in a blood pressure reading. This is the blood pressure when heart muscle is between beats.

Table 1: Normal statistical data of systolic and diastolic pressure for various BP Category

Systolic (mmHg)	Diastolic (mmHg)	BP Category
Below 120	Below 80	Normal BP
120 – 139	80 – 89	Prehypertension
140 – 159	90 – 99	Stage 1 Hypertension
160 or higher	100 or higher	Stage 2 Hypertension

iv. **Cholesterol:**

It is a type of fat found in blood. Damages caused to liver leads to an increased blood cholesterol level. Body acquires cholesterol from various diets like meat, fish, eggs, butter, cheese, and whole or low-fat milk. Bad cholesterol contributes to plaque, a thick, hard deposit that clogs arteries. This can result in heart attack or stroke. Good cholesterol helps to remove bad cholesterol from the arteries.A healthy level of cholesterol may protect against heart attack.

Table 2: Normal statistical data of cholesterol levels

Children	Adult	Category
Below 170	Below 200	Good
170 – 199	200 – 239	Border Line
200 or higher	240 or higher	High

v. **Heart Rate:**

The number of heartbeats per unit of time usually counted per minute. The heart rate is based on the number of contractions of the ventricles.

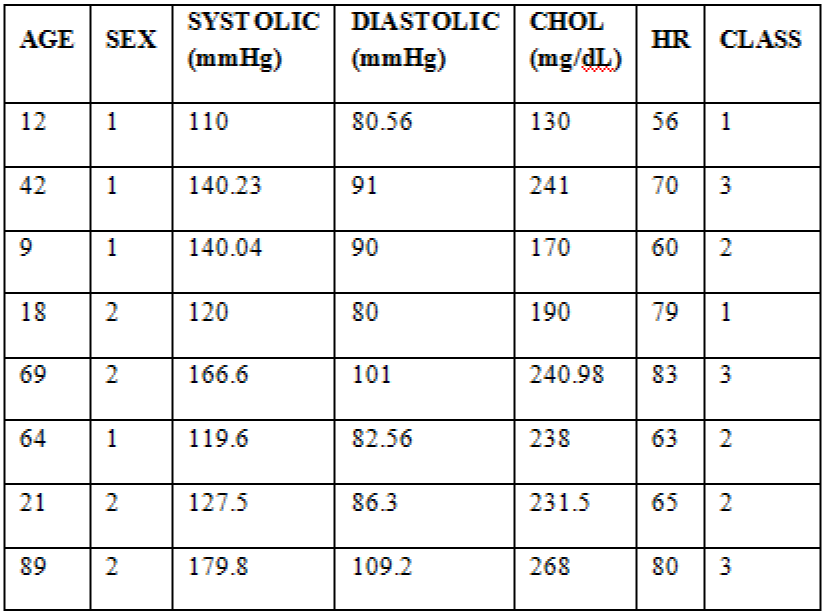
Table 3: The statistical data for heart rate with respect to different age group

Age	18-25	26-35	36-45	46-55	56-65	65+
Good	55-65	55-65	57-66	58-67	56-67	56-65
Average	60-81	66-81	67-82	68-83	68-81	66-79
Poor	82+	82+	83+	84+	82+	80+

vi. Class: Represents the code used to identify healthy, average and unhealthy category of people prone to heart disease Healthy – 1, Average – 2, Unhealthy – 3.

A set of questionnaire is prepared to collect the data that was distributed in area in northern Bengaluru. The queries were based on the information regarding the health conditions, frequent visit to the doctors, reasons for ailments and etc. Hundred different data from various backgrounds were collected. The required details for analysis is sampled and shown in the table 4.

Table 4: Description of sample data set



AGE	SEX	SYSTOLIC (mmHg)	DIASTOLIC (mmHg)	CHOL (mg/dL)	HR	CLASS
12	1	110	80.56	130	56	1
42	1	140.23	91	241	70	3
9	1	140.04	90	170	60	2
18	2	120	80	190	79	1
69	2	166.6	101	240.98	83	3
64	1	119.6	82.56	238	63	2
21	2	127.5	86.3	231.5	65	2
89	2	179.8	109.2	268	80	3

Data set shown in the table 4 is filtered, processed and classified using K-Nearest Neighbor and Naive Bayes algorithm and interpreted using KNIME tool.

III. K-NEAREST NEIGHBOR ALGORITHM

This is one of the classifier algorithm based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space.

When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance between two points,

$$X=(x_1,x_2,\ldots\ldots, x_n) \text{ and } Y=(y_1,y_2,\ldots.,y_n) \quad \text{is}$$

$$D(X, Y) = 2$$

$$D(X, Y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

The k-nearest neighbor algorithm is one of the simplest machine learning algorithms. In this an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. It is helpful to choose k to be an odd number as this avoids tied votes. As well as having a k value that is too small it is important to choose a value that isn't too large as it can also lead to misclassification.

IV. NAIVE BAYESIAN ALGORITHM

Naive Bayes builds and scores models extremely rapidly; it scales linearly in the number of predictors and rows. Naive Bayes makes predictions using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence.

Bayes' Theorem: The probability of event A occurring given that event B has occurred (P(A|B)) is proportional to the probability of event B occurring given that event A has occurred multiplied by the probability of event A occurring. Thus this is represented as (P (B|A) P (A)).
Hence P (A|B) = (P (B|A) P (A)) / P (B)

This algorithm makes the assumption that each attribute is conditionally independent of the others. That is, given a particular value of the target, the distribution of each predictor is independent of the other predictors. In practice, this assumption of independence in case when violated does not degrade the model's predictive accuracy significantly. Hence makes a difference between a fast, computationally feasible algorithm and an intractable one.

V. KNIME Tool

It is a graphical user interface that allows assembly of nodes for data preprocessing, modeling, data analysis and visualization. Here the data set comprises of 100 records attributed numerically.

Stages of processing
Initially the data is cleaned by filling the missing values. Data is transformed into a form appropriate for mining. Here aggregation operations are applied to the data. Attributes are constructed to help the mining process. This data are normalized so that it falls in a smaller range. Min-Max normalization is performed for transforming the original data using the formula.

Min-Max Normalization formula is given as follows:

$$V_{i\,=\,(}v_{i\text{-}}\mathrm{min}_{A/}\mathrm{max}_{A\text{-}}\mathrm{min}_{A)(}\mathrm{newmax}_A - \mathrm{newmin}_A) + \mathrm{newmin}_A$$

K-Nearest Neighbor

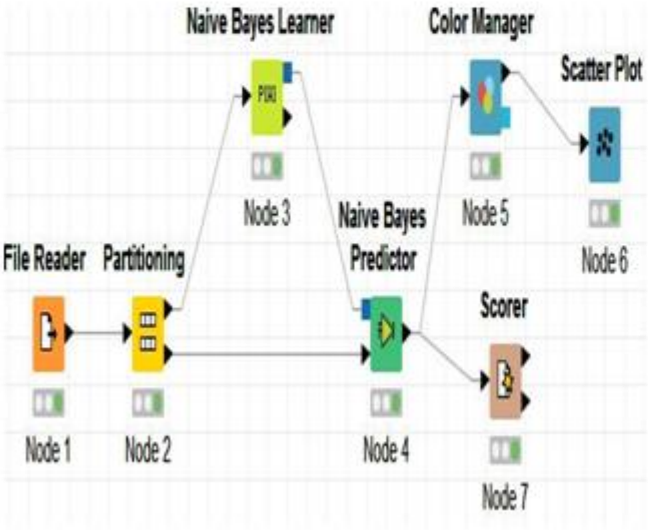


Figure 2: K-Nearest neighbor algorithm in KNIME

Naive Bayesian

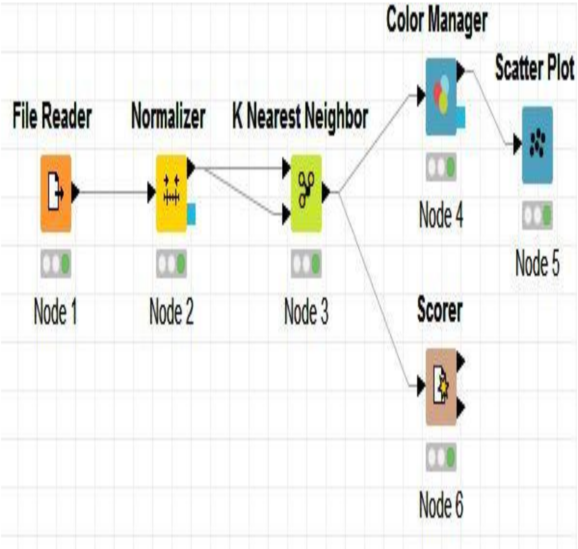


Figure 3: Naive Bayesian algorithm in KNIME

ISSN 2348 – 8034
Impact Factor- 4.022

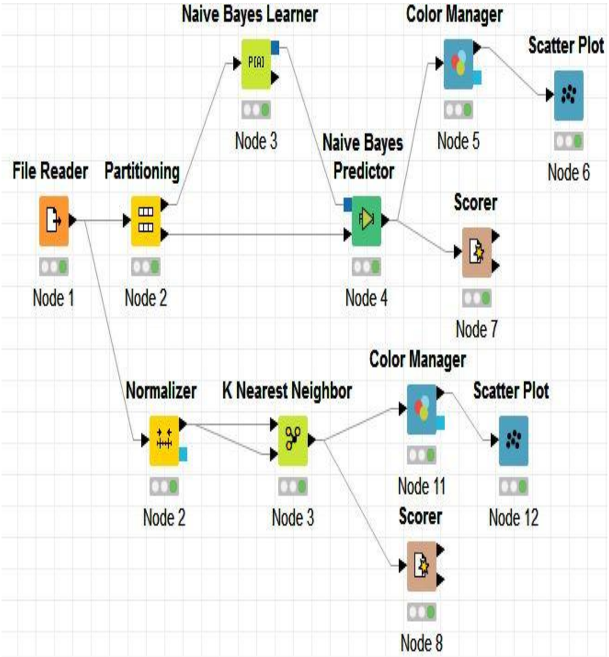


Figure 4: Combining K-Nearest neighbor and Naive Bayesian algorithms in KNIME

VI. REPORT

The results obtained from the two algorithms are shown in table 5 and table 6

Table 5: Outcome of K-nearest neighbor

Actual\Predicted	Healthy	Border	Unhealthy	
Healthy	32	0	0	T_A_H=32
Border	0	33	0	T_A_B=33
Unhealthy	0	1	34	T_A_U=35
	T_P_H = 32	T_P_B =34	T_P_U = 34	

Accuracy: 99.1% Error: 0.9%

Table 6: Outcome of Naive Bayesian

Actual\Predicted	Healthy	Border	Unhealthy	
Healthy	30	0	2	T_A_H=32
Border	0	28	6	T_A_B=34
Unhealthy	1	0	33	T_A_U=34
	T_P_H = 31	T_P_B = 28	T_P_U = 41	

Accuracy: 91% Error: 9%

The figure 5 represents the comparison of actual and predicted values of the end class. We can see that K-nearest neighbor predicted 99.1% of correct report whereas Naive Bayesian has lesser accuracy.

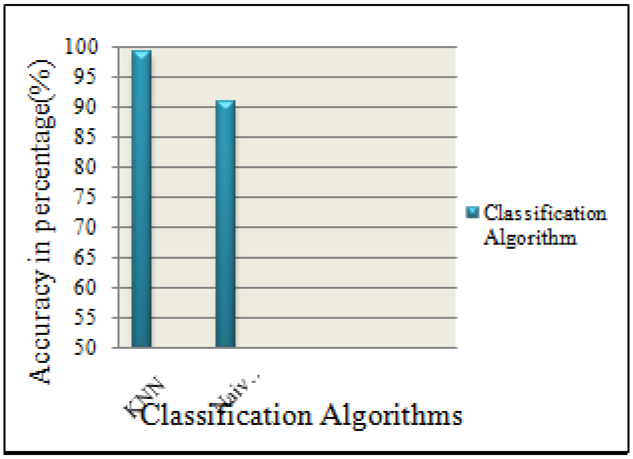


Figure 5: Graphical representation of accuracy of K-nearest neighbor and Naive Bayesian algorithm

VII. CONCLUSION

The application of data mining algorithms to detect hypertensive heart disease compares the output from the two algorithms and shows the best method of prediction. The predictive accuracy determined by both the algorithms suggests that K-nearest neighbor gives better result than Naive Bayes algorithm. This paper helps the health care professionals in the diagnosis of heart disease. It is made sure that parameters used are reliable indicators to predict the presence of heart disease. Using a conservative definition of essential hypertension (160 mm Hg systolic or 100 mm Hg diastolic) it is estimated that people are hypertensive. One can use this cost effective tool and obtain the results and there by precautionary measures can be taken to control on hypertension.

REFERENCES

1. *Jiawei Han, Jian Pei and Micheline Kamber, Data Mining Concepts and Techniques, University of Illinois at Urbana-Champaign, 2011*
2. *www.healthline.com*