International Journal of Computer Applications (0975 - 8887)
Volume 175 - No.14, August 2020

Web Support System for Prediction of Heart Disease using k-Nearest Neighbor Algorithm

Savitha Kamalapurkar
CSE, UVCE, Bangalore University
Bengaluru, Karnataka, India

Samyama Gunjal G.H.
CSE, UVCE, Bangalore University
Bengaluru, Karnataka, India

ABSTRACT

Cardiovascular diseases are a major threat to human life. The death rates are increasing day by day due to heart problems, such as cardiac arrest, clot or damage in blood vessels, arrhythmia, etc.. Most of the times heart related problems are emergency cases, hence there is a requirement of the on-line system in place to find the symptoms of heart problem and prevent it by detecting it in early stage. On-line system always helps the patients as well as doctor when there is an casualty. The proposed work supports health-care industry by providing a web based system for prediction of cardiac problems using Nearest Neighbors algorithm.

General Terms

Machine Learning, Heart Disease

Keywords

K-Nearest Neighbor, Machine Learning, Heart Disease, Prediction

1. INTRODUCTION

Cardiovascular diseases are becoming more in common in India. As per world health organization it is approximated that more than 12 million people die worldwide due to heart diseases every year. Various works is carried out in the health-care field using machine learning algorithms for predicting the heart diseases in the early stage. Existing systems does not provide user-friendly interface to be used by ordinary people. When heart Diseases are predicted using machine learning algorithms, it should be made available to common man as well as to doctors, only then data mining models fulfills the purpose. Data mining is used to find useful and valuable information in large datasets to turn raw data into meaningful and useful information. In the current work a system using K-Nearest Neighbor (KNN) algorithm is designed for prediction of heart disease as Nearest Neighbor algorithm is one of the top machine learning prediction algorithms used for disease predictions. KNN classification locates a collection of k samples from the training data which are very close to the record which needs to be classified, and based on majority voting by k nearest samples, label is given to test data.

User who uses the application needs to feed the information needed in the fields of the application which is implemented as Internet based application. It matches the parameters entered with the trained attributes of the model. This work can help doctors to take right decisions at right time and system is useful in case of emergency. In the designed system inputs given by user are passed to trained model and result predicted for test records are stored in catalog along with the test record for future report generation. Proposed work uses KNN algorithm for doing classification to detect cardiac problem. This approach predicts the test data as cardio positive or negative. Accuracy of the designed system always relies on k value chosen in the algorithm, dataset used and the number of attributes used in the dataset.

2. LITERATURE SURVEY

Summary of previous works done by different researchers is as shown in Table 1 .

Table 1. Literature Review

Year	Author	Algorithm	No of Attributes used	Accuracy
2017	Priyanka [1]	Naive Bayes	13	86
2013	M.Akhil jabbar [2]	KNN and GA	14	100
2016	S. Rajathi [3]	KNN+ ACO	15	68
2019	Divya Krishnani [4]	KNN, Random Forest(RF), Decision Tree(DT)	16	92
2018	Rathnayakc [5]	KNN + Naive Bayes + Decision Tree + Neural Network	15	75
2019	Buettner R [6]	Random Forest	13	84
2019	Mohan S [7]	Hybrid RF	13	88.7
2015	V. Krishnaiah [8]	Fuzzy KNN	13	80

Priyanka et al., proposed a Internet based technique for finding presence of cardiac problem using Weighted Naive Baye's method, here user needs to feed the required details to the application which are matched to the trained fields of the model built and accuracy of 86% achieved.

M. Akhil jabbar et al., proposes a different algorithm which combines KNN algorithm with genetic algorithm for effective classification. Genetic algorithms provide optimal solution by performing global search and experiments show that accuracy in diagnosis of heart disease can be improved by combining KNN with genetic algorithms [2].

Rajath et al., showed that when KNN is combined with ant colony optimization, improvement in the KNN algorithms effectiveness is seen, where the research is carried out in two levels, in first level KNN is implemented and in second level ACO is performed. Accuracy is increase by 2 percentage by merging ACO with KNN [3]

Divya Krishnani et al., predicted the endanger of cardiac problem by using RF, DT and KNN. And cross validation method is applied in the work and accuracy of 96.8%, 92.7%, and 92.89% respectively are obtained, when tested with dataset having more than 4000 samples [4].

Rathnayakc et al., observed better accuracy when more attributes and combinations of different machine learning techniques are used [5].

Buettner R et al., showed the possibility of presence of heart disease by using random forest algorithm. In this work cross validation is used which is nothing but, it divides the dataset into multiple datasets and each subset of dataset is used for both testing purpose and training purpose. By this he assured optimal result can be achieved for prediction [6].

Mohan S et al., introduced a hybrid model using machine learning algorithm like random forest and deep learning method like artificial neural networks. Further his model uses different combination of attributes of dataset. This model can be used for accurate diagnosis. The results of this model is proved to be very good in comparison with traditional methods [7].

Krishnaiah et al., emphasis that, conventional methods do not remove uncertainty in the data and makes an attempt to remove uncertainty in data by introducing the Fuzzy logic concept, creates membership function and include it with the calculated value to eliminate obscurity. Fuzzy K-NN classifier works well when classifiers of parametric techniques are used. System can be improved by raising the number of parameters in the proposed system [8].

3. PROPOSED WORK

In the proposed work online based heart disease prediction is developed using K-Nearest Neighbor algorithm as it is one of the top clanssificatio algorithms used in machine learning for disease predictions [13]. Heart disease can be predicted using some basic features like family history, lifestyle, body mass index, etc.. The common factors which are considered as most heartwarming [6,11] in cardiac disease are as in Table 2.

Table 2. Risk Factors

Sl No	Factor
1	High blood pressure
2	Intake of Alcohol
3	Cholesterol
4	Overweight
5	Physical fitness
6	Glucose level
7	Age of the person
8	Gender
9	Disease record in family

3.1 Dataset Description

The dataset used in the work holds the information of heart disease patients with 11 features and a target. Dataset contains 3 different types of attributes. They are:

Objective: These type of attributes are true data of patient.

Subjective: It is the information provided by the patient.

Test: The value for this type of attribute is obtained by medical examination;

The dataset contains data collected from 300 patients. The attributes used in the work are listed in Table 3. All of the values present in the dataset were collected while doing medical examination of the patients. Description of each attribute is as follows.

(1) **age:** Age of patient in years.

(2) **gender:** Two different numbers are used to represent male and female patients.
1 - Female
2 - Male

(3) **height:** Height of the person in cm, which gives the proportional body mass index.

(4) **weight:** Body mass of patient in kgs. This attribute plays major role as more weight is the cause of many diseases.

(5) **ap_lo:** It is diastolic blood pressure(DBP) which is nothing but force on arteries when blood is taken into heart from all the parts of the body. Its normal value is 80 mmHg or below.

(6) **ap_hi:** It is systolic blood pressure(SBP) which is nothing but force on arteries when blood is ejected from heart to all the parts of the body. Its normal value is 120 mmHg or below.

(7) **cholesterol :** Cholesterol is the main risk parameter to measure the risk of heart disease. There are two types of cholesterols: High Density Lipoprotein(HDL) and Low Density Lipoprotein(LDL) . Cholesterol of a person is measured by considering these two cholesterols, which has 3 categories:
3 - Major with above regular margin;
2 - Minor with regular margin; and
1 - Regular

(8) **gloc :** It is blood glucose level of the patient. High glucose is also one of major risk item of heart diseases. Glucose is measured by taking the reading before food and after food, and glucose level is measured by considering both readings, which has 3 categories:
3 - Major with above regular margin;
2 - Minor with regular margin; and
1 - Regular

(9) **smoke:** It gives information about lifestyle of the person as lifestyle also impacts the level of risk of heart disease, where
1 - Smoker
0 - Non- smoker

(10) **active:** It gives information about patient's lifestyle like he does regular physical exercise or not. This factor plays a role in heart disease prediction as regular excersise helps to keep the heart healthy
1 - Does physical exercise regularly
0 - No physical exercise regularly

(11) **cardio :** It is the target variable in the dataset.
1 - patient has heart disease
0 - patient does not have heart disease.

Figure. 1 shows the pictorial representation of presence of heart disease in male and female patients and Figure. 2 shows the

Table 3. Dataset Description

Sl No	Attribute	Value	Type of Feature
1	age	in years	Objective
2	gender	1 - women, 2 - men	Objective
3	height	Cm	Objective
4	weight	Kg	Objective
5	ap_lo	Diastolic blood pressure(DBP)	Test
6	ap_hi	Systolic blood pressure(SBP)	Test
7	cholesterol	3: well above regular margin 2: above regular margin 1: regular	Test
8	gluc	3: well above regular margin 2: above regular margin 1: regular	Test
9	smoke	whether patient smokes or not	Subjective
10	alco	Binary feature	Subjective
11	active	Binary feature	Subjective
12	cardio	Target variable	Presence or absence of cardiovascular disease

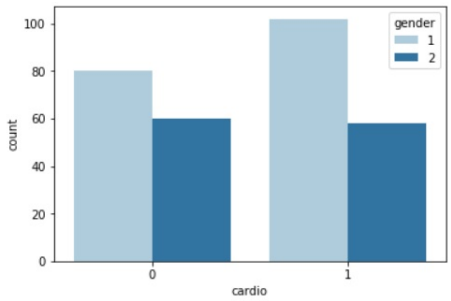

Fig. 1. Gender based pictorial representation of heart disease

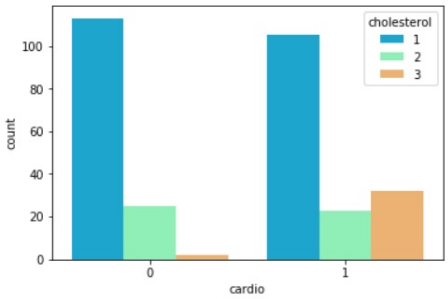

Fig. 2. Cholesterol based pictorial representation of heart disease

pictorial representation of presence of heart disease based on cholesterol in patient.

Different approaches are available in machine learning for doing prediction and classification [9, 10]. KNN is one among the simplest supervised prediction algorithms in machine learning. Any data to be classified is assigned with a label which is the majority vote given by its k neighbors. It uses different kinds of distance measures to find the distance between test instance and its neighbors.

3.2 Best way to fit the right value for k

Selecting the right value for k is very important as performance of the system depends on the k value. For any data, experiment needs to be done to see if it works well with low k value or bigger k value. Because, there is no assurance that it works same with all datasets, as few researchers have showed that KNN works well with lower k value, and few have proved that KNN works well with bigger k value. To find the correct k value, a patch of code is written and executed for different values of k i.e between 1 and 40. And comparison between error rate and predicted value is done. A graph is plotted for the same. Figure. 3 shows that error rate is less when k value is 7 or 13. Hence 7 is chosen as k value in the work.

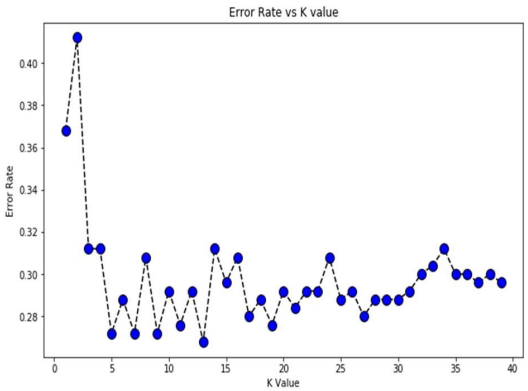

Fig. 3. Pictorial representation of Error rate Vs k value

Types of distances used in KNN to find nearest neighbors are Euclidean distance, Manhattan distance and Minkowski and the equations of each distance measure is as follows.

$$Euclidian = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2}$$

$$Manhattan = \sum_{i=1}^{k}|p_i - q_i|$$

$$Minkowski = \left[\sum_{i=1}^{k}(|q_i - q_i|)^r\right]^{1/r}$$

Where k is the number of dimensions, p is the record from dataset and q is the new record to be predicted and r is integer. Usually Euclidean distance is used in KNN to get the distance between test and training data records. In the proposed work Euclidean distance is used. Equation for finding distance (d) between two data points in the search space having coordinates (a1, b1) and (a2, b2) is given by

$$d = \sqrt{(a2 - a1)^2 + (b2 - b1)^2}$$

Algorithm used in KNN for classification is mainly based on k(Integer) value. It finds the distances between test and all training data points, sorts the records in ascending order of distances, and selects top k data points. Then gets the majority votes from these points and attaches label to the test record. Working of KNN algorithm is as shown in Algorithm 1.

Algorithm 1 KNN Algorithm Classification(U,V,t)

Begin
Input: U: Training data, V: class labels of U , and t: unknown sample
Output: Classification(U,V,t)
Load the training data U and test record t
do experiment and finalize the value for k
for every data point i in test data U:
  compute the distance d (Ui,t) from test record t to all training data points(using type of distance chosen)
  record all the distances in some data structure

end for
arrange the distance list in the ascending order
select the top k points from the sorted data structure
attach a label V to the test record t on the basis of higher number of votes present in k neighbors
End

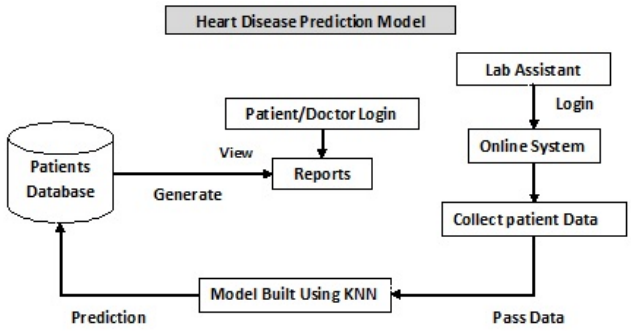

Fig. 4. System Architecture

3.3 System Architecture

Figure. 3.3 explains the whole architecture of the web support system built for predicting heart disease.
The proposed system is a web based application in which two tasks are performed.

(1) Prepare model using KNN for prediction

(2) Communicating the prediction result through online system to patient as well as to doctor

For building the model csv file having real data of patients is used. Correlation coefficient of the attributes in file are as shown in Table 4. It clearly shows that age, cholesterol, DBP, SBP, weight and glucose are impacting more in heart disease prediction. Required attributes are selected from file and then dataset is split into test and train subsets. Model is trained using KNN algorithm and when test patient's record is passed to the model, majority voting is done and final class for the test record is finalized. Flow of data in KNN model for prediction is as shown in Figure. 5. For doing prediction lab assistant logs into the system and enters patient details using user interface. Data entered is passed to the model built using KNN and predicted result is stored in the patient database along with the result. stored data is viewed in the form of report by patient or doctor.

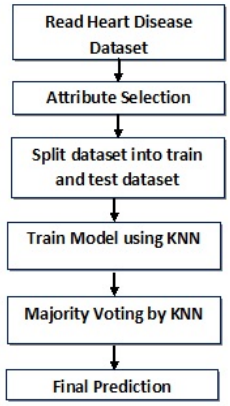

Fig. 5. Data flow in KNN model

Table 4. Correlation coefficient of attributes

Attribute	Value
age	0.238159
gender	0.008109
height	0.010821
weight	0.181660
ap_hi	0.054475
ap_lo	0.065719
cholesterol	0.221147
gluc	0.089307
smoke	0.015486
alco	0.007330
active	0.035653

4. RESULTS

Developed system is tested with actual data of patients through graphical user interface. When patient details are uploaded into the system, it finds existence of heart problem in the sufferers and prediction result of the patient along with his data is stored in the database. Same is displayed in the form of report to end user who can be either patient or doctor. Format of report generated is as shown in Figure.6.

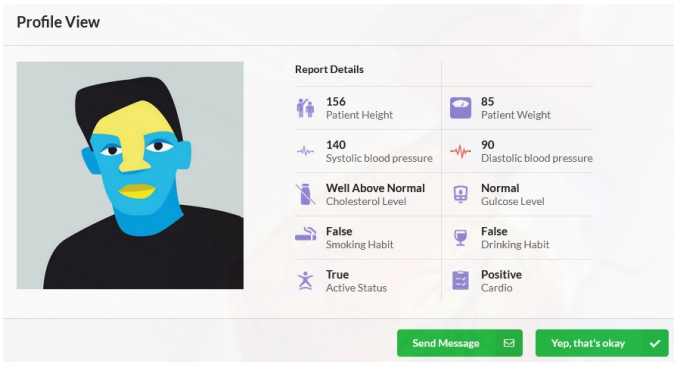

Fig. 6. Prediction result of test record

When model is trained with all the attributes present in the dataset, accuracy of the proposed model is 97% which is as shown in Figure. 7. Accuracy score changes as attributes are added or removed while training the model and accuracy also depends on k value.

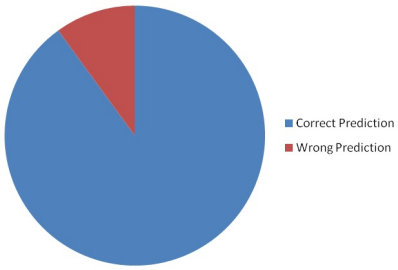

Fig. 7. Accuracy of proposed model

Again accuracy of the model is always dependent on the dataset used and k value chosen for the nearest neighbor algorithm. It is also studied that results are different when same algorithm is applied on another dataset from another source.
Figure.8 shows ROC curve (Receiver operating characteristic curve) of the model. In binary classification problems to visualize how well the model is performing ROC curves are plotted. In these type of curves X axis gives the false positive rate and Y axis gives the true positive rate of the model [12].

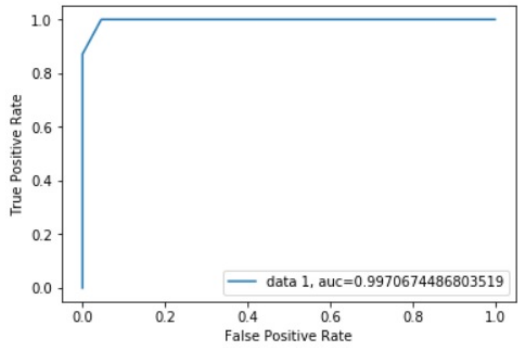

Fig. 8. ROC curve of model

Model is tested with 50 patient's data. Out of these 50 patients, 40 had heart disease and correctly classified as cardio positive. 5 patients had no heart problem and correctly classified as no cardio. 3 samples had no cardio but wrongly classified as cardio positive and 2 patients had cardio but wrongly classified with no cardio label. Same is summarized in Table 5. Precision, recall and F1-Score of the model is summarized in the Table 6.

Table 5. Prediction Classes

Actual Class	Predicted class (With cardio)	Predicted class (Without cardio)
With cardio	40	2
Samples having no cardio	3	5

Table 6. Accuracy of the system

	Precision	Recall	F1- Score
Without cardio	1.00	0.95	0.98
With cardio	0.94	1.00	0.97

5. CONCLUSION

The proposed work builds a web application for finding the heart problem using supervised machine learning algorithm KNN. It helps the patient in case of emergency. Doctor can check the details of patients through online portal and diagnose remotely in case of unavailability. Application can be enhanced by combining KNN with other prediction algorithms which are established to be very accurate. And application can be used to predict different type of health illness like cancer, diabetes etc. Research work can be enhanced by testing it with other real datasets. The accuracy of the system developed rely on the dataset, its attributes and the k value chosen.

6. REFERENCES

[1] Priyanga, Dr. Naveen , " Web Analytics Support System for Prediction of Heart Disease Using Naive Bayes Weighted Approach (NBwa)," IEEE 2017 Asia Modeling Symposium DOI 10.1109/AMS.2017.12

[2] M.Akhil jabbar, " Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm," International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013

[3] Rajathi, S Radhamani, G," Prediction and analysis of Rheumatic heart disease using KNN classification with ACO," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).doi:10.1109/sapience.2016.7684132

[4] Divya Krishnani, " Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)

[5] Rathnayakc, B. S. S., and Ganegoda,G. U," Heart Diseases Prediction with Data Mining and Neural Network Techniques," 2018 3rd International Conference for Convergence in Technology (I2CT).

[6] Buettner, R., Schunter, M," Efficient machine learning based detection of heart disease," In IEEE Healthcom Proceedings: IEEE International Conference on E-health Networking

[7] Mohan S., Thirumalai C., & Srivastava G. (2019)," Effective Heart Disease Prediction using Hybrid Machine Learning Techniques," IEEE Access, 1?1. doi:10.1109.access.2019.2923707

[8] Krishnaiah V. , Narsimha G., & Chandra N. S. "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach," Emerging ICT for Bridging the Future - Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1, 371?384. 2015

[9] Sajetha T, Samyama Gunjal G. H. An Approach to Face Recognition Using Feed Forward Neural Network, International Journal of Computer Applications Technology and Research (IJCATR), Volume - 6, Issue - 4, Pages : 172-212, Feb-2017.

[10] Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar, "Air Pollution Prediction Using Machine Learning Supervised Learning Approach," International Journal

Of Scientific & Technology Research (IJSTR), Volume 9, Issue 04, April 2020, ISSN: 2277-8616118

[11] F BrainBoudi, "Risk Factors for Coronary Artery Disease,2016,"https://emedicine.medscape.com/article/164163 overview.

[12] Hajian-Tilaki, Karimollah," Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," in Caspian Journal of Internal Medicine Vol.4, pp. 627-635, 2013.

[13] Wu, Xindong & Kumar, Vipin & Quinlan, Ross & Ghosh, Joydeep & Yang, Qiang & Motoda. Top 10 algorithms in data mining. Knowledge and Information Systems. 14. 10.1007/s10115-007-0114-2.