International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Machine Learning Techniques: Performance Analysis for Prevalence of Heart Disease Prediction

Sachin Kamley, R.S. Thakur

Abstract: ***In these days, heart disease has become most dominating problem for medical professionals as well in India and abroad. However, heart disease is a major factor for behind the most of the people deaths today. An efficient and effective machine learning technique is required in order to reduce large scale of deaths due to this problem. In this direction, data mining and machine learning techniques play prominent role for pre-stage detection from heart disease problem. This study focuses on three most important machine learning techniques Support Vector Machine (SVM), Naive Bays (NB) and K-Nearest Neighbor (K-NN) for heart disease prediction. The machine learning tool Statistica is used for result generation purpose. Finally, experimental results stated that SVM method has excellent accuracy (86.12%) over other methods.***

Index Terms: Data Mining, Heart Disease, Machine Learning, Naïve Bays, Support Vector Machine, Prediction, Statistica.

I. INTRODUCTION

Due to the changing lifestyle, most of the people are suffering from heart disease problem increasing rapidly day by day. In 2008, the World Health Organization (WHO) had revealed the figures that 17.3 million people (estimated) were died globally due to Cardiovascular disease (World Bank, 2008) [1]. Therefore, lives of millions of people can only be saved by precise diagnosis at early stage and appropriate treatment might be provided to the patients.

But, at primary stage diagnosis of heart disease problem is very challenging task because heart disease problem depends on several factors [2] [3]. Thus, addressing these challenges and issues, healthcare industry is needed to develop such kind of medical diagnosis system in order to help medical stockholders to improve diagnostic process or decision quality.

Meanwhile, Data mining process has mystery to extract meaningful information from large amount of database [4]. Previously, various data mining software's and techniques are utilized for medical diagnosis purposes. However, these techniques play very crucial role for identification of disease as well as improving the diagnosis quality [5].

Machine Learning (ML) is another sub field of computer science which helps to gain insight from huge amount of database [6]. ML techniques have already shown their usefulness in healthcare or medical industry. In the past, various ML techniques like Decision Tree (DT), Naïve Bays (NB), K-Nearest Neighbor (KNN) etc. are used for diagnosing purpose [7] [8].

This study presents a comparative study about machine learning classification techniques like SVM, KNN and NB on heart disease dataset. The remaining sections are structured as follows:

Brief literature review of significant researchers is described in section 2. Proposed methodologies are described in section 3. Experimental results are described in Section 4 and at last conclusion and future scopes of the study are described in section 5.

II. LITERATURE REVIEW

This study presents some significant researchers work in brief.

Gudadhe et al. (2010) [9] have designed machine learning model based on Multilayer Perceptron (MLP) network and SVM approach. The experimental results stated that overall accuracy is achieved by model is 80.41%.

Nahar et al. (2013) [10] have applied Apriori algorithm for rule generation from heart disease dataset. However, the dataset is classified in two parts i.e. men and women.

Ishtake and Sanap (2013) [11] have recorded the performance of Naïve Bays, Decision Tree (DT) and Neural Network (NN) for heart disease dataset. However, the Naïve Bays method performs outstanding than other approaches for short data size.

Dey and Rautary (2014) [12] have compared the performance of MLP and Naïve Bays for healthcare dataset. Finally, experimental results showed that Naïve Bays had highest correctly classified instances.

Chaki et al. (2015) [13] have used supervised machine learning techniques like Naive Bays, C4.5 decision tree and SVM for classification purpose. Finally, experimental results showed that SVM method performs outstanding and has highest no of correctly classified instances than others.

Khanna et al. (2015) [14] have conducted the performance of three different classification techniques like Neural Network (NN), Logistic Regression (LR) and SVM (Linear Kernel) respectively. The dataset obtained from UCI repository and consists of 303 instances with 14 attributes. However, the experiential results stated that SVM method with linear kernel type had outstanding performance than others.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Sachin Kamley, Department. of Computer Applications, S.A.T.I,. Vidisha (M.P.), India.
R.S. Thakur, Department of Computer Applications, M.A.N.I.T., Bhopal (M.P.), India.

Sen (2017) [15] has designed machine learning model for prediction and diagnosis for heart disease dataset. The dataset downloaded from UCI repository and contain 303 samples. However, the dataset consists of 14 parameters where 13 input parameters and 1 output parameter. He has used four different machine learning techniques like Decision Tree, Naive Bays, K-Nearest Neighbor and SVM for this purpose. Finally, results are obtained using WEKA data mining tool and experimental results stated that SVM method has highest correctly classified rate (84.15%) and incorrectly classified rate (15.84%) respectively.

Ravathi and Kavitha (2017) [16] have downloaded heart disease dataset from UCI repository. However, the techniques used like Naive Bays, K-Nearest Neighbor (KNN) and Random Forest for analyzing the data set. They have considered 270 instances and WEKA data mining tool to obtain the results. Finally, Naive Bays method gives better performance than others.

Rabbi et al. (2018) [17] have analyzed heart disease dataset using various classification techniques. Therefore, UCI machine learning repository is used to obtain the dataset. They have used three different classification techniques like Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN) respectively. Their experimental results stated that SVM method had highest classification accuracy (85.18%) than others.

Finally, based on significant researches work this study is inspired by SVM, KNN and Naive Bays methods.

III. Proposed Methodology

A. Dataset Description

The Kaggle data repository is used to obtain the dataset in this study [18]. However, the dataset contain 209 instances and 8 attributes of healthy person. Table 1 describes these attributes in brief.

Table I. Description of Heart Disease Dataset [18]

S. No.	Name	Type	Description
1	Age	Continuous	Age in Years
2	Sex	Discrete	{Male, Female}
3	BPL	Continuous	Blood Pressure Level (in mm Hg)
4	FBS	Discrete	Fasting Blood Sugar >120 mg/dl
5	MHR	Continuous	Maximum Heart Rate
6	Chol	Continuous	Serum Cholesterol in mg/dl
7	Old-Peak	Continuous	Depression Induced by Exercise Relative to Rest
8	Disease	Discrete	Having Disease in Yes or No

Fig. 1 shows the flowchart of the proposed methodologies.

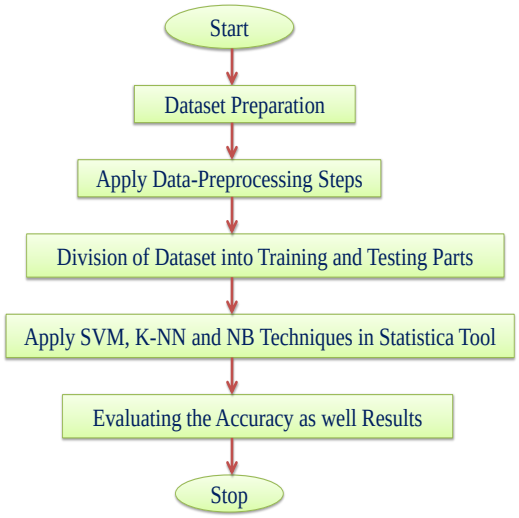

Fig. 1. Flowchart of the Proposed Methodologies [2] [3] [6].

B. Machine Learning Techniques

Machine learning techniques have extensively used in diagnosing of heart disease prediction. In this study, three important and most popular machine learning techniques are discussed for the purpose of comparative analysis. These techniques are Naïve Bays (NB), Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN).

Support Vector Machine (SVM): SVM model was first proposed by Cortel and Vapnik in 1995 [6] [8]. However, it is basically used for classification and regression analysis purpose and it is based on the statistical learning concept. In multidimensional environment, SVM model creates optimum boundary that is used to distinguish data points into different classes [6] [8] [19]. However, the optimum boundary is known as hyper plane. Thus, SVM is a classification technique which distinguishes the various classes of data with the use of a hyper plane. Generally, SVM model is prepared with training data and accuracy of the model is checked with test data with respect to hyper plane. Moreover, the SVM model tries to find the space in matrix of data where different classes of data can be widely differentiated and draws a hyper plane. Fig. 2 shows simple representation of SVM model.

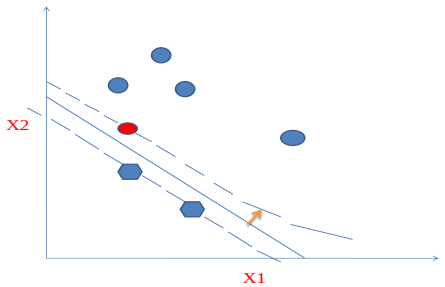

Fig. 2. Representation of SVM Model [6] [8].

Fig. 2 shows the basic representation of SVM model, where optimum hyper plane always maximize the margin between the classifier.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Naive Bays (NB): NB or NB classifier is a simple machine learning technique which uses the Bays theorem to classify the data. However, the theorem based on the assumption that attributes value on a given class is totally independent of another attribute [6] [20]. This assumption is also called conditional independence. However, this made the commutation process easy involved in. so it is called Naïve.
Bays theorem is an important theorem which is used to calculate conditional probability [21]. Bays theorem states that for two events H and E with the probability P (E)>0 (probability of a statement is always greater than 0 and less than unity), the conditional probability of event H given that, evidence E has occurred is defined by equation (1).

$$P(H/E) = \frac{P(H\&E)}{P(E)} \quad (1)$$

Similarly, the conditional probability of evidence E given that hypothesis H occurred can likewise be shown by equation (2) .

$$P(H/E) = \frac{P(E/H_i)*P(H_i)}{\sum_{i=1}^{n} P(E/H_i)*P(H_i)} \quad (2)$$

Now Bays theorem formula can be given by equation (3) [8].

$$P(H/E) = \frac{P(E/H).P(H)}{P(E)} \quad (3)$$

K-Nearest Neighbor (K-NN): K-NN is one of the most popular and simplest among all machine learning techniques. Therefore, the K-NN algorithm classifies an instances based on given training set and predicts any new instance based on majority vote of its closest neighbors [2] [3] . Simply, objects are classified based on some distance measure. It comes under the lazy learning category because function approximation done at local level and all computation is deferred until classification [22].
In this method, distance calculation between two data points or objects, Euclidean distance formula is used respectively. The equation (4) shows Euclidean distance formula [6] [7].

$$D = \sqrt{\sum_{j=1}^{n} (X_j - \overline{X_j})} \quad (4)$$

IV. EXPERIMENTAL RESULTS

In this section, different machine learning techniques are experimented using heart disease dataset. The Statistica machine learning software is used to generate the results. However, the experimental results are based on the performance and efficiency of the algorithms i.e. which algorithm gives the best performance. Table II shows sample of performance comparison of actual and predicted disease.

Table II. Sample of Performance Comparison of Actual Vs Predicted Disease

S.No.	Actual Disease	Predicted Disease	Accuracy
1	positive	positive	Correct
2	negative	positive	Incorrect
3	negative	negative	Correct
4	negative	negative	Correct
5	positive	negative	Incorrect
6	positive	positive	Correct
7	negative	negative	Correct
8	positive	negative	Incorrect
9	negative	negative	Correct
10	positive	negative	Incorrect
11	negative	negative	Correct
12	positive	negative	Incorrect
13	positive	positive	Correct
14	positive	positive	Correct
15	negative	negative	Correct
16	positive	negative	Incorrect
17	positive	negative	Incorrect
18	positive	negative	Incorrect
19	positive	positive	Correct
20	negative	negative	Correct
21	positive	negative	Incorrect
22	positive	positive	Correct
23	negative	negative	Correct
24	positive	negative	Incorrect
25	positive	positive	Correct
26	negative	negative	Correct
27	negative	negative	Correct
28	negative	negative	Correct
29	negative	negative	Correct
30	negative	negative	Correct

Table III shows confusion matrices of the proposed classifiers.

Table III. Confusion Matrices of the Proposed Classifiers

Classifier	Confusion Matrix	
SVM	69	14
	7	43
K-NN (K=1)	41	19
	13	31
NB	58	11
	10	34

Table IV shows the performance comparison of proposed algorithms.

Table IV. Performance Comparison of proposed algorithms

Method	Classified	Misclssified	Accuracy (%)
Naïve Bayes	165	44	78.94%
K-NN	155	54	74.16%
SVM	180	29	86.12%

Fig. 3 shows the accuracy comparasion of proposed algorithms.

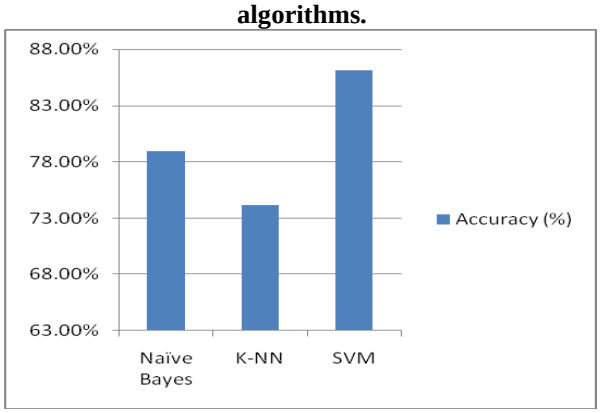

Fig. 3. Accuracy Comparasion Of Proposed Algorithms.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

It is clearly shown by results (based on Table II, III and Fig. 3) that SVM method has highest classified rate and accuracy i.e. 86.12% respectively.

V. CONCLUSION AND FUTURE SCOPES

In this study, performance of different machine learning techniques like K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) and Naïve Bays (NB) are recorded for the heart disease dataset. As the results suggested that for small dataset size, SVM method outperformed than other techniques. The benefit of this study is that it will be helpful for medical professionals for diagnosing the heart problems. The major drawback of this study is also noticed that accuracy of proposed methods is below 90%.

In near future, rich dataset and ensemble machine learning techniques will be adopted for accuracy improvement as well as performance comparison.

REFERENCES

1. World Bank Health Statistics and Information Systems Retrieved from http://www.who.int/healthinfo/global_burden_disease/estimates.
2. S. Chaitrali, D. Sulabha and S. Apte (2012). Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques. International Journal of Computer Application, Vol. 47, pp. 44-48.
3. J. Nahar (2013). Computational Intelligence for Heart Disease Diagnosis: A Medical Knowledge Driven Approach. Expert Systems with Applications, Vol. 2, pp. 96-104.
4. P. Domingos and M. Pazzani, "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", In Proceedings of the 13th Conference on Machine Learning, pp. 105-112, Bari, Italy, 1996.
5. C. Kalaiselvi and G.M. Nasira (2015). Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques. Indian Journal of Science and Technology, Vol.8 (14), pp. 21-30.
6. E. Alapydin, "Introduction to Machine Learning", 2nd ed. Cambridge Massachusetts, MIT Press: 2010.
7. G. James. Introduction to Statistical Learning", New York, Springer: 2013.
8. V.N. Vapnik. The Nature of statistical Learning Theory", New York, Springer: 1995.
9. M. Gudadhe, K. Wankhede and S. Dongre (2010). Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network. In Proceedings of the International Conference on Computer and Communication Technology (ICCCT), pp. 741-745.
10. J. Nahar, T. Imama, K.S. Tickle and Y.P. Chen (2013). Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females. Elsevier, Vol. 40, pp. 1086-1093.
11. Ms. Ishtake and S.H. Sanap (2013). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Journal of Healthcare & Biomedical Research, Vol. 1(3), pp. 94-101.
12. M. Dey and S.S. Rautaray (2014). Study and Analysis of Data mining Algorithms for Healthcare Decision Support System. International Journal of Computer Science and Information Technologies. Vol. 5(1), pp. 470-477.
13. D. Chaki, A. Das and M.I. Zaber (2015). A Comparison of Three Discrete Methods for Classification of Heart Disease Data. Bangladesh J. Sci. Ind. Res. Vol. 50 (4), pp. 293-296.
14. D.Khanna, R. Sahu, V. Baths and B. Deshpande (2015). Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. International Journal of Machine Learning and Computing. Vol. 5(5), pp. 414-419.
15. S.K. Sen (2017). Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering and Computer Science, Vol. 6(6), pp. 21623-21631.
16. K.K. Revathi and K.K. Kavitha (2017). Comparison of Classification Techniques on Heart Disease Dataset. International Journal of Advanced Research in Computer Science, Vol. 8(9), pp.276-280.
17. M.F. Rabbi, M.P. Uddin, M.A. Ali, M.F. Kibria, M.I. Afjal , M.S. Islam and A.M. Nitu (2018). Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction. American Journal of Engineering Research (AJER). Vol. 7(2), pp. 278-283.
18. Heart Disease Dataset Downloaded from Data source "http://www.kaggle.com" on Date 4/12/2018.
19. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, Vol. 15, pp. 104-116.
20. N. A. Sundar, P. P. Latha, and M. R. Chandra (2012). Performance Analysis of Classification Data Mining Techniques over Heart Disease Data Base. International Journal of Engineering Science & Advanced Technology, Vol. 2(3), pp. 470– 478.
21. S.A. Pattekari and Parveen Asma (2012). Prediction System for Heart Disease Using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, Vol. 3(3), pp. 290-294.
22. J. Liu, Y. T. HSU and C. L. Hung (2012). Development of Evolutionary Data Mining Algorithms and Their Applications to Cardiac Disease Diagnosis. In WCCI IEEE World Congress on Computational Intelligence, pp. 10–15.

AUTHORS PROFILE

Sachin Kamley completed his Masters degree from S.A.T.I., Vidisha in 2006 affiliated to Rajiv Gandhi Technological University, Bhopal (M.P.). He is working as an assistant professor in Department of Computer Applications since May 2007 at Samrat Ashok Technological Institute (S.A.T.I), Vidisha and also completed Ph. D from Barkatullah University, Bhopal in the year 2015. He has published more than 18 papers in International journals and conferences and attended many workshops of National repute.

Ramjeevan Singh Thakur is working as an Associate Professor in the Department of Computer Applications at Maulana Azad National Institute of Technology, Bhopal, since 2010. He has guided several Ph.D. Research Scholars and handling various Government Research Projects of about Rs. One Crore. He has published more than 95 Research Paper in National, International Journals and Conferences. He has visited several countries like USA, Hong Kong, Iran, Thiland, Malaysia, and Singapore.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)