**Machine Learning Approach for Heart Disease Prediction**

Naeeb Md Ridwan1,FuadTanvir1

BRAC University, Computer Science And Engineering Department

66 Mohakhali, Dhaka,Bangladesh

{md.naeeb, fuad.tanvir}@gmail.com

**Abstract.** Heart disease is one of the most common and catastrophic health conditions which causes major damages to the health of a number of people all over the world. But the damage, caused by this fatal disease, can be prevented if there is any prognosis done to make the patient more concern about his/her health condition. Prognosis means pre-diagnosis and by the implementation of machine learning, it can be done. This paper demonstrates the implementation of a SVM algorithm (linear SVC) to determine the accuracy in terms of predicting heart disease possibility of a patient. The accuracy was measured 0.616 on the scale of 1 while we have learnt the implementation of such algorithm for this specific task.

**Keywords**-Machine Learning, Artificial Intelligence, Heart Disease Prediction, SVM Algorithm, Prognosis

**1 Introduction**

According to World Health Organization (WHO), cardiovascular disease or heart disease is the number one cause of death globally. Around 17.7 million people died due to health disease in 2015 all over the world, which holds about 31% of overall deaths in that year. One of the major factors that aggregates the overall situation is that in most cases heart disease are diagnosed following a severe damage caused by heart attack or stroke. In this circumstance, a prognosis process can be very much helpful to reduce the damage to a great extent. Prognosis is a medical term which basically means to find out the likely course of a medical condition. To be more precise, prognosis provides a forecast of the likely outcome of the situation. But existing prognosis and diagnosis processes are often costly and out of the reach of the middle class to poor people. However, development of an AI system with the implementation of machine learning can be a solution in this context. In this project, our goal is to develop a system with the implementation of machine learning model and algorithm that can predict the likeliness of heart disease of a person based on the given data about him/her. For this, we have used UCI data repository for Machine Learning for Heart Disease diagnosis. The dataset is Cleveland Clinic Foundation (Cleveland data) and the algorithm is linear SVC, a classification of SVM algorithm.

Implementation of AI or machine learning to predict heart disease is not new. However, these systems have not been developed as a software which can be accessible from any device. We will hopefully extend this project in near future and will try to come up with a software solution.

**2. Methods and Approach**

Methods and processes used in this system have been mentioned below:

**Dimensionality Reduction Using PCA Method:**

PCA is a method which is used to convert the data from high dimensional space to a lower dimensional space. It is a statistical method in which the observed raw data set, which might be correlated in nature, is converted into linearly uncorrelated variables.

Centering the data and performing an SVD on it is the process of computing PCA. Here first few columns of the V matrix produced by SVD as the principal components. The equation that is used in this process is:

X=U^VT

Here, X is the data matrix, is the diagonal eigenvalue matrix and U and V are unitary matrices.

**Linear SVM:**

Support Vector Machines (SVM) is an algorithm which is widely used as a machine learning algorithm to solve multiclass classification problem. Set of examples are given in one or two classes and the SVM algorithm builds model that assign each new example of the dataset to one of those classes.

**RBF kernel SVM:**

RBF kernel is used in SVM for non-linearly separable problem. It is a non-linear kernel function as no hyperplane is sufficient enough to accurately classify data in this case.

**Work Procedure**

Our target was to test different classifier and compare which one is working better in terms of heart disease prediction. We have compared Linear SVM and RBF Non-linear SVM on the given vector representation of Cleveland dataset. For the purpose of experiment, we run the classifiers for 80/20 splits while we used 80% for training our classifiers and 20% for testing their predictions.

With the presence of 303 instances and 14 attributes, we have implemented Principal Component Analysis (PCA) on the original X value where X is the feature set. Our feature set is then reduced to X\_new which is a vector representation of 303 samples. Then we tried 80/20 split of X\_new where 80% of X\_new is used to train the SVM classifier and 20% of X\_new is used to test.

**3. Results**

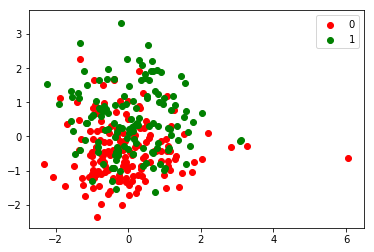


Fig 1: Outcome of the prediction. 1 indicates presence of heart disease while 0 nullifies the existence of it.

Based on the given data we predicted whether symptoms of heart disease is present or not. The outcome is represented with the value 1-presence and 0-absence. However, our aim was to classify the existence of disease in four values from 1 to 4. But we have simplified the process and reduced the outcome range to 0 and 1 for now. The accuracy was measured 0.61 on the scale of 1.

**4. Discussion**

**Previous Work**

Earlier a few approaches were made in the field of AI or machine learning to predict heart disease is not new. Chaurasia and Pal [1] did applied data mining techniques to predict heart disease at an early level. Hazra, Mandal, Gupta, Mukherjee and A. Mukherjee [2] also implemented machine learning and data mining processes to diagnose heart disease.

**Generalization of the Result**

In a generalize form, the result can be classified into two types, represented by 0 and 1. Analyzing the presence of various factors such as- blood pressure, cholesterol level, heart bit and so on the result has been derived in the form of presence or absence of the disease.

**Scope of Development**

The accuracy is not which is not up to the mark of our target. That happened mainly because the data given here are loosely bounded even though they are coo-related. However, modification of this algorithm is needed to increase the accuracy level or different algorithm is need to be implemented for better outcome. Overall, this project can be a baseline to go for further betterment to make an expert system ultimately.

**Significance of the Result**

The output derived from the experiment clearly states that based on the given sets of data decision can be made applying machine learning approach. Therefore, the result of this experiment is important to form an efficient algorithm which can predict the possibility of something depending on the given data.

**5. Conclusion**

It is quite evident that input data plays an integral role in machine learning techniques and based on that reliable outcome can be produced. We have now labels from 0 and 1 where the labels can tell us about the existence of heart disease in a human body. Based on this, further multiclass labels can be derived to be more specific about the likeliness of this disease. Implementation of PCA has again proven that we can get rid of similar feature set and still obtain predictions with great efficiency. The experiment helped us in predicting the outcome but also gave us valuable insights about the nature of data. The knowledge and experience, gained through this project, will enable us to go for further development and also to reach towards our main goal which is to build an expert software system for predicting the existence of heart disease.

**References**

1. V. Chaurasia and S. Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib. J. SciTech, vol. 1, pp. 208-2017, 2013.

2. A. Hazra, S. K. Mandal, Gupta, Mukherjee, A.Gupta, A. Mukherjee and A. Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques”, Jvol.10, pp.2137-2159, 2017.

3. P. V. Ankur Makwana, “Identify the patients at high risk of re-admissionin hospital in the next year,” International Journal of Science andResearch, vol. 4, pp. 2431–2434, 2015.

4. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare", Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241-266, 2013.

5. M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, N. and W. Treis-ter. "Transforming Health Care Through Big Data" in, Washington DC, USA:Institute for Health Technology Transformation, 2013.

6. H. C. Koh, G. Tan. "Data Mining Application in Healthcare", Journal of Health care Information Management, vol. 19, no. 2, 2015.

7. C. S. Dangare and S. S. Apte, “Improved study of heart disease predictionsystem using data mining classification techniques,” InternationalJournal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.

8. Y. E. Shao, C.-D. Hou, and C.-C. Chiu, “Hybrid intelligent modelling schemes for heart disease classification,” Applied Soft Computing,vol. 14, pp. 47–52, 2014.

9. Y. Xing, J. Wang, Z. Zhao, and Y. Gao,“Combination data miningmethods with new medical data to predicting outcome of coronary heartdisease,” pp. 868–872, 2007.

10. Y. E. Shao, C.-D. Hou, and C.-C. Chiu, “Hybrid intelligent modelling, schemes for heart disease classification,” Applied Soft Computing,vol. 14, pp. 47–52, 2014.