Detection of Chronic Disease using Machine Learning Techniques

[1]Amruta H Raut, [2]Tejashree H Jadhav, [3]Yogita G Borole, [4]Dr M A Pradhan

[1]Student, [2]Student, [3]Student, [4]Associate Professor
Department of Computer Engineering
AISSMS College of Engineering, Kennedy Road, Pune-411001, India

***Abstract*: According to World Health Organization (WHO), chronic diseases are the leading cause of death worldwide. With the improvement of living standards, the incidences of chronic disease is increasing. Hence it is important to perform risk assessments for chronic diseases. Since there is enormous amount of data available from medical industry, it could be useful for discovering hidden patterns and developing a classifier using existing machine learning techniques. Accurate analysis of medical data benefits early disease detection and patient care which in turn benefits the society. Pre-processing of data plays a significant role for enhancing accuracy of classification systems. In this paper, we use machine learning algorithms for effective prediction of chronic disease. We first did the comparative study of different machine learning algorithm then we propose a new framework based on two machine learning algorithms: Naive Bayes Algorithm and Decision Tree Algorithm. We apply both these algorithms on the dataset and predict the outcome. We also compare the results obtained by applying these algorithms to conclude which of the algorithm is better and more accurate.**

***Index Terms:* Machine Learning; Chronic Disease; Naive Bayes Algorithm; Decision Tree**

I. INTRODUCTION

Managing the increasing burden of chronic diseases is a major health problem worldwide. Some of the leading chronic diseases include diabetes, strokes, cardiovascular disease, arthritis, cancer, hepatitis. It is important to diagnose chronic disease in its early stage and help spread awareness about it worldwide. There is strong evidence that delayed diagnosis and management are frequent and often adversely affect patients and society. Living with this kind of chronic diseases is perhaps the most difficult challenge that people can face in their life. Chronically ill people should learn more details about the disease. It is very important to have knowledge about the illness, its types, its stages and the ways of its treatment. Chronic Disease Diagnosis (CDD) systems can prove to be valuable tools for proper control and management of the chronic disease.

II. LITERATURE REVIEW

Min Chen, Yixue Hao, Kai Hwang et al. , [1] *"Disease Prediction by Machine Learning over Big Data from Healthcare Communities"* In this paper the authors have proposed a new convolutional neural network based multimodal disease risk prediction algorithm using data from hospital. They have used machine learning algorithms for the predicting if the person is suffering from cerebral infarction (a chronic disease) or not. They have experimented the prediction model over real life hospital data collected from central China in 2013-2015. They have used Naive Bayesian, K- Nearest Neighbor and Decision Tree algorithms for classifying Structured Data and Convolutional Neural Network Algorithm for Unstructured Data. Compared to other typical algorithms the prediction accuracy of their proposed algorithm reaches 94.8 percent

Messan Komi, Jun Li ,Yongxin Zhai et al. , [2] *"Application of Data Mining Methods in Diabetes Prediction "* in their paper the authors have proposed that methods based on data mining can be effectively used to predict diabetes. They have explored the early prediction of diabetes using five different data mining methods which includes GMM, SVM, Logistic regression, ELM, ANN Algorithms. As per the experiment conducted the results prove that ANN (Artificial Neural Network) provides the highest accuracy of 0.89 to predict the disease which is more than other data mining techniques. The Logistic regression and SVM are less able to obtain an expected result if compared with other methods and due to the complexity and the variety of the data set.

Syed Immamul Ansarullah, Pradeep Kumar Sharmaet al.,[3] *"Heart Disease Prediction System using Data Mining Techniques: A study"* in this paper authors have highlighted the importance of data mining techniques in healthcare sector. According to them applying Data Mining techniques on healthcare data can help in predicting the likelihood of patients getting heart disease. In the paper the authors have analyzed the data mining techniques. The methodology that was used for this paper was survey of journals and publications in the fields of medicine. This research also concluded that good performance of the system can be obtained only by preprocessed and normalized dataset. The classification accuracy can be improved by using some feature selection techniques. This study also highlights the important role played by data mining tools in analyzing huge volumes of healthcare related data in prediction and diagnosis of disease.

Veena Vijayan V, Anjali C. , [4] *"Prediction and Diagnosis of Diabetes Mellitus –a Machine Learning Approach"* in this authors have proposed a decision support system that uses AdaBoost algorithm with Decision Stump as base classifier for classification.

The experiment focused on prediction of diabetes using AdaBoost algorithm. They have also used Support Vector Machine, Naive Bayes and Decision Tree algorithms to implement as base classifiers for AdaBoost algorithm for verification of accuracy. The dataset that was used was taken from UCI repository of machine learning for training that contained 768 instances and 9 attributes and used local dataset for validation which had been collected from various places in Kerala. The proposed classifier was found to be 80.72 percent which is greater compared to Support Vector Machine, Naive Bayes and Decision Tree.

Seema Sharma, Jitendra Agrawal et al. , [5] *"Machine Learning Techniques for Data Mining: A Survey"* in this paper they have compared machine learning algorithms like Decision Tree, Bayes algorithm, Support Vector Machine and Nearest Neighbor. These algorithms are used for classification mainly. They are used for predicting group membership for data instances. They provide a relative analysis of various algorithms. In data mining they extract the hidden predictive data from the large database. They have analyzed machine learning calculations like Decision Tree, Bayes algorithm, Support Vector Machine (SVM) and Nearest Neighbor. These figuring are used all together generally. They are used for anticipating group enlistment for data illustrations. They give a relative examination of various calculations. In information mining they remove the covered insightful data from the sweeping database.

Jyoti Soni et al. , [6] *"Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction"* here in this paper author presented three classifiers based on Decision Tree, Naïve Bayes and Classification using clustering to diagnose the presence of heart disease in patients. They experimented using WEKA 3.6.0 tool on the dataset of 909 records with 13 different attributes. The attributes were made categorical and inconsistencies were resolved for simplicity. In order to enhance the prediction of classifiers, genetic search was incorporated. Results concluded that the Decision Tree performs well than other two data mining techniques after incorporating feature subset selection but with high model construction time.

III. CLASSIFICATION ALGORITHMS

For doing comparative study of different machine learning algorithm we have used following algorithms:

1. Decision Tree(J48):

A Decision Tree represents rules and it is very popular tool for classification and prediction. J48 algorithm uses divide and conquer strategy for creation of decision tree using greedy algorithm. It creates tree with the help of recursion. A DT contains leaf node which is class label and decision nodes.

2. Naive Bayes:

The naïve bayes classification algorithm is a probabilistic classifier. Naïve Bayes Classifier uses Bayes theorem for prediction of class label of instance.

Bayes Theorem: $P(x|Y) = \frac{P(Y|x)*p(x)}{P(Y)}$

3. IBK or KNN :

K-Nearest Neighbors is used in the field of Pattern Recognition. All training tuples are stored in n-dimensional space. KNN searches the pattern space for k training tuple which are closest to the unknown testing tuple using measure like Eculidean distance.

$$d(p,q) = \sqrt{(\mathrm{p1}-\mathrm{q1})^2 + (\mathrm{p2}-\mathrm{q2})^2 + \cdots + (\mathrm{pn}-\mathrm{qn})^2}$$

4. Logistic:

Logistic classification technique is used for Binary classification problem. It is type of regression technique. Regression analysis is measure of strength of relationship between one dependent variable and more independent variable using some statistical tools.

5. SMO:

Sequential Minimal Optimization is a way to solve the Support Vector Machine training problem that is more efficient than standard QP solvers.

Dataset: For our system we have used Pima Indian Diabetes dataset, which is collected from UCI repository. Datasets are analyzed under different classification techniques. All information about the dataset is given in following table.

Table1. Total Instances

Dataset	Instance	Attributes
Pima Indian Diabetes	**768**	**9**

Algorithm	Accuracy %	True Positive Rate %	False Positive Rate %	Error %
Decision Tree(J48)	**77**	**80.3**	**29.4**	**23**
Naïve Bayes	**77**	**83.3**	**35.3**	**23**
KNN	**70.18**	**79.4**	**47**	**29.82**
Logistic	**74**	**81.8**	**41.2**	**26**
SVM	**74**	**84.8**	**47.1**	**26**

Table2. Performance Analysis of algorithms without pre-processing dataset

- Some performance evaluation factors are given below:

1. **Accuracy:-**correctly classified instances / total number of instances.
2. **True Positive Rate:- TP/(TP+FN)**
3. **False Positive Rate:-FP/(FP+TN)**
4. **Error rate:-** 1-Accuracy

Algorithm	Accuracy %	True Positive Rate %	False Positive Rate %	Error %
Decision Tree(J48)	**84.11**	**93.6**	**33.6**	**15.89**
Naïve Bayes	**77.50**	**84.2**	**38.4**	**22.5**
KNN	**70.18**	**79.4**	**47**	**29.82**
Logistic	**77.21**	**88.0**	**42.9**	**22.79**
SVM	**77.47**	**90.4**	**46.6**	**22.53**

Table3. Performance Analysis of algorithms by pre-processing dataset

Applying all above explained algorithms on dataset without pre-processing it. Finding out the performance of algorithms.

Now Preprocessing techniques are applied on dataset such as Normalization, Discretization and replacing missing values and finding out the performance of machine learning algorithm.

IV. CONCLUSION

By our analysis of literature survey, it was observed that the prediction done earlier did not use large dataset. On the basis of our analysis the result that we got is that the Naive Bayes and Decision Tree are the best classification algorithms.

REFERENCES

[1]Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities"2169-3536 (c) 2016 IEEE.

[2]Messan Komi, Jun Ki, Y ongxin Zhai, Xianguo Zhang, "Application of Data Mining Methods in Diabetes Prediction" 978-1-5090-6238-6/17/$31.00 ©20 17 IEEE

[3]Syed Immamul Ansarullah, Pradeep Kumar Sharma, Abdul Wahid, Mudasir M Kirmani, "Heart Disease Prediction System using Data Mining Techniques: A study", © 2016, IRJET

[4]Veena Vijayan V, Anjali C., "Prediction and Diagnosis of Diabetes Mellitus -AMachine Learning Approach" 978-1-4673-6670-0/15/$31.00 ©2015 IEEE

[5]Seema Sharma , Jitendra Agrawal, Shikha Agarwal., "Machine Learning Techniques for Data Mining: A Survey"978-1-4799-1597-2/13/$31.00 ©2013 IEEE

[6] Jyoti Soni et al., "Predictive Data Mining Diagnosis: An Overview of Heart Disease Prediction; International Journal of Computer Applications (0975-8887) Volume 17-No. 8, March 2011