International Conference on Advances in Materials, Computing and Communication Technologies
In Association withInternational Journal of Scientific Research in Science and Technology
Volume 9 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

Comparative Study on Classification Algorithms for Disease Prediction

A. Anisha[1], C. Renit[1], A. Anitha[2]

[1]Assistant Professor, Department of CSE, St. Xavier's Catholic College of Engineering, Tamil Nadu, India

[2]Associate Professor, Department of CSE, Noorul Islam Centre for Higher Education, Tamil Nadu, India

ABSTRACT

Data Mining plays an important role in data analysis process intended to discover data. There is huge amount of medical data but there is lack of powerful analysis tools to discover the hidden relationships and trends within the data. A disease prediction system forecasts the presence of a disease in a patient based on their symptoms. Also, it will recommend essential preventive measures required to treat the disease predicted. Application of data mining in disease prediction helps to predict the most possible disease based on the given symptoms and can avoid the aggression of disease. This paper presents a comparative study on application of classification algorithms for disease prediction. The Findings show that the proposed system can predict disease with an accuracy of 95.67%.

I. INTRODUCTION

It is estimated that more than 70% of people in India are prone to general body diseases like viral, flu, cough, cold etc., in every 2 months. General body diseases may be symptoms for more harmful diseases. Because of ignoring the early general body symptoms, 25 % of our population succumbs to death. This leads to a hazardous situation for our population. So, early prediction of diseases is very crucial to avoid such undesirable casualties.

II. LITERATURE SURVEY

M. Gandhi and S. N. Singh1 gave a comparitative study on heart disease prediction system and the proposed model can be enhanced. S. Palaniappan and R. Awang2 presents a comparison on data mining classification modeling techniques for heart disease prediction and the experiments were conducted on limited size data set. D. Dahiwade et.al 3 performed a comparison on K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) for general disease prediction system and achieved an accuracy of 84.5% using CNN.

S. Meesri et.al 4 proposed a hybrid model for the diagnosis of heart diseases and achieved good results. C. Raju et.al 5 presented a detailed study on prediction of heart diseases using various classification algorithms and their findings show that Support Vector Machine gives the best result. B. D. Kanchan and M. M. Kishor 6 performed a study on supervised machine learning algorithms to predict heart diseases. F. Huang et.al 7 applied the data mining algorithms to predict hypertension from patient medical records. Kaur et.al 8 provides an insight on future trends on application of data mining tools for healthcare. Verma AK et.al 9 presented a study on prediction of various classes of skin disease with different classification algorithms. Taranu I discussed on the methodologies of data mining that can be used in healthcare.

This paper presents a general disease prediction system that predicts presence of a disease present in a patient on the basis of their symptoms. The model is implemented with three classification algorithms namely, K-Nearest Neighbor, Decision tree and Naïve Bayes algorithms. The objective is to predict possible disease from the patient data-set using the algorithms and determines which model gives the highest percentage of accurate predictions for the diagnoses.

III. THE PROPOSED GENERAL DISEASE PREDICTION MODEL

The steps involved in developing the proposed model are as follows.

1. Pre-processing
2. Model Training and Evaluation
3. Prediction

The system will initially be fed data from different sources i.e. patients, the data will then be pre-processed before further process is carried out; this is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. The data is preprocessed using data wrangling methods like data binning and data normalization. This data will be processed using Data mining algorithms, the system will be trained so as to predict the disease based on the input data given by the user. Fig.1. shows the architecture Diagram of the proposed disease prediction system.

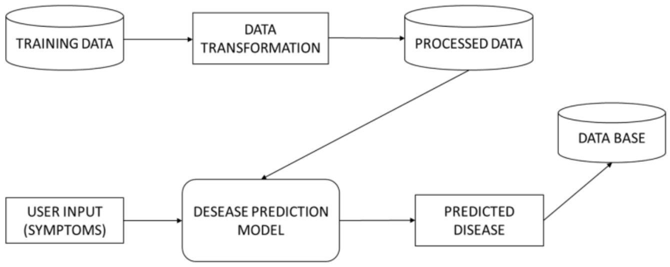


Fig.1. Architecture Diagram of the proposed disease prediction system

IV. DATA MINING ALGORITHMS

K-Nearest Neighbour algorithm

K-Nearest Neighbour algorithm is a supervised classification algorithm. It predicts the target label with the help of nearest neighbour class. Closeness is measured with distance measures like Euclidean distance.

PSEUDO CODE:

```
1.  Load the training and test data
2.  Choose the value of K
3.  for each point in test data:
      -  find the Euclidean distance to all training
         data points
      -  store the Euclidean distances in a list and
         sort it
      -  choose the first k points
      -  assign a class to the test point based on the
         majority of classes present in the chosen
         points

4.  End
```

Decision tree algorithm

Decision tree algorithm is a supervised classification algorithm. The Decision Tree algorithm creates a model based on the training data and generates learning decision rules based on the model which can be used to predict class or value of new test data.

The basic algorithm used in decision trees is known as the ID3 which builds decision trees by a top-down approach. First the best attribute is selected (let it be X). For each value of X, create a sub tree. Then arrange the training data according to the value of X. If training data are classified accurately, then the processing is over, else iterate over the new leaf nodes.

Naïve Bayes algorithm

Naïve Bayes algorithm is probabilistic Machine learning algorithm that works based on Bayes theorem.

$P(A|B) = P(A)P(A)$ (1)

P(A): The probability of hypothesis H being true.

P(B): The probability of the evidence. P(A|B): The probability of the evidence given that hypothesis is true.

P(B|A): The probability of the hypothesis given that the evidence is true.

The Bayes classifier is a function that assigns a class label l= Ck for some k as,

$l = p(C\ ) \prod p(x\ |C\ )(2)$

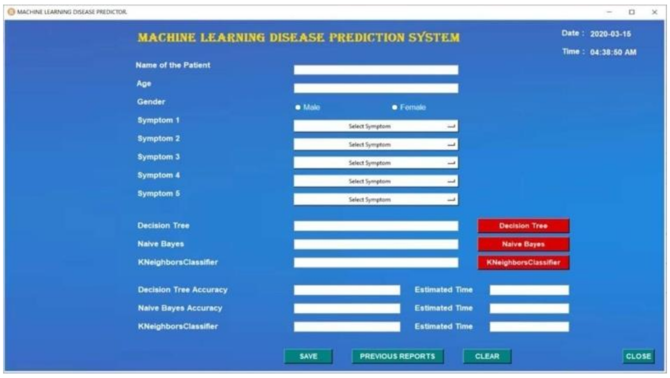
k

i=1

i k

5. Experimental Results and Discussion

The proposed system is implemented with python as front end and SQLite3 as backend and tested on a sample dataset with the data of 4952 patients. In the main page, the User/Medical representative can enter the information about the patient along with a maximum of five symptoms they have. Then they entered data is validated for correctness to reduce the usage of pre-processing which help in faster analysis of data. After pre-processing, the disease of the given patient is predicted with each algorithm and the accuracy of the algorithm and the time taken to predict is recorded. Fig.2. shows the user interface of the Disease Prediction system. Fig. 3 presents the accuracy of the classification algorithms. It was observed that both Decision tree and Naïve Bayes algorithm predicts the diseases with the same accuracy leaving the Nearest Neighbour algorithm behind. Fig. 4 presents the time taken to predict diseases of the classification algorithms. It was evident that the Naïve Bayes algorithm was able to predict disease within 0.031 milliseconds. So, among the three classification algorithms, Naïve Bayes algorithm was efficient in terms of both time taken and accuracy for the general prediction of diseases.



MACHINE LEARNING DISEASE PREDICTOR

MACHINE LEARNING DISEASE PREDICTION SYSTEM

Date : 2020-03-15
Time : 04:38:50 AM

Name of the Patient
Age
Gender ● Male ● Female
Symptom 1 Select Symptom
Symptom 2 Select Symptom
Symptom 3 Select Symptom
Symptom 4 Select Symptom
Symptom 5 Select Symptom

Decision Tree
Naive Bayes
KNeighborsClassifier

Decision Tree Accuracy Estimated Time
Naive Bayes Accuracy Estimated Time
KNeighborsClassifier Estimated Time

Fig.2.User Interface of the Disease Prediction system

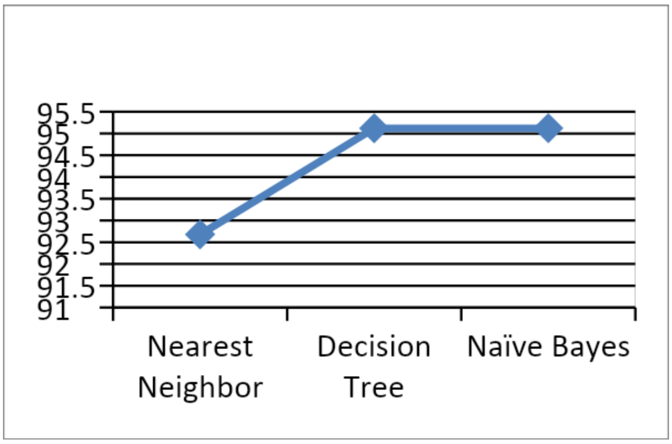


Fig 3. Comparison of Accuracy for prediction

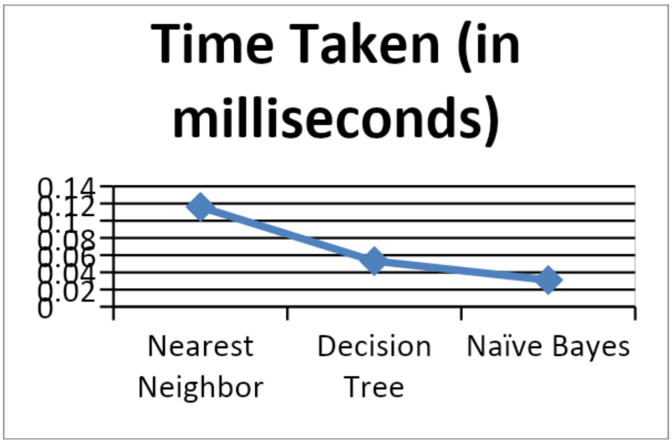


Fig 4. Comparison of Time Taken for prediction

V. CONCLUSION

Early Prediction of diseases helps us to quickly decide the severity of diseases and to save lives before they are in real danger. A general disease prediction system has been implemented, which accepts the symptoms as input and predicts the presence of diseases. The system could also be used for keeping track of patients record and generating in- sights of their records as report that may help in their future medications and health condition to be diagnosed. Classification algorithms such as, K-Nearest Neighbor, Decision tree and Naïve Bayes algorithm are implemented and their effectiveness is measured. The findings clearly show that Naïve Bayes algorithm outperforms the other algorithms in terms of both time complexity and accuracy. The system can be enhanced in the future for the early prediction of fatal diseases like cancer too.

VI. REFERENCES

[1]. M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), Greater Noida, India, 2015, pp. 520-525, doi:10.1109/ABLAZE.2015.7154917.

[2]. S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 2008, pp. 108-115, doi:10.1109/AICCSA.2008.4493524.

[3]. D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1211-1215, doi:10.1109/ICCMC.2019.8819782.

[4]. S. Meesri, S. Phimoltares and A. Mahaweerawat, "Diagnosis of Heart Disease Using a Mixed Classifier," 2017 21st International Computer Science and Engineering Conference (ICSEC), Bangkok, Thailand, 2017, pp. 1-5, doi: 10.1109/ICSEC.2017.8443940.

[5]. C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh and S. Deepa Rajan, "A Survey on Predicting Heart Disease using Data Mining Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, India, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.

[6]. B. D. Kanchan and M. M. Kishor, "Study of machine learning algorithms for special disease prediction using principal of component analysis," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, India, 2016, pp. 5-10, doi: 10.1109/ICGTSPICC.2016.7955260.

[7]. F. Huang, S. Wang and C. Chan, "Predicting disease by using data mining based on healthcare information system," 2012 IEEE International Conference on Granular Computing, Hangzhou, China, 2012, pp. 191-194, doi: 10.1109/GrC.2012.6468691.

[8]. Kaur, Shubpreet & Bawa, R, "Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System", International Journal of Energy, Information and Communications.2015, 6. 17-34. 10.14257/ijeic.2015.6.4.02.

[9]. Verma AK, Pal S, Kumar S. Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method-a Comparative Study. Appl Biochem Biotechnol. 2020 Feb;190(2):341-359. doi: 10.1007/s12010-019-03093-z. Epub 2019 Jul 27. PMID: 31350666.

[10]. Taranu I.,"Data mining in healthcare: decision making and precision", Database Systems Journal. 2015; VI (4).