International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 09 | Sept - 2021 ISSN: 2582-3930

HUMAN DISEASE PREDICTION MODEL USING MACHINE LEARNING APPROACH

[1]DR. Kalaivani D Hareesh,[2] S.Nivedha,[3]Monisha K, [4]Roshini P, [5]Rushika Bali

[1]Dean Research and Development, 8th semester, Department of Information Science &l Engineering, New Horizon College of Engineering, Bangalore, India
[2,3,4]UG students, Department of Information Science& Engineering, New Horizon College of Engineering, Bangalore, India

ABSTRACT

Now-a-days, people face various diseases due to the environmental condition and their living habits. So the prediction of disease at earlier stage becomes important task. But the accurate prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. We proposed general disease prediction based on symptoms of the patient. For the disease prediction, we use (DECISION TREE AND NAVE BAYES) machine learning algorithm for accurate prediction of disease. For disease prediction required disease symptoms dataset. In this general disease prediction the living habits of person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using decision tree and nave bayes is 84.5%.

1. INTRODUCTION

Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Many works have been applied data mining techniques to pathological data or medical profiles for prediction of specific diseases. These approaches tried to predict the reoccurrence of disease. Also, some approaches try to do prediction on control and progression of disease. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre processing and produce more accurate results. With the development of big data technology, more attention has been paid to disease prediction from the perspective of big data analysis; various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification rather than the previously selected characteristics.

The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients.

The healthcare industry produces large amounts of healthcare data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision making for patient's health. Also, this areas need improvement by using the informative data in healthcare. One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made.

2. LITERATURE SURVEY

M. Chen proposed [1] a new DT based multimodal disease risk prediction algorithm by using structured and unstructured data of hospital. M. Chen ,Y. Hao, K. Hwang, L. Wang, and L. Wang invented disease prediction system for the numerous regions. They performed disease prediction on three diseases like diabetics, cerebral infraction and heart disease. The disease prediction is carried out on structured data. Prediction of heart disease, diabetes and cerebral infraction is carried out by using different machine learning algorithm like naïve bayes, Decision tree and NB algorithm. The result of Decision tree algorithm is better than Naïve bayes and NB algorithm. Also, they predict that whether a patient experiences from the high risk of cerebral infarction or low risk of cerebral infarction. For the risk prediction of cerebral infraction, they utilized DT based multimodel disease risk prediction on text data. The accuracy comparison takes place between DT based unimodel disease risk predictions against DT based multimodel disease risk prediction algorithm. The accuracy of disease prediction reaches up to the 94.8% with faster speed than DT based unimodal disease risk prediction algorithm. The DT based multimodel disease risk prediction algorithm steps is similar as that of the DT-UDRP algorithm only the testing steps consist of two additional steps. This paper work on both the type of dataset like structured and unstructured data. Author worked on unstructured data. While previous work only based on structured data, none of the author worked on unstructured and semi- structured data. But this paper depends on structured as well as unstructured data.

.

3.EXISTING SYSTEM:

Prediction using traditional disease risk model usually involves a machine learning and supervised learning algorithm which uses training data with the labels for the training of the models. High-risk and Low-risk patient classification is done in groups test sets. But these models are only valuable in clinical situations and are widely studied. A system for sustainable health monitoring using smart clothing by Chen et.al. He thoroughly studied heterogeneous systems and was able to achieve the best results for cost minimization on the tree and simple path cases for heterogeneous systems. The information of patient's statistics, test results, and disease history is recorded in EHR which enables to identify potential data-centric solutions which reduce the cost of medical case studies. Bates et al. propose six applications of big data in the healthcare field. Existing systems can predict the diseases but not the subtype of diseases. It fails to predict the condition of people. The predictions of diseases have been non-specific and indefinite.

4.PROPOSED SYSTEM:

In this Project , we have combined the structure and unstructured data in healthcare fields that let us assess the risk of disease. . Medical science has large amount of data growth per year. Due to increase amount of data growth in medical and healthcare field the accurate analysis on medical data which has been benefits from early patient

care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. We proposed general disease prediction based on symptoms of the patient. For the disease prediction, we use (DECISION TREE AND NAVE BAYES) machine learning algorithm for accurate prediction of disease. For disease prediction required disease symptoms dataset. In this general disease prediction the living habits of person and checkup information consider for the accurate prediction. The approach of the latent factor model for reconstructing the missing data in medical records which are collected from the hospital. And by using statistical knowledge, we could determine the major chronic diseases in a particular region and in particular community. To handle structured data, we consult hospital experts to know useful features. In the case of unstructured text data, we select the features automatically with the help of k-mean algorithm. We propose a k-mean algorithm for both structured and unstructured data.

5.SYSTEM ARCHITECTURE

Fig -1: System Architecture

Fig -1: System Architecture

6.ALGORITHM

6.1.DECISION TREE

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated

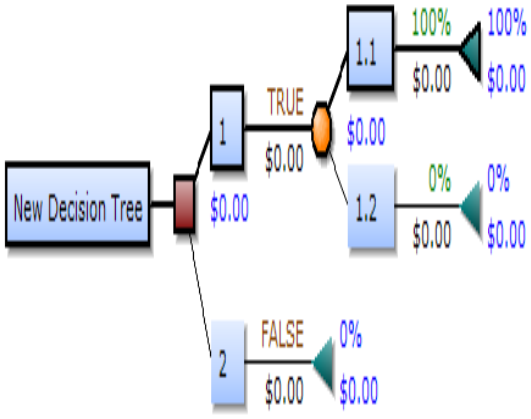
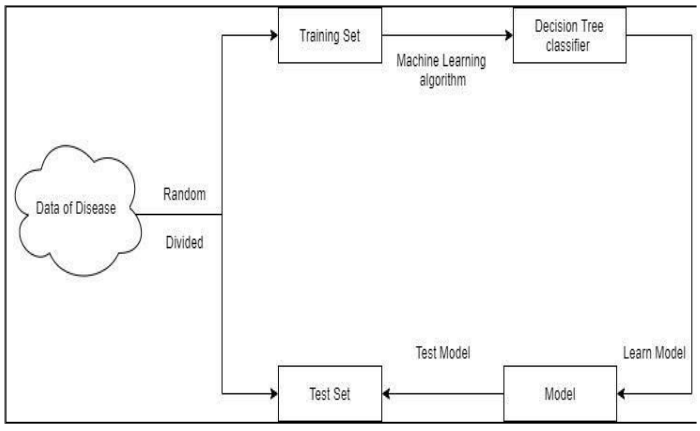


FIG 2 DECISION TREE ELEMENTS

The decision tree can be linearized into decision rules,[2] where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form:

if condition1 *and* condition2 *and* condition3 *then* outcome.



Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy

most likely to reach a goal, but are also a popular tool in machine learning.

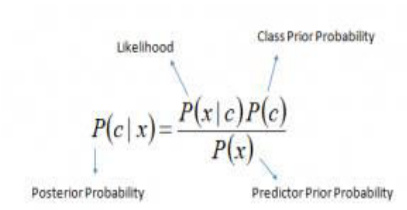
6.2 NAVIE BAYES ALGORITHM:

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

7. CONCLUSIONS

With the proposed system, higher accuracy can be achieved. We not only use structured data, but also the text data of the patient based on the proposed k-mean algorithm. To find that out, we combine both data, and the accuracy rate can be reached up to 95%. None of the existing system and work is focused on using both the data types in the field of medical big data analytics. We propose a K-Mean clustering algorithm for both structured and unstructured data. The disease risk model is obtained by combining both structured and unstructured features.

ACKNOWLEDGEMENT

We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.
We would like to thank our Dean and HOD,
Dr.R.J Anandhi for guiding us in every step. The guidance and support received from all teaching and non-teaching staff of Information Science & Engineering department, **New Horizon college of Engineering** who contributed to this project. We are grateful for their constant support and help.
We are also grateful to our project guide **Dr.Kalaivani D Hareesh,** Dean (R&D) New Horizon college of Engineering, for his tremendous support and help. Without his encouragement and guidance this project would not have materialized.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang,"Disease prediction by machine learning over big data from healthcare communities", ," IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[3] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.

[4] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

[5] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for telehealth in cloud computing," in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), Nov. 2016, pp. 184–189.

[6] Disease and symptoms Dataset –www.github.com.

[7] Heart disease Dataset-WWW.UCI Repository. com

[8] Ajinkya Kunjir, Harshal Sawant, Nuzhat F.Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare," in IEEE big data analytics and computational intelligence, Oct 2017 pp.2325.