International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Developing a Social Media Platform for Disease Prediction

Aditya Khandelwal[1], Ajay Kandi[2], Aditya Vyas[3], Hrithvik Ranka[4], Archana Khandekar[5]

[1,2,3,4]*Students , CSE Department, MIT-WPU, Pune, India*
[5]*Assistant Professor , CSE Department, MIT-WPU, Pune, India*

Abstract - *Hospitals are the foremost widely used ways by which a sick and diseased person gets medical check-ups, gets disease diagnosis and therefore the treatment recommendation. This has been a practise by the majority of people over the planet. People treat it as the most reliable means to test their status of health. The fundamental purpose for developing this system is disease prediction using various data processing and mining techniques. This Thesis describes a new approach to disease prediction which helps to forestall future life losses. This thesis emphasizes on every single symptom associated with each and every disease. This will be accomplished by building a web platform in which medical professionals can interact with each other, share their experiences and knowledge. This ends up in a dynamically growing online survey, which ultimately helps in data collection that can be used to identify various diseases and helps to stop them. This portal can be used for various purposes where different medical institutions can use our data to launch their products moreover to as acquire feedbacks. This paper will going to be helpful for students perceiving medical studies they can collect the correct information from the acceptable source and in precise manner.*

***Key Words*: Data Processing, Data Mining, Disease prediction, Mining Techniques , KNN , Neural Network.**

1.INTRODUCTION

Since the past few decades, humans are tirelessly working day and night that they fail to prioritize their health on a daily basis. Within the longer run, this problem results in jeopardizing the standard of life. Nevertheless, with the help of AI, we will now going to provide health care services to individuals at their convenience at reasonable prices. One among the most important blessings we possess may be a healthy body. A healthy body and enhanced quality of life are some things all folks looks up to. Disease prediction is one among the most important goal of the researchers supported the facts of massive data analysis which successively improves the accuracy of risk classification supported the info of an outsized volume. E-healthcare facilities generally, are an important resource to developing countries but are often difficult to determine due to the shortage of awareness and development of infrastructure. Variety of internet users depend upon the web for clearing their healthcare-based queries. So in this paper we are going to talk about designing a platform for providing online disease prediction to patients with a goal to supply assistance to healthcare professionals. The user also can seek medical guidance in a neater way and obtain exposure to various diseases and diagnosis available for it.

2. LITERATURE SURVEY

S. Apte (2012) proposed a data mining classification technique for the prediction of heart related diseases. He applied data preprocessing technique to remove missing values and those missing value were replaced by mean mode method. Later the multi-layer perceptron neural network was used for the mapping of the data. Hence data mining techniques such as naive Bayes, neural network and decision trees were analyzed and used on Heart disease database.

S. Pal (2013) describe the model for predicting heart disease using data mining techniques. The methodology which were proposed by him were surveyed on three different classifiers i.e ID3 (Iterative Dichotomized 3), Decision tree, and CART (Classification and Regression tree). He collected the dataset from Cleveland Clinic Foundation. His observation and comparison showed that Classification and Regression tree (CART) achieved the accuracy of 83.49% which was comparatively better then ID3 (Iterative Dichotomized 3) and Decision tree.

R. Ceylan (2013) proposed a model for identifying biomedical pattern classification using artificial neural network based on rotation forest (RF-ANN). The model used multilayer perceptron neural network as the base classifier. RF algorithm was used as ensemble classifier. Different feature sets were obtained from original data set using the principal component analysis technique. The dataset was obtained from Wisconsin breast cancer from the University of California Irvine (UCI) ML repository and the accuracy of the system found out to be 98.05%.

3. OVERVIEW OF SYSTEM

The basic purpose for developing this system is disease prediction using various data mining techniques. This paper describes a new approach to disease prediction which helps to prevent future life losses. This will be accomplished by building a web platform in which medical professionals and medical practitioners can interact with each other, share their experiences and knowledge. People can also interact with our system just like they do with another human and through a series of queries our system will identify the symptoms of the user and thereby predicts the disease and will recommends treatment. This paper can be of great use to people in conducting daily checkups , make people aware of their health status and encourages people to make proper measures to remain healthy.

4. SYSTEM MODULES

4.1 *Test data acquisition module* :

Test data acquisition module is test data acquisition or Data selection which is the process of determining the acceptable data type and source, also as suitable instruments to gather data. Data selection precedes the particular practice of knowledge collection. This definition differentiates data selection and selective data reporting from each other.

4.2 *Data pre-processing module* :

Data preprocessing describes any sort of processing performed on data to organize it for an additional processing procedure. Commonly used as a preliminary data processing practice, data preprocessing transforms the info into a format which will be more easily and effectively processed for the aim of the user.

4.3 *Data mining module* :

Data mining is the process which includes extraction of valid information from large databases. Data processing is utilized in many research areas like in medical processing. There are many techniques of knowledge mining which are developed for deciding and making decisions (DSS). If decision making techniques are properly applied on data then these data will get stored in databases. These data are basically used to understand the hidden correlation between symptoms and diseases.

4.4 *Result analysis And testing* :

Disease predicting models must be thoroughly tested and validated after being developed. Interest has arisen lately in model validation through the quantification of the economic costs of false positives and false negatives, where disease prevention measures could also be used.

5. DATA MINING CLASSIFICATION TECHNIQUES

5.1 *KNN Algorithm*:

K-Nearest Neighbour is an algorithm which is based on the similarity with other items. The items which are similar to each other are called neighbours. If a new item is found then its distance from other items present in the model is calculated. This classifies the partition and item to the nearest neighbour which is also the most similar one, then that item is placed in a group that includes the nearest neighbours.

5.2 *Neural Network* :

Artificial Neural Network is inspired from the knowledge that is considered as a data processing . In this algorithm many microprocessors are responsible for data processing and they are acting as an interconnected and parallel network with each other to solve a problem. By using data science in this network a data structure can be designed that can act as a neuron and this data structure further will be called as neuron. By setting the network b/w the neurons and using a learning algorithm, the network is trained. In this Neural Network neurons are divided into two modes i.e enable (NO or 1) or disable (OFF or 0) and each edge (synapses or connections between nodes) will have a weight. Edges with positive weight stimulates/enable the next disable nodes and edges with negative weights disable/inhibit the next connected nodes (if they are enabled).

5.3 *Naïve Bayes* :

Naive Bayes classifier is based on Bayes theorem. This classifier uses conditional independence in which the value of the attribute is independent of the values of other attributes.

The Bayes theorem is as follows:

Let X = {x_1, x_2, x_3, ... , x_n} be a set of n attributes.
In Bayesian,
X is considered as evidence
H be some hypothesis means
And the data of X belongs to specific class C.

We have to determine P (H|X) which means the probability that the hypothesis H holds given evidence i.e. data sample X. According to the Bayes theorem the P(H|X) is expressed as :

P(H|X) = P(X|H) x P(H) / P(X)

6. SYSTEM DESIGN

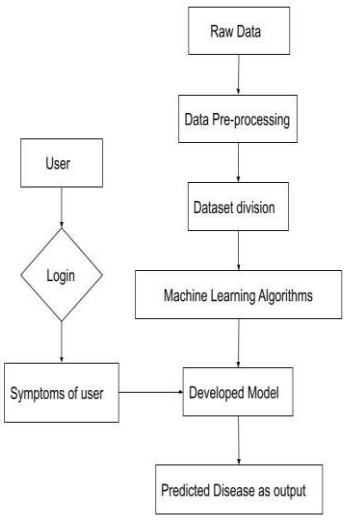

Fig-1 : System Flow Diagram

This system provides an interactive portal where infected person needs to register his details. After successful login he will be given access to submit his symptoms from which he is suffering. User can submit multiple symptoms for better accuracy in disease prediction. These symptoms will be processed and used to predict disease using Machine Learning algorithms. User can also contact to expert for seeking advice and to solve his queries and get necessary solutions.

7. METHODOLOGY

The platform in our project will be constructed for information acquisition. It acquires the patient's information along with the symptoms and therefore the disease is predicted on the idea of the symptoms. The disease prediction system is meant using the concepts of NLP and machine learning algorithms. The system will comprise of two concepts of NLP namely tokenization and wordnet. Upon receiving the symptoms from the user, tokenization is performed and therefore the symptoms are extracted from the sentences the user has entered. Synset may be a simple interface that's present in NLTK (Natural Language Toolkit) to see whether the words are present or not in WordNet. Synset instances will be used to extract the words synonymous with the symptoms.

If the symptom entered is wrong, an invalid response is generated and will be sent to the user and if the symptoms entered are valid, then the symptoms which are extracted will be sent to the classifier. The second phase will be comprised of classification which will be consisting of two processes, which are prediction and learning. According to the first process, a model will be made based on the training data and within the second process, the best model will be used to predict the response for the data. Our model will going to use the approach of Supervised Learning.

8. CONCLUSIONS

This research paper tells about the prediction model driven system that predicts accurate diseases based on the provided symptoms. The concept of NLP will be used to design an interactive platform which uses symptoms provided by the user. The prediction model will be designed using ML algorithms such as KNN. The algorithm will be selected and applied on the dataset based on the confidence and accuracy rate and according to that further changes will be made. We believe that if this approach is incorporated into existing strategies in the field of healthcare then it will provide great assistance to medical professionals and practitioners.

REFERENCES

[1] Cios K J,Pedryz W,Swiniark R M.Data Mining methods for knowldge discovery.IEEE Trans Neural Netw.1998,9(6): 15331534

[2] K.Sinivas," Mining Assocition Rules from Large Datasets towards Disease Prediction", 2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT.

[3] Fayad U M, Pitetsky-Shapiro G, Smith P. From data mining to knowledge discvery: an overview. Advances in knowledge discvery and data mining , American Association for AI, 1996, 1-34.

[4] M. Chin, Y. Hao, K. Hwong, L. Wangg and L. Wangg, "Disease Prediction by ML Over Bigg Data From Healthcare Comunities," in IEEE Access, vol. 5, pp. 8869-8879, 2017.

[5] Yu, Hangg. (2019). Experimntal Diseases Prediction Research on Combning NLP and ML.145-1500.1109/ICCSNT47585.2019.8962507.

[6] S. Vinita and S. Swetlin and H. Viunsha and SSajini, Disease Prediction Using Machine Learning Over Big Data (February 2018). Computer Science and Enginering: An International Journal (CSEIJ), Vol.8, No.1, February 2018.

[7] Chaitrali S. Dangore , Sulabha S. Apte , Improved Study of Heart Diseases Prediction System using Data Mining Clasification Techniques (2012), Internatinal Journal of Computer Application (0975 – 888) Volume 47– No.10.

[8] Chaurasia V, Pal S (2013) Early prediction of heart diseases using data mining techniqes. Carib J Sci Technol 1:208–217

[9] Hasan Koyuncu , Rahime Ceylan Artifical neural network based on rotation forest for biomedical pattern clasification , 2013 IEEE 36th International Conference on Telecommunications and Signal Processing (TSP).