International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Disease Prediction using Machine Learning

Kedar Pingale[1], Sushant Surwase[2], Vaibhav Kulkarni[3], Saurabh Sarage[4], Prof. Abhijeet Karve[5]

[1]Kedar Pingale Zeal College of Engineering &Research Department of Information Technology
[2]Sushant Surwase Zeal College of Engineering &Research Department of Information Technology
[3]Vaibhav Kulkarni Zeal College of Engineering &Research Department of Information Technology
[4]Saurabh Sarage Zeal College of Engineering &Research Department of Information Technology
[5]Prof Abhijeet C. Karve Zeal College of Engineering &Research Department of Information Technology

---

Abstract - *Disease Prediction system is based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the Naive Bayes Classifier. Naive Bayes Classifier calculates the probability of the disease. With big data growth in biomedical and health care communities, accurate analysis of medical data benefits early disease detection, patient care. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.*

Key Words: Logistic Regression, Machine Learning, chronic Diseases, Python.

1. INTRODUCTION

Machine learning is programming computers to optimize a performance using example data or past data. Machine learning is study of computer systems that learn from data and experience .Machine learning algorithm has two passes: Training, Testing .Prediction of a disease by using patient's symptoms and history machine learning technology is struggling from past decades. Machine Learning technology gives a good platform in medical field, so that a healthcare issues can be solved efficiently.

We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyze data and deliver results faster, with the use of machine learning technology doctors can make good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare services. Healthcare is the most prime example of how machine learning is use in medical field.

To improve the accuracy from a large data, the existing work will be done on unstructured and textual data .For prediction of diseases the existing will be done on linear, KNN, Decision Tree algorithm. The order of reference in the running text should match with the list of references at the end of the paper.

2. OBJECTIVE

There is a need to study and make a system which will make it easy for an end users to predict the chronic diseases without visiting physician or doctor for diagnosis. To detect the Various Diseases through the examining Symptoms of patient's using different techniques of Machine Learning Models. To Handle Text data and Structured data is no Proper method. The Proposed system will consider both structure and unstructured data. The Predictions Accuracy will Increase using Machine Learning.

3. EXISTING SYSTEM

The system predicts the chronic diseases which is for particular region and for the particular community. The Prediction of Diseases is done only for particular diseases. In this System Big Data & CNN Algorithm is used for Diseases risk prediction. For S type data, system is using Machine Learning algorithm i.e K-nearest Neighbors, Decision Tree, Naïve Bayesian. The accuracy of the System is upto 94.8%.

Existing paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. We experiment the modified prediction models over real-life hospital data collected from central China. We propose a new convolutional neural net-work based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital.

4. PROPOSED SYSTEM

This system is used to predict most of the chronic diseases. It accepts the structured and textual type of data as input to the machine learning model. This system is used by end users. System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. For predicting diseases Naïve Bayes algorithm, for clustering KNN algorithm, final output will be in the form of 0 or 1 for which Logistic tree is used.

5. DATASET AND MODEL DESCRIPTION

In this we describe dataset which is being use to train the machine learning model. The dataset will contain symptoms of various diseases.

5.1 DATASET OF HOSPITAL

The hospital data will be in the form of textual format or in the structural format. The dataset used in this project is real life data. The structural data contains symptoms of patients while unstructured data consist of textual format.

The dataset used is contains real-life hospital data, and data stored in data center. The data provided by the hospital contains symptoms of the patients

6. EVALUATION METHOD

To calculate performance evaluation in experiment, first we denote TP, TN, Fp and FN as true positive (the number of results correctly predicted as required), true negative (the number of results not required), false positive (the number of results incorrectly predicted as required), false negative (the number of results incorrectly predicted as not required) respectively. We can obtain four measurements: recall, precision, accuracy and F1 measure as follows:

accuracy -:

$$\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

$$\text{Precision} = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$\text{Recall} = \frac{TruePositive}{TruePositive+FalseNegative}$$

$$\text{F1-Measure} = \frac{2\times precision\times recall}{precision+recall}$$

7. SYSTEM ARCHITECTURE

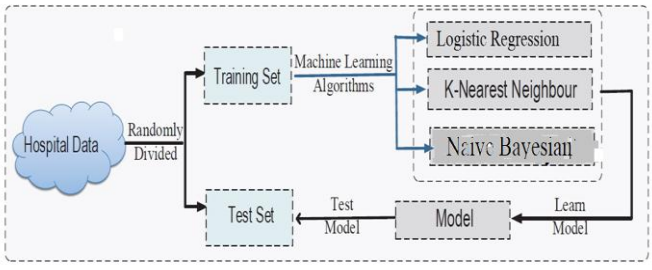

Fig -1: System Architecture

8. ALGORITHM

8.1 KNN

K Nearest Neighbor (KNN) could be a terribly easy, simple to grasp, versatile and one amongst the uppermost machine learning algorithms. In Healthcare System, user will predict the disease. In this system, user can predict whether disease will detect or not. In propose system, classifying disease in various classes that shows which disease will happen on the basis of symptoms. KNN rule used for each classification and regression issues. KNN algorithm based on feature similarity approach.

A case is classed by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is just assigned to the category of its nearest neighbor

$$Euclideandistance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

It ought to even be noted that every one 3 distance measures square measure solely valid for continuous variables. In the instance of categorical variables, the Hamming distance must be used. It conjointly brings up the difficulty of standardization of the numerical variables between zero and one once there's a combination of numerical and categorical variables within the dataset.

$$HammingDistance = \sum_{i=1}^{k}|x_i - y_i|$$

8.2 NAIVE BAYES

Naive Bayes is a easy however amazingly powerful rule for prognosticative modelling. One of the simplest ways that of choosing the foremost probable hypothesis given the info that we've that we are able to use as our previous information regarding the matter. Bayes' Theorem provides how that we are able to calculate the likelihood of a hypothesis given our previous information.

Naive Bayes classifier assumes that the presence of a specific feature in an exceedingly class is unrelated to the presence of the other feature. Bayes theorem provides some way of calculative posterior chance P(b|a) from P(b), P(a) and P(a|b). Look at the equation below:

$$P(b \vee a) = \frac{P(a \vee b)P(b)}{P(a)}$$

Above,

- P(b|a) is that the posterior chance of class (b, target) given predictor (a, attributes).
- P (b) is the prior probability of class.
- P(a|c) is that chance that is that the chance of predictor given class.
- P(a) is the prior probability of predictor.

8.3 LOGISTIC REGRESSION

Logistic regression could be a supervised learning classification algorithm accustomed predict the chance of a target variable that is Disease. The nature of target or variable is divided, which means there would be solely 2 potential categories.

In easy words, the variable is binary in nature having information coded as either 1 (stands for success /yes) or 0 (stands for failure / no). Mathematically, a logistic regression model predicts P(y=1) as a function of x.

Logistic regression can be expressed as:

$\log(p(X)/(1-p(X)))=\beta_0+\beta_1 X$

Where, the left hand side is called the logiest or log odds function, and p(x)/(1-p(x)) is called odds. The odds signifies the ratio of probability of success to probability of failure. Therefore in logistic Regression, linear combination of inputs are mapped to the log(odds) - the output being adequate to 1.

9. CONCLUSIONS

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease.

In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data

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" (2017).

[2] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[3] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The 'big data' revolution in healthcare: Accelerating value and innovation," 2016.

[4] S.-H. Wang, T.-M. Zhan, Y. Chen, Y. Zhang, M. Yang, H.-M. Lu, H.-N. Wang, B. Liu, and P. Phillips, "Multiple sclerosis detection based on biorthogonal wavelet transform, rbf kernel principal component analysis, and logistic regression," IEEE Access, vol. 4, pp. 7567–7576, 2016.

[5] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.