PREDICTION OF DISEASE USING MACHINE LEARNING

Vaibhav Kulkarni[1], Sushant Surwase[2], Kedar Pingale[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 using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.*

Key Words: Logistic Regression, Naïve Bayes Classifier, Decision Tree, Machine Learning.

1. INTRODUCTION

Machine Learning is the domain that uses past data for predicting. Machine Learning is the understanding of computer system under which the Machine Learning model learn from data and experience. The machine-learning algorithm has two phases: 1) Training & 2) Testing. To predict the disease from a patient's symptoms and from the history of the patient, machine learning technology is struggling from past decades. Healthcare issues can be solved efficiently by using Machine Learning Technology.

We are applying complete machine learning concepts to keep the track of patient's health. ML model allows us to build models to get quickly cleaned and processed data and deliver results faster. By using this system doctors will make good decisions related to patient diagnoses and according to that, good treatment will be given to the patient, which increases improvement in patient healthcare services. To introduce machine learning in the medical field, healthcare is the prime example.

To improve the accuracy of large data, the existing work will be done on unstructured or textual data. For the prediction of diseases, the existing will be done on linear, KNN, Decision Tree algorithm.

2. OBJECTIVE

Currently, the scenario is if the patient is suffering from any symptoms then he/she must visit to the doctor or to the hospital to diagnose the disease. But, our main objective is to reduce such efforts taken by patients only to diagnose the disease. Many patients are losing their life only because of the late diagnosis of their disease. So our main aim is to reduce such deaths.

3. EXISTING SYSTEM

The existing system predicts the chronic diseases which are for a particular region and for the particular community. Only particular diseases are predicted by this system. In this System, Big Data & CNN Algorithm is used for Disease risk prediction. For S type data, the system is using Machine Learning algorithm i.e K-nearest Neighbors, Decision Tree, Naïve Bayesian. The accuracy of the existing System is up to 94.8%.

In the existing paper, they streamline machine learning algorithms for the effective prediction of chronic disease outbreak in disease-frequent communities. They experiment with the modified prediction models over real-life hospital data collected from central China. They propose a convolutional neural network-based multimodal disease risk prediction(CNN-MDRP) algorithm using structured and unstructured data from the hospital.

4. PROPOSED SYSTEM

Most of the chronic diseases are predicted by our system. It accepts the structured type of data as input to the machine learning model. This system is used by end-users i.e. patients/any user. In this system, the user will enter all the symptoms from which he or she is suffering. These symptoms then will be given to the machine learning model to predict the disease. Algorithms are then applied to which gives the best accuracy. Then System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. Naïve Bayes algorithm is used for predicting the disease by using symptoms, for classification KNN algorithm is used, Logistic regression is used for extracting features which are having most impact value, the Decision tree is used to divide the big dataset into smaller parts. The final output of this system will be the disease predicted by the model.

5. DATASET AND MODEL DESCRIPTION

Dataset used in this system is in a structured format. The dataset which is used contains the disease name with its all symptoms. As our system is based on supervised learning machine algorithms, the dataset is having the label with 0 or 1. Then we divide the dataset into a Training dataset and Testing dataset. The model is trained by a training dataset. All algorithms were applied to this training dataset and then the machine learning model is trained. Then the testing dataset was provided to the trained model to test the accuracy of the model.

5.1 DATASET OF HOSPITAL

The hospital data will be in the form of structural format. The dataset used in this project is real life data. The structural data contains symptoms of patients. Any dataset is converted into either 0 or 1. Zero value represents feature/symptom impacts on disease and value one represents that it does not impact on disease.

6. EVALUATION METHOD

To calculate performance evaluation in the experiment, first, we denote TP, TN, Fp and FNias 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 measures 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

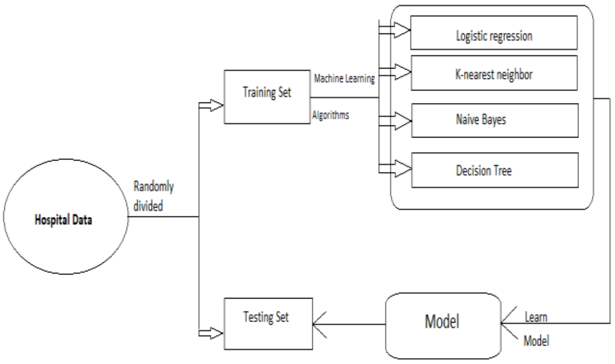


Fig -1: System Architecture

8. ALGORITHM

8.1 KNN

K Nearest Neighbor (KNN) could be terribly easy, simple to grasp, versatile and one amongst the uppermost machine learning algorithms. In the Healthcare System, the user will predict the disease. In this system, the user can predict whether the disease will detect or not. In the proposed 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 issue. KNN algorithm is based on feature similarity approach.

It is the best choice for addressing some of the classification related tasks. K-nearest neighbor classifier algorithm is to predict the target label of a new instance by defining the nearest neighbor class. The closest class will be identified using distance measures like Euclidean distance. If K = 1, then the case is just assigned to the category of its nearest neighbor.

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

The value of 'k' has to be specified by the user and the best choice depends on the data. The larger value of 'k' reduces the noise on the classification. If the new feature i.e in our case symptom has to classify, then the distance is calculated and then the class of feature is selected which is nearest to the newer instance. 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.

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

8.2 NAIVE BAYES

Naive Bayes is an easy however amazingly powerful rule for prognosticative modeling. The independence assumption that allows decomposing joint likelihood into a product of marginal likelihoods is called as 'naive'. This simplified Bayesian classifier is called as naive Bayes. The Naive Bayes classifier assumes the presence of a particular feature in a class is unrelated to the presence of any other feature. It is very easy to build and useful for large datasets. Naive Bayes is a supervised learning model.

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.

In our system, Naïve Bayes decides which symptom is to put in classifier and which is not.

8.3 LOGISTIC REGRESSION

Logistic regression could be a supervised learning classification algorithm accustomed to predict the chance of a target variable that is Disease. The nature of the 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(y=1) as a function of x.

Logistic regression can be expressed as:
log(p(X)/(1-p(X)))=β_0+β_1 X

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

8.4 Decision Tree

A decision tree is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision tree.

With each successive division, the members of the resulting sets become more and more similar to each other. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous (mutually exclusive) groups with respect to a particular target.

The target variable is usually categorical and the decision tree is used either to:

- Calculate the probability that a given record belong to each of the category and,
- To classify the record by assigning it to the most likely class (or category).

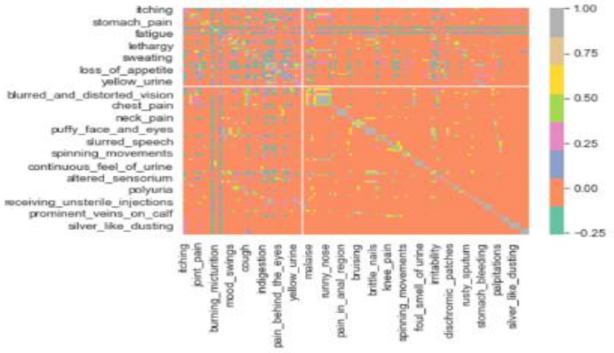
In this disease prediction system, decision tree divides the symptoms as per its category and reduces the dataset difficulty.

9. CONCLUSION

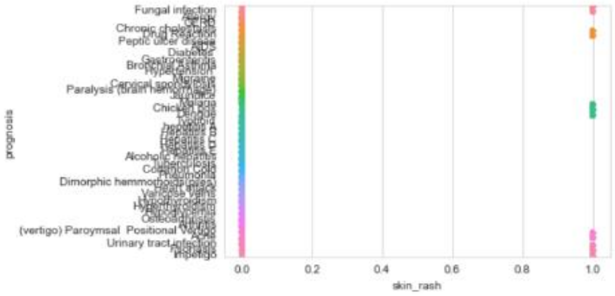
The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained. Disease Predictor was successfully implemented using the grails framework. This system gives a user-friendly environment and easy to use. As the system is based on the web application, the user can use this system from anywhere and at any time.

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

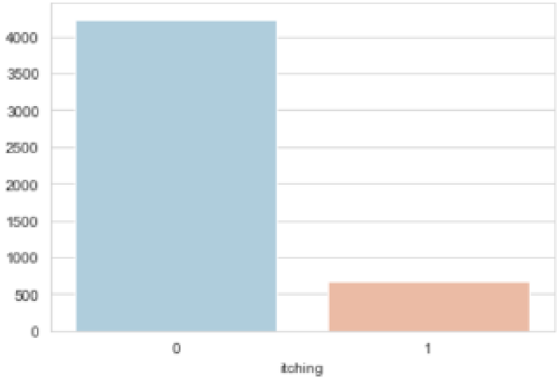
10 PREDICTION GRAPH



Heat-Map



Swarmp Plot



Count Plot

REFERENCES

[1] D. Kornack and P. Rakic, “Cell Proliferation without Neurogenesis in Adult Primate Neocortex,” Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[2] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[3] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] K. Elissa, “Title of paper if known,” unpublished.