IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10663

Chronic Disease Prediction Using Machine Learning

Kaushik Kulkarni[1], Manjunath B[2], Mayur Hebbar T M[3], Meghana M[4], Shashank S[5], Tojo Mathew[6]

Student, Computer Science and Engineering, The National Institute of Engineering, Mysuru, India[1, 2, 3, 4]

Assistant Professor, Computer Science and Engineering, The National Institute of Engineering, Mysuru, India[5, 6]

Abstract: Technological development, including machine learning, has a huge impact on health through an effective analysis of various chronic diseases for more accurate diagnosis and successful treatment. In the field of biomedical and healthcare communities the accurate prediction plays the major role to find out the risk of the disease in the patient. The only way to overcome with the mortality due to chronic diseases is to predict it earlier so that the disease prevention can be done. Such model is a Patient's need in which Machine Learning is highly recommendable. But the precise prediction on the basis of symptoms becomes too difficult for doctor. The correct prediction of disease is the most stretching task. To overcome this problem data mining plays an important role to predict the disease. This study analyzes chronic diseases using machine learning techniques based on a chronic diseases dataset from the UCI machine learning data warehouse. We use Heart disease, Kidney disease, Cancer disease and Diabetes disease datasets, In order to build reliable prediction models for these chronic diseases using data mining techniques. The most relevant features are selected from the dataset for improved accuracy and reduced training time. 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 Logistic Regression. By using logistic regression, random forest and decision tree we are predicting diseases like Diabetes, Heart, Cancer and Kidney. For each chronic disease, diverse models, techniques, and algorithms are used for predicting and analyzing. The paper comprises a conceptual model that integrates the prediction of most common chronic diseases.

Keywords: Logistic Regression, Chronic Diseases, Machine Learning, Diseases Prediction and Accuracy.

I. 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. ML is categorized as supervised (i.e., consists of output variables that are predicted from input variables) or unsupervised (i.e., deals with clustering of different groups for a particular intercession). ML is used to determine complex models, and extract medical knowledge, exposing novel ideas to professionals, and specialists. In clinical practice, ML predictive models can highlight strengthen rules in the decision-making regarding individual patient care. These are also capable of autonomous diagnosis of different diseases under clinical rules. The incorporation of these models in drug prescription can save doctors and offer new medical opportunities in identification.

Machine learning has been shown to be effective in supporting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We streamline machine learning algorithms for effective prediction of chronic disease breakout. Various studies give only a glimpse into predicting disease with ML techniques. We propose a novel method that aims at finding significant features by applying machine learning techniques such as K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Logistic Regression, Random Forest and Naive Bayes (NB) resulting in improving the accuracy in the prediction of disease. Multiple such algorithms are carried out to improve the accuracy of the learning process. It can then be tested with the available datasets. The prediction model is introduced with different combinations of features and various known classification techniques.

With ML models, it can also be possible to improve quality of medical data, reduce variation in patient rates, and save in medical costs. Therefore, these models are frequently used to investigate diagnostic analysis when compared with other conventional methods. To reduce the death rates caused by chronic diseases (CDs), early detection and effective therapy are the only solutions. Therefore, most medical scientists are attracted to the new technologies of predictive models in disease estimation. These new advancements in medical care have been spreading the accessibility of electronic data and opening new doors for decision support and productivity improvements. ML methods have been effectively utilized in the computerized elucidation of pneumonic capacity tests for the differential analysis of CDs. It is expected that the models with the highest accuracies could gain large importance in medical diagnosis.

II. LITERATURE SURVEY

The name machine learning was coined in 1959 by Arthur Samuel. Tom Mitchell states machine learning as “Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience”. It is a combination of correlations and relationships. Most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. There are various types of algorithms such as Linear Regression, Logistic Regression, Naive Bayes Classifier, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, SVM (Support Vector Machines), K-means Algorithm, Random Forest etc.

Machine learning examine the study and construction of algorithms that can learn from and make predictions on data. It is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties with mathematical optimization which delivers methods, theory and application domains to the field. Machine learning is sometimes merged with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning.

Machine learning tasks are typically classified into several broad categories:

A. *Supervised learning*:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from marked training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

B. *Unsupervised learning*:

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human observation. In similar to supervised learning that usually makes use of human-labeled data, unsupervised learning, also known as self-organization allows for modeling of probability densities over inputs.

III. CURRENT PROCESS

Chronic diseases are growing to be one of the prominent causes for deaths worldwide. There is an increasing percentage of the world population facing the adverse health effects of living. In general, the patient's reports have to be carefully scrutinized by doctors to make a diagnosis of the disease. Since the diagnosis is manual sometimes it is difficult for the doctors to treat patients efficiently.

The number of people suffering from Chronic Diseases is rising day by day. The conventional Health Care is passive. Due to this type, patients can die due to a lack of proper treatment during emergencies such as cardiac arrest. The key to improving Health Care efficiency is to reduce the mortality rate due to lack of proper treatment and to transform the passive Health Care program into a continuous one at a reduced cost.

IV. PROPOSED SYSTEM

Due to the low-progress nature of Chronic Diseases, it is important to make an early prediction and provide effective medication. Therefore, it is essential to propose a decision model which can help to diagnose chronic diseases and predict future patient outcomes. While there are many ways to approach this in the field of AI, the present study focuses distinctly on ML predictive models used in the diagnosis of Chronic Diseases. In comparison to the conventional data analysis techniques, we will be able to find promising results that enhance the quality of patient data and inspect of specific items that are related to ML algorithms in medical care.

The main purpose of our project is to make hospital tasks easy and to develop an efficient and feasible software that replaces the manual prediction system into an automated healthcare management system. Our project enables healthcare providers to improve operational effectiveness, reduce medical errors and time consumption. If disease can be predicted, then early treatment can be given to the patients which can reduce the risk of life and save life of patients. The cost to get treatment of diseases can also be reduced up to an extent by early recognition.

The diagnose will be done based on various Classification Machine Learning Models such as,

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 6, June 2021

DOI 10.17148/IJARCCE.2021.10663

- Logistic Regression,
- Naive Bayes Classification and
- KNN algorithm.

V. SYSTEM DESIGN

A. *Design Goals*

The design goals consist of various designs which we have implemented in our system "Chronic Disease Prediction Using Machine Learning". This system is built with various designs such as data flow diagram, sequence diagram, class diagram, use case diagram, activity diagram.

We have designed our system in such a way that the registration process is solely done by administrator. After the registration process, the users i.e. doctors can login into the system using their credentials. Based on the inputs/attributes given, doctors will be able to predict the chronic disease accordingly

B. *System Architecture*

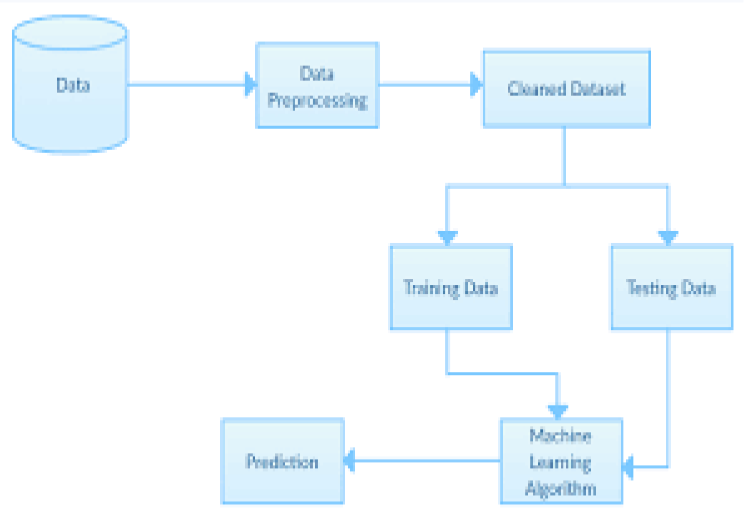

Figure 1. System Architecture

An architecture diagram is a graphical representation of a set of concepts that are part of architecture, including their principles, elements and components. The diagram explains about the system software in perception of overview of the system

C. *Activity Diagram*

The activity diagram is presented in Fig 2. It represents the order in which a particular task of the system is performed to obtain the result.
The registration process of a User/Doctor is carried out by the Administrator. After the registration, the user i.e. doctor will login to the system using the credentials provided by the admin. Once the user successfully logs in, the system will take him to the desired page based on the specialization. Here, in order to get the desired prediction, the user has to enter the attributes (independent variables) accordingly. System uses the Machine Learning Model that is built using available datasets and various ML algorithms (classification algorithms) to generate the desired predictions and visualization.

DOI 10.17148/IJARCCE.2021.10663

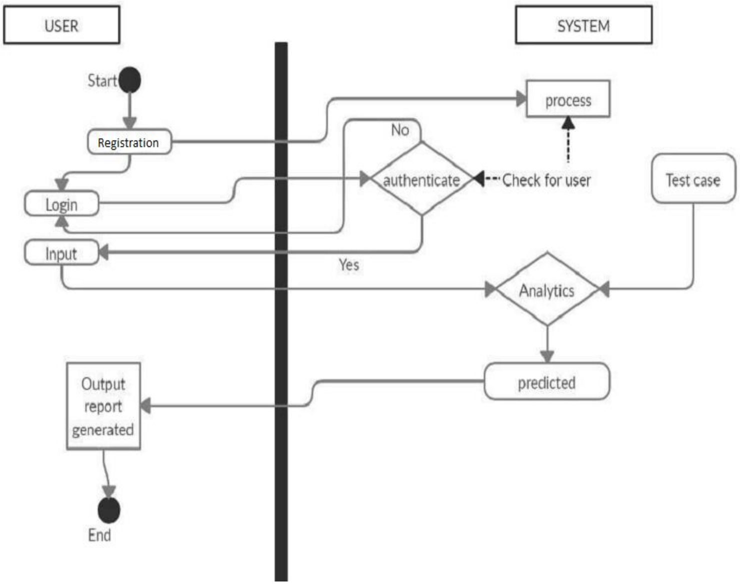

Figure 2. System Architecture

VI. ALGORITHM

A. *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 comparable approach.

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

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

It ought to even be noted that every one 3 distance measures square measure solely valid form continuous variables. In the instance of categorical variables, the Hamming distance must be used. It combinedly 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.

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

B. *Naïve Bayes*

Naive Bayes is an 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:

$$\text{P (bVa)} = \frac{P(aVb)\,P(b)}{P(a)}$$

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.

C. *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 simple 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.

VII. RESULTS AND DISCUSSION

The metrics provided below gives us information on the quality of the results that we get in this study.

Precision: Precision or positive predictive value here is the ratio of all patients actually with chronic diseases to all the patients predicted with chronic disease (true positive and false positive).

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: It is also known as sensitivity and it is the ratio of actual number of chronic diseases patients that are correctly identified to the total no of patients with chronic diseases.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F- Measure: It measures the accuracy of the test. It is the harmonic mean between precision and recall.

$$\text{F - Measure}=2 * \frac{Recall*Precision}{Recall+Precision}$$

Accuracy: It is the ratio of correctly predicted output cases to all the cases present in the data set.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

Table I: Result on accuracy with correctly and incorrectly classified instances.

Disease	Accuracy	Correctly Classified Instances	Incorrectly Classified Instances
Cancer	81.8182	108	24
Heart	99.8243	568	1

Diabetes	78.9272	206	55
Kidney	77.2121	309	91

Table II: Result of Precision, Recall and F – Measure.

Disease	Precision	Recall	F - Measure
Cancer	1.000	0.333	0.500
Heart	1.000	0.995	0.998
Diabetes	0.812	0.899	0.853
Kidney	0.857	0.666	0.749

VIII. CONCLUSION AND FUTURE SCOPE

Machine Learning has brought major improvements to the healthcare sector. With the aid of Machine Learning, the difficult and life-critical tasks such as chronic disease diagnosis are made easy and reliable. It has brought about revolutionary changes in hospital, clinic, and laboratory procedures. By analyzing historical and real-time data, doctors can predict the future situation of patients. Our method has been tested with various datasets for Heart, Kidney, Cancer and Diabetes disease. The main objective of this study was to predict the chronic disease using attributes while maintaining a higher accuracy (here we obtain an accuracy of about 90%). Also, our model generates the report consisting of possibilities of occurrence of disease. The results demonstrate the robustness of the approach proposed. Future research should analyze different supervised and unsupervised machine learning technique with additional performance metrics for better chronic disease prediction.

REFERENCES

[1] Hamet P., Tremblay J. Artificial intelligence in medicine. Metabolism. 2017; 69:S36–S40. doi: 10.1016/j.metabol.2017.01.011. [PubMed] [CrossRef] [Google Scholar].

[2] Johnson K.W., Soto J.T., Glicksberg B.S., Shameer K., Miotto R., Ali M., Dudley J.T. Artificial intelligence in cardiology. J. Am. Coll. Cardiol. 2018; 71:2668–2679. doi: 10.1016/j.jacc.2018.03.521. [PubMed] [CrossRef] [Google Scholar].

[3] Bini S. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J. Arthroplast. 2018; 33:2358–2361. doi: 10.1016/j.arth.2018.02.067. [PubMed] [CrossRef] [Google Scholar].

[4] Kotsiantis S.B., Zaharakis I., Pintelas P. Supervised machine learning: A review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. 2007; 160:3–24. [Google Scholar].

[5] Deo R.C. Machine Learning in Medicine. Circulation. 2015; 132:1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593. [PMC free article] [PubMed] [CrossRef] [Google Scholar].

[6] Battineni G., Sagaro G.G., Nalini C., Amenta F., Tayebati S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. Machines. 2019; 7:74. doi: 10.3390/machines7040074. [CrossRef] [Google Scholar].

[7] Polat H., Mehr H.D., Cetin A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. J. Med. Syst. 2017; 41:55. doi: 10.1007/s10916-017-0703-x. [PubMed] [CrossRef] [Google Scholar].