International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X doi : https://doi.org/10.32628/IJSRSET23103189

Performance Evaluation of Various Machine Learning Algorithms in Prediction of Chronic kidney disease (CKD)

Sandhu Sarojini[1], S. Akhilesh[2]

[1]M.Tech Student, Department of Computer Science And Engineering, Priyadarshini Institute of Technology and Sciences for Women, Guntur (Dt), , Andhra Pradesh, India

[2]Assistant Professor, Department of Computer Science And Engineering, Priyadarshini Institute of Technology and Sciences for Women, Guntur (Dt), , Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 10 June 2023
Published: 04 July 2023

Publication Issue
Volume 10, Issue 4
July-August-2023

Page Number
23-29

ABSTRACT

With a high rate of morbidity and mortality as well as the ability to spread other diseases, chronic kidney disease (CKD) is a major worldwide health concern. Patients sometimes overlook the disease in the early stages of CKD since there are no evident symptoms. Early diagnosis of CKD enables patients to receive effective treatment in time to slow the disease's progression. Due to their quick and precise detection capabilities, machine learning models can help therapists accomplish this goal efficiently. In this research, we suggest a machine learning approach to CKD diagnosis. The website KAGGLE provided the CKD data set, which has a significant number of missing values.. The mean value is used to fill in the blanks; for object data types (strings), we utilized the most frequent object (string) to replace the missing values. Since patients may overlook particular measurements for a variety of reasons, missing values are typically observed in real-world medical scenarios. Four machine learning algorithms—Logistic Regression, SVM, Random Forest Classifier, and Decision Tree Classifier—were applied to create models after successfully completing the incomplete data set. Random Forest has the highest accuracy of these machine learning models.

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

I. INTRODUCTION

Programming computers to perform better using example data or historical data is known as machine learning. The study of computer systems that learn from data and experience is known as machine learning. ML can be classified as either unsupervised (which deals with clustering of various groups for a specific function) or supervised (which uses output variables that are predicted from input factors). With

the use of machine learning (ML), complicated models can be determined and medical knowledge extracted, presenting new concepts to experts. In clinical practice, machine learning (ML) predictive models can emphasize and enhance rules in the selection of specific patient treatment options. These are also capable of independently diagnosing various disorders in accordance with clinical guidelines. The use of these models in drug prescription can save medical costs and provide new potential for discovery.

Machine learning has proven to be successful in helping with decision-making and prediction from the massive amount of data generated by the healthcare sector. We simplify machine learning techniques for accurate chronic illness outbreak prediction. Only a few studies have looked into using ML to predict disease. By utilizing machine learning techniques including the K-Nearest Neighbour Algorithm (KNN), Decision Trees (DT), Logistic Regression, Random Forest, and Naive Bayes (NB), we offer a novel strategy that tries to uncover meaningful features with the goal of increasing the accuracy of disease prediction. To increase the learning process' accuracy, multiple such algorithms are used. Then, using the datasets at hand, it may be tested. A variety of combinations are used to introduce the prediction model.

With ML models, it may be feasible to decrease patient rate variation, enhance the quality of medical data, and lower medical expenses. In contrast to other traditional techniques, these models are thus commonly utilized to examine diagnostic analyses. The only ways to lower the death rates brought on by chronic illnesses (CDs) are early detection and efficient treatment. Therefore, the majority of medical scientists are drawn to the new technologies of disease estimate using predictive models. These recent developments in healthcare have increased electronic data accessibility, opened up new avenues for decision support, and increased productivity. The computerized explanation of pneumonic capacity tests for the differential analysis of CDs has successfully applied ML approaches. We anticipate 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

The prevalence of chronic diseases as a cause of death is rising globally. A growing portion of the global population is dealing with the negative health repercussions of living. In general, doctors must thoroughly review the patient's records in order to diagnose the ailment. The manual nature of diagnosis can occasionally make it challenging for doctors to effectively treat patients. Chronic disease patients are becoming more numerous every day. Traditional healthcare is apathetic. Because of this kind, people may pass away from improper care during emergencies like cardiac arrest. Reducing the mortality rate due to improper care and transforming health care delivery systems are the keys to enhancing health care efficiency and to transform the passive Health Care program into a continuous one at a reduced cost.

IV. PROPOSED SYSTEM

Because chronic diseases have a slow rate of progression, it's critical to provide timely diagnosis and efficient treatment. Therefore, it is crucial to suggest a decision model that can aid in the diagnosis of chronic diseases and forecast patient outcomes in the future. In the realm of AI, there are numerous methods to go about doing this, however the current work specifically focuses on ML prediction models employed in the diagnosis of Chronic Diseases. We will be able to identify promising outcomes that improve the quality of patient data and inspect particular items that are related to ML algorithms in medical treatment when compared to the traditional data analysis methods. Our project's major goal is to produce software that will simplify hospital tasks and replace the manual prediction system with an automated healthcare management system. With the help of our initiative, healthcare providers can operate more efficiently while consuming less time and lessening medical errors. If an illness can be predicted, early treatment can be offered to the patients, lowering the risk of death and potentially saving their lives. Early detection of diseases can also help to some degree lower the cost of medical care.

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

- 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

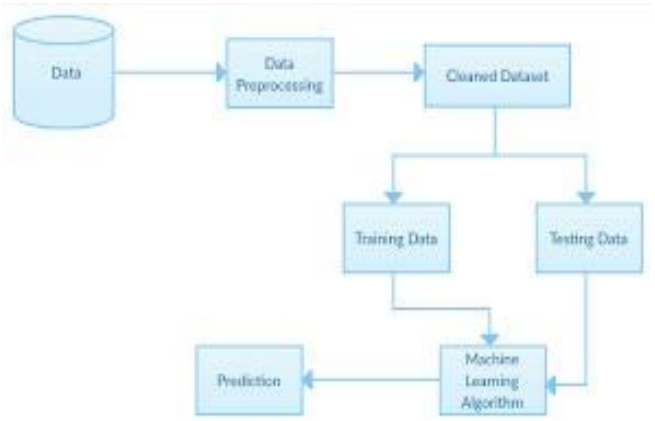

Figure 1. System Architecture

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.

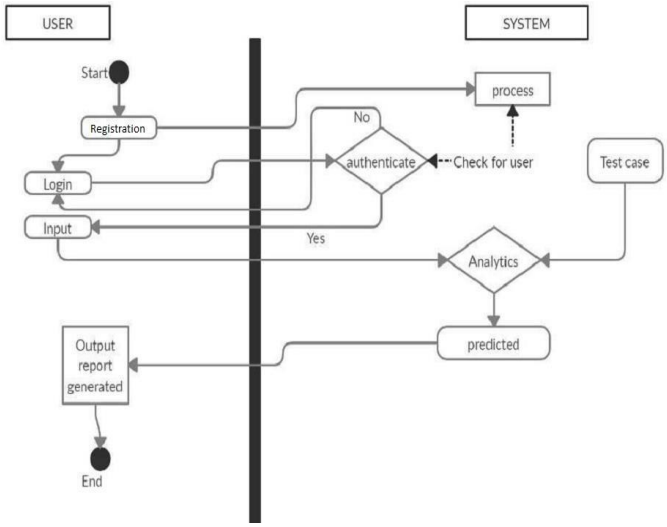

Figure 2. System Architecture

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 nearestneighbor.

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

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 categoricalvariables within the dataset.

A. Naïve Bayes

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

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 ourprevious 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:

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

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.

A. *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 (standsfor 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 inputsare mapped to the log (odds) - the output being adequate to 1.

B. *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 (standsfor 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 inputsare mapped to the log (odds) - the output being adequate to 1.

VI. 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 correctlyidentified 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 ClassifiedInstances
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

VII. CONCLUSION AND FUTURE SCOPE

The healthcare industry has seen significant advancements because to machine learning. The complex and life-threatening jobs, like diagnosing chronic diseases, are made simple and accurate with the help of machine learning. The techniques used in hospitals, clinics, and laboratories have undergone revolutionary modifications as a result. Doctors can forecast their patients' future conditions by analyzing previous and current data. Various datasets for heart, kidney, cancer, and diabetes disorders have been used to test our technique. The major goal of this study was to use qualities to predict chronic disease with a greater level of accuracy (in this case, we achieve an accuracy of nearly 90%). Additionally, our model creates a report with potential disease occurrences. The outcomes show how reliable the suggested strategy is. Future studies should

examine. Future research should analyze different supervised and unsupervised machine learning technique with additional performance metrics for better chronic disease prediction.

VIII. 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].

[8]. Pascanu, Razvan and Bengio, Yoshua. Revisiting natural gradient for deep networks.arXiv preprint arXiv:1301.3584, 2013.

[9]. Tieleman, T. and Hinton, G. Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning.

[10]. Technical report, 2012.

[11]. Zeiler, Matthew D. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.

[12]. Kevin Bache and Moshe Lichman. 2013. UCI machine learning repository. (2013).

Cite this article as :

Sandhu Sarojini, S. Akhilesh, "Performance Evaluation of Various Machine Learning Algorithms in Prediction of Chronic kidney disease (CKD)", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 4, pp. 23-29, July-August 2023. Available at doi : https://doi.org/10.32628/IJSRSET23103189

Journal URL : https://ijsrst.com/IJSRSET23103189