Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Chronic Kidney Disease Prediction using Machine Learning

Reshma S[1], Salma Shaji[2], S R Ajina[3], Vishnu Priya S R[4], Janisha A[5]
[1,2,3,4,5]Dept of Computer Science and Engineering
[1,2,3,4,5]LBS Institute Of Technology For Women, Thiruvananthapuram, Kerala

Abstract: **Chronic Kidney Disease also recognized as Chronic Renal Disease, is an uncharacteristic functioning of kidney or a failure of renal function expanding over a period of months or years. Habitually, chronic kidney disease is detected during the screening of people who are known to be in threat by kidney problems, such as those with high blood pressure or diabetes and those with a blood relative Chronic Kidney Disease(CKD) patients. So the early prediction is necessary in combating the disease and to provide good treatment. This study proposes the use of machine learning techniques for CKD such as Ant Colony Optimization(ACO) technique and Support Vector Machine(SVM) classifier. Final output predicts whether the person is having CKD or not by using minimum number of features.**

Keywords: Chronic kidney, SVM, Ant colony optimization

I. INTRODUCTION

Chronic Kidney Disease (CKD) is considered as an important threat for the society with respect to the health in the present era. Chronic kidney disease can be detected with regular laboratory tests, and some treatments are present which can prevent development, slow disease progression, reduce complications of decreased Glomerular Filtration Rate(GFR) and risk of cardiovascular disease, and improve survival and quality of life. CKD can be

caused due to lack of water consumption, smoking, improper diet, loss of sleep and many other factors. This disease affected 753 million people globally in 2016 in which 417 million are females and 336 million are males. Majority of the time the disease is detected in its final stage and which sometimes leads to kidney failure.

The existing system of diagnosis is based on the examination of urine with the help of serum creatinine level. Many medical methods are used for this purpose such as screening, ultrasound method. In screening, the patients with hypertension, history of cardiovascular disease, disease in the past, and the patients who have relatives who had kidney disease are screened. This technique includes the calculation of the estimated GFR from the serum creatinine level, and measurement of urine albumin-to-creatinine ratio (ACR) in a first morning urine specimen. This paper focuses on machine learning techniques like ACO and SVM by minimizing the features and selecting best features to improve the accuracy of prediction.

II. LITERATURE REVIEW

[J. Snegha, 2020][10] proposed a system that uses various data mining techniques like Random Forest algorithm and Back propagation neural Network. Here they compare both of the algorithm and found that Back Propagation algorithm gives the best result as it uses the supervised learning network called feedforward neural network.

[Mohammed Elhoseny, 2019] described a system for CKD in which it uses Density based feature selection with ACO. The system uses wrapper methods for feature selection.

[Baisakhi Chakraborty, 2019][9] proposed development of CKD prediction system using machine learning techniques such as K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine and Multi-Layer Perceptron Algorithm. These are applied and their performance are compared to the accuracy, precision, and recall results. Finally, Random forest is chosen to implement this system.

[Arif-Ul-Islam, 2019] proposed a system in which prediction of disease is done using Boosting Classifiers, Ant-Miner and J48 Decision Tree. The aim of this paper is two fold that is, analyzing the performance of boosting algorithms for detecting CKD and deriving rules illustrating relationships among the attributes of CKD. Experimental results prove that the performance of AdaBoost was less that of LogitBoost by a fraction.

[S.Belina V, 2018] proposed a system that uses extreame learning machine and ACO for CKD prediction. Classification is done using MATLAB tool and ELM has few constraints in the optimization. This technique is an improvement under the Sigmoid additive type of SLFNs.

[Siddheshwar Tekale, 2018][8] described a system using machine learning which uses Decision tree SVM techniques. By comparing two techniques finally concluded that SVM gives the best result. Its prediction process is less time consuming so that doctors can analyze the patients within a less time period.

[Nilesh Borisagar, 2017] described a system which uses Back Propagation Neural Network algorithm for prediction. Here Levenberg, Bayesian regularization, Scaled Conjugate and resilient back propagation algorithm are discussed. Matlab R2013a is used for the implementation purpose. Based on the training time, scaled conjugate gradient and resilient back propagation are found more efficient than Levenberg and Bayesian regularization.

[Guneet Kaur, 2017][7] proposed a system for predicting the CKD using Data Mining Algorithms in Hadoop. They use two data mining classifiers like KNN and SVM. Here the predictive analysis is performed based upon the manually selected data columns. SVM classifier gives the best accuracy than KNN in this system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

[Neha Sharma, 2016] proposed a system in which the kidney disease of a patient is analyzed and the results are to compute automatically using the data set of the patient. Here Rule based prediction method is used. This system uses neuro-fuzzy method and obtained the outcome by mathematical computation.

[Kai-Cheng Hu, 2015][6] proposed a system which uses a multiple pheromone table based on ACO for clustering. Here they divided the problem into a set of several different patterns based on their features. Two pheromone tables are used here one for keeping the track of the promising information and the other to hold the details of unpromising information which in turn increases the probability of searching directions.

III. DATASET AND METHODS

A. *Dataset*

The Dataset here we use is the publically available CKD Dataset from UCI repository. It contains 400 samples of two different classes. Out of 25 attributes, 11 are numeric and 13 are nominal and one is class attribute. The data set contains number of missing values. Here the information of dataset uses the patient's data like age, blood pressure, specific gravity, albumin, sugar, red blood cells etc.

Table.1 List of attributes present in the CKD dataset

Attributes	Type
Age	Numeric
Blood Pressure	Numeric
Specific Gravity	Numeric
Albumin	Numeric
Sugar	Numeric
Red Blood Cells	Nominal
Pus Cell	Nominal
Pus Cell clumps	Nominal
Bacteria	Nominal
Blood Glucose Random	Numeric
Blood Urea	Numeric
Serum Creatinine	Numeric
Sodium	Numeric
Potassium	Numeric
Hemoglobin	Numeric
Packed Cell Volume	Numeric
Red Blood Cell count	Numeric
White Blood Cell Count	Numeric
Hypertension	Nominal
Diabetes Mellitus	Nominal
Coronary Artery Disease	Nominal
Appetite	Nominal
Pedal Edema	Nominal
Anemia	Nominal
Class	Class

CKD is caused due to diabetes and high blood pressure. Due to Diabetes our many organs get affected and it will be followed by high blood sugar. So it is important to predict the disease as early as possible. This study improvises some of the machine learning techniques to predict the disease.

B. *Steps*

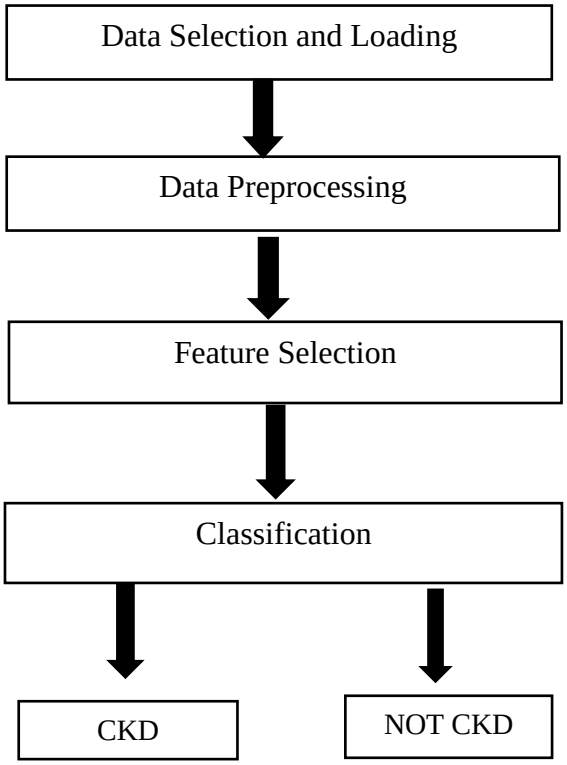

Figure.1 Flowchart of the proposed system

a. Pre-Processing

Data Pre-Processing is that stage where the data that is distorted, or encoded is brought to such a state that the machine can easily analyze it. A dataset can be observed as a group of data objects. Data objects are labeled by a number of features, that ensures the basic features of an object, such as the mass of a physical object or the time at which an event ensured. In the dataset there may be missing values, they can either eliminated or estimated. The most common method of dealing with missing values is filling them in with mean, median or mode value of respective feature. As object values cannot be used for the analysis we have to convert the numeric values with type as object to float64 type. Null values in the categorical attributes are changed with the most recurrent occurring value current in that attribute column. Label encoding is done to translate categorical attributes into numeric attribute by conveying each unique attribute value to an integer. This automatically changes the attributes to int type. The mean value is premeditated from each column and is used to replace all the missing values in that attribute column. For this function we are using a function called imputer which is used to find the mean value in each column. After the replacing and encoding is done, the data should be trained, validated and tested. Training the data is the part on which our algorithms are actually trained to build a model. Validation is the part of the dataset which is used to validate our various model fits or improve the model. Testing the data is used to test our model hypothesis.

b. Feature Selection

Feature Selection is the method where we computationally select the features which contribute most to our prediction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

variable or output. In this study we used Ant Colony Optimization (ACO) for selecting the best features from the dataset. It is a technique for solving computational problems which can be condensed to finding good paths through graphs. Artificial Ants stand for multi-agent methods enthused by the behavior of real ants. The pheromone–based communication of biological ants is often the main paradigm used. Combinations of Artificial Ants and local search algorithms have become a method of choice for numerous optimization tasks involving some sort of graph. This algorithm evaluates the intensity of pheromone during each iteration rather than accumulating them. The proposed algorithm will change a small number of features in subsets which are selected by choosing the best ants. A classification algorithm has to be used to evaluate the performance of the subsets that is wrapper evaluation function.

ACO: To apply an ant colony algorithm, the optimization problem needs to be transformed into the problem of finding the shortest path on a weighted graph. In the first step of each iteration, each ant stochastically builds a solution, i.e. the order in which the edges in the graph should be followed. In the second step, the paths found by the different ants are equated. The last step consists of updating the pheromone levels on each edge. Each ant needs to construct a solution to move through the graph. To select the next edge in its tour, an ant will consider the length of each edge available from its present position, as well as the consistent pheromone level.

The package used for Ant Colony Optimization is ACO Pants. Using pants we can easily determine the way to visit the interconnected nodes to minimize the path. Nodes represents data and edges represents the work done to travel from one node to another. Using the list of nodes and a function returning the length of the edge between any two given nodes. It may not provide the actual length of the path. Here length refers to the amount of work for moving between the nodes. Iterative process is to be done to obtain the solution. In each iteration, several ants traverse through the path covering each and every node to find a solution. The amount of pheromone is updated on each edge according to the length of the solution used. The local best solution is estimated as the ant that traversed through the least distance. Each local best solutions are recorded. If the local solution has least distance compared to that of the best from any of the previous iterations, it is then considered as the global best solution. The best ant thus found then deposits its pheromone on the global best solution path so as to strengthen the path more. This process is done repeatedly.

c. Classification

For classification we use Support Vector Machine(SVM) to predict the disease and its performance. As a first step we have to import the libraries for classification and prediction. We import SVM and datasets from the scikit-learn library. NumPy for carrying out efficient mathematical computations. Accuracy-score from sklearn.metrics to predict the accuracy of the model. We have divided the data into training and testing sets. Now is the time to train our SVM on the training data. scikit-learn contains the SVM library, which contains built-in classes for various SVM algorithms. Since we are going to perform a classification task, we will use the support vector classifier class, which is written as SVC in the scikit-learn's SVM library. This class takes one parameter, which is the kernel form. The fit method of SVC class is called to train the algorithm on the training data, which is passed as a parameter to the fit method. To make predictions, the predict method of the SVC class is used. For evaluating the algorithm, we use the confusion matrix.

SVM: In machine learning, Support Vector Machine (SVM) are supervised learning models with related learning algorithms that examine data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM works by mapping data to a high-dimensional feature space so that data points can be classified, even when the data are not otherwise linearly separable.

IV. RESULTS AND DISCUSSION

The metrics provided below gives us information on the quality of the outcomes that we get in this study. A confusion matrix helps us with this by describing the performance of the classifier.

Table.2 Confusion Matrix

Confusion Matrix	CKD (Predicted)	Not CKD (Predicted)
CKD(Actual)	TP	FN
Not CKD (Actual)	FP	TN

Precision: Precision or positive predictive value here is the ratio of all patients actually with CKD to all the patients predicted with CKD (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 CKD patients that are correctly identified to the total no of patients with CKD.

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

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

$$\text{F-Measure}=2*\frac{\text{Re}\,call*\text{Pr}\,ecision}{\text{Re}\,call+\text{Pr}\,ecision}$$

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

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

Support: Support is the correct number of outcomes or responses that are present in each class of the predicted outcome.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Table.3 Results of CKD prediction using ACO and SVM

	PRECISION	RECALL	F1-SCORE	SUPPORT
CKD	1.00	0.94	0.97	83
Not CKD	0.88	1.00	0.94	37
Accuracy			0.96	120
Macro-Average	0.94	0.97	0.95	120
Weighted Average	0.96	0.96	0.96	120

V. CONCLUSION

This paper deals with the prediction of CKD in people. A wrapper method used here for feature selection is ACO. ACO is a meta-heuristic optimization algorithm. Out of the 24 attributes present 12 best attributes are taken for prediction. Prediction is done using the machine learning technique, SVM. In this classification problem SVM classifies the output into two class with CKD and without CKD. The main objective of this study was to predict patients with CKD using less number attributes while maintaining a higher accuracy. Here we obtain an accuracy of about 96 percentage.

REFERENCES

[1] Hussein Abbass, “Classification Rule Discovery with Ant Colony Optimization”, Research Gate Article, 2004
[2] Mohammed Deriche, “Feature Selection using Ant Colony Optimization”, International Multi-Conference on Systems, Signals and Devices, 2009
[3] X. Yu and T. Zhang, ”Convergence and runtime of an Ant Colony Optimization”, Information Technology Journal 8(3) ISSN 1812-5638, 2009
[4] David Martens, Manu De Backer, Raf Haesen, “Classification with Ant Colony Optimization”, IEEE Transactions on evolutionary computation, Vol.11, No.5, 2010.
[5] Vivekanand Jha, ”Chronic Kidney Disease Global Dimension and Perspectives”, Lancet, National Library of Medicine, 2013
[6] Kai-Cheng Hu, “Multiple Pheromone table based on Ant Colony Optimization for Clustering”, Hindawi, Research article, 2015.
[7] Guneet Kaur, “Predict Chronic Kidney Disease using Data Mining in Hadoop, International Conference on Inventive Computing and Informatics, 2017.
[8] Siddeshwar Tekale, “Prediction of Chronic Kidney Disease Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering, 2018.
[9] Baisakhi Chakraborty, “Development of Chronic Kidney Disease Prediction Using Machine Learning”, International Conference on Intelligent Data Communication Technologies, 2019.
[10] J. Snegha, “Chronic Kidney Disease Prediction using Data Mining”, International Conference on Emerging Trends, 2020.