IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101266

Prior Stage Kidney Disease Prediction Using AI & Supervised Machine Learning Techniques

Barot mitisha[1], prof. barkha bhavsar[2]

Student, Master Of Computer Engineering, LDRP Institute Of Technology & Research Center, Gandhinagar, Gujarat[1]

Assistant Professor, Engineering, LDRP, Gandhinagar, Gujarat[2]

Abstract: Chronic kidney disease, also known as chronic kidney disease, is a featureless disorder of kidney function or kidney function that lasts months or years. Chronic kidney disease is usually found by screening people who are known to be at risk for kidney problems such as: Therefore, early prediction is needed to combat illness and provide good treatment. This study suggests the use of CKD machine learning techniques such as KNN, DT, NB, and SB classifiers.

Keywords: Chronic kidney, KNN, DT, NB, and SB classifiers

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.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101266

[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.

[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.

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

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

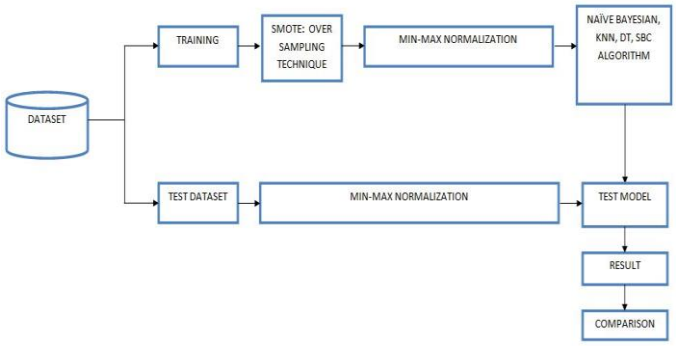
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101266

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



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. KNN

In pattern recognition, the K-Nearest Neighbor algorithm (K-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In KNN Classification, the output is a class membership. Classification is done by a majority vote of neighbours. If K = 1, then the class is single nearest neighbor. In a common weighting scheme, individual neighbour is assigned to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a straight line and the distance is known as Euclidean distance

[7]. The limitation of the K-NN algorithm is it's sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as Feature extraction. In Feature space, extraction is taken place on raw data before applying K-NN algorithm. The steps involved in a K-NN algorithm

NB: Naive Bayes is a classification method based upon Bayes Theorem which computes the likelihood for every attribute. It selects the outcome with highest probability. This classifier assumes the features are independent and that the existence of a specific feature in a class is not linked to the existence of any other feature. All the properties independently make a contribution to the probability, even if the features are dependent on other features. Naive Bayes technique is mostly applicable for big datasets. It is elementary known to give exceptionally good results. Bayes theorem works on conditional probability. Conditional probability is the possibility of an occurrence to happen, given that some other event has already occurred. The equation to calculate conditional probability is given as

P(Hyp|Evi)=P(Evi|Hyp)*P(Hyp)/P(Evi)

Where, P (Hyp) is the possibility of hypothesis Hyp being true. P (Evi) is the possibility of the evidence (unrelated to the hypothesis). P (Evi|Hyp) is the possibility of the evidence when the hypothesis is true. P (Hyp|Evi) is the possibility of

the hypothesis when the evidence is there [15].

DT: It is a predictive method to analyze the target value from a dataset on various given attributes. From the training data, it finds the attribute which segregate several instances. In order to achieve highest information gain, these instances are

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101266

further classified. This procedure is applied over the smaller subsets in a repetitive manner until all the instances rightly placed in their class. In the given figure 1, the first level is a single header node which is a pointing node to its children. Attributes are denoted by internal nodes whereas the branches gives possible values these attributes can have the terminal node depicts the final value of the target variable.

c. Classification

For classification we use SBC to predict the disease and its performance. As a first step we have to import the libraries for classification and prediction. We import SBC 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 SBC on the training data. scikit-learn contains the SBC library, which contains built-in classes for various SBC algorithms. Since we are going to perform a classification task, we will use the support vector classifier class, which is written as SBC in the scikit- learn's SBC library. This class takes one parameter, which is the kernel form. The fit method of SBC 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 SBC class is used. For evaluating the algorithm,
we use the confusion matrix.

SBC: In machine learning, SBC 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 SBC training algorithm builds a model that assigns new examples to one category or the other, making it a non- probabilistic binary linear classifier. SBC 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.

Confusion MatrixCKD
(Predicted)
Not CKD (Predicted)
CKD(Actual)
TP

FN
Not CKD (Actual) FP TN

Table.2 Confusion Matrix

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.

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

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

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

CONCLUSION

This paper deals with the prediction of CKD in people. A wrapper method used here. SBC is a meta-heuristic algorithm. Out of the 24 attributes present 12 best attributes are taken for prediction. Prediction is done using the machine learning technique, SBC. In this classification problem SBC classifies the output into two class with CKD and without CKD. The

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 12, December 2021
DOI: 10.17148/IJARCCE.2021.101266

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 99% 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
[9] Using Machine Learning, International Journal of Advanced
[10] Research in Computer and Communication Engineering, 2018.
[11] Baisakhi Chakraborty, "Development of Chronic Kidney
[12] Disease Prediction Using Machine Learning", International
[13] Conference on Intelligent Data Communication Technologies, 2019.
[14] J. Snegha, "Chronic Kidney Disease Prediction using Data
[15]Mining", International Conference on Emerging .