International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-03, JUNE 2021

Analyzing risk factor in kidney disease using Supervised Algorithms

[1]K.Jhansi Lakshmi, [2]Jessy, [3]K.Hemanth, [4]J.Madhu Sudhana Rao, [5]Mrs.Y.Aditya

[1,2,3,4]Student, [5]AssistantProfessor, CSE Department & GEC, Gudlavalleru, Krishna, India.

[1]jhansilakshmikollipara99@gmail.com, [2]jessykattupalli@gmail.com,

[3]hemanthkatta1999@gmail.com, [4]madhuj.9977@gmail.com, [5]adityalu@gmail.com

***Abstract*--- Chronic kidney disease (CKD) is a type of kidney disease in which there is a gradual loss of kidney function over a period of months or years. Prediction of this disease is one of the most important problems in medical fields. So, automated tool which will use machine learning techniques to determine the patient's kidneycondition that will be helpful to the doctors in prediction of chronic kidney disease and hence better treatment. The proposed system extracts the features which are responsible for CKD, then machine learning process can automate the classification of the chronic kidney disease in different stages according to its severity. The objective is to use machine learning algorithm and predicting the chronic kidney disease using clinical data.**

***Keywords*--- Chronic kidney disease, Machine learning algorithms, Classification.**

I. INTRODUCTION

The kidneys are a two of a kind of organs placed towards the lower back of the abdomen. Kidney job is to strain the blood by moving out the toxic substance from the body using the bladder through urination. Kidney failure can cause death if the kidneys do not remove waste which is affected toxins. Acute and chronic are two types of difficulties of kidney. Chronic kidney diseases include circumstances that harm kidneys and reduce their capability to keep us fit.CKD can be caused by diabetes, hypertension, coronary heart disease, lupus, anemia, bacteria and albumin in the urine, complications of some drugs, sodium and potassium deficiency in the blood and a family history and many others. Early disclosure and medical care can usually avoid the worsening of chronic kidney disease. If it can get worse, it can result into kidney failure requiring dialysis.

Chronic kidney disease (CKD) is a universal public wellbeing difficulty. 10% of the world's population is affected by CKD. However, there is a little direct evidence on how to diagonize CKD systematically. Worldwide, CKD is a serious reason of demise and disability. It was the 27th focal reason in 1990 and became 18th focal reason in 2010.Despite that, people of developing countries are being affected by CKD. According to the Kidney Foundation, out of 18 million people about 35,000 to 40,000 patients suffer with chronic renal failure in Bangladesh each year.

Director of the National Institute of Renal Diseases and Urology (NIKDU) Nurul Huda Lenin said that women are more comparing to men in case of by CKD. There are three main causes which are responsible for CKD among women: neglect, social barriers, and lack of consciousness. President of Bangladesh kidney foundation said that the patients of diabetes and hypertension are increasing because poor food habit and undisciplined daily life which led them kidney disease

II. RELATED WORK

Gunarathne W.H.S.D et.al.Has compared results of different models. And finally they concluded that they good accuracy for Multiclass Decision forest algorithm which is around 99% for the reduced dataset of 14 attributes.[1]. S.Ramya and Dr.N.Radha worked on diagnosis time and improvement of diagnosis accuracy using different classification algorithms of machine learning. The analysis results indicate that RBF algorithm gives better results than the other classifiers and produces 85.3%[2]. S.DilliArasu and Dr. R. Thirumalaiselvi[3] has worked on missing values in a dataset of chronic kidney disease. They replaced missing values with mean.median values. Asifsalekin and john Stankovicthey use novel approach to detect CKD using machine learning algorithms like KNN and random forest and neural networks to get results[4].

Yildrim searches the effect of class imbalance when we train the data by using development of neural network algorithm for making medical decision on CKD.Sahilsharma, vinod, and anulhas assessed 12 different classification algorithm on dataset having 400 records and 24 attributes. They used assessment metrics like accuracy, sensitivity, precision and specificity.

III. METHODOLGY

This section illustrates the entire concept of the research

 DOI : 10.35291/2454-9150.2021.0332

work which will aid to understand the wholenotionof the paper. At first, we collected the data and preprocess it. After preprocessing the data, we handled the missing data of the dataset using Anaconda. Feature selection is conducted to extract the most significant features. Then we apply different machine learning algorithms on the dataset. Figure 3.1 shows the system structure of proposed method.

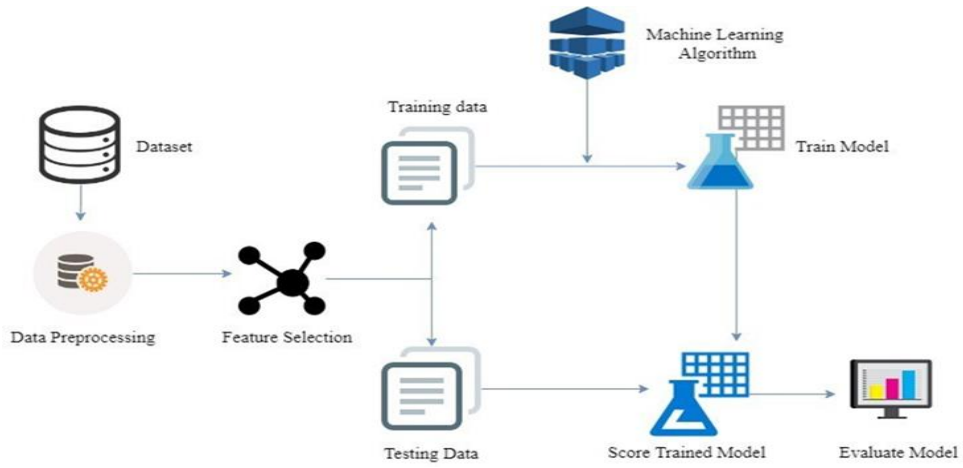

Figure:1 Architecture Model

Data Collection:

The dataset of prediction of chronic kidney disease using machine learning algorithm is downloaded from Kaggle website. In that dataset there are 400 patient records are included. Also, they include 25 attributes but we take only some attributes to build the model.

Data Preprocessing:

It is an essential step to gain better accuracy. Before going out in a journey it is important to be prepared. It is same for the machine learning journey. If there will be no data pre-processing than machine learning model won't work appropriately. "Pandas" and "Numpy" library is used for processing the data.

Handling Missing Data:

While processing the data, there were so many incomplete data means there were numerous missing data which occurs myriad of time in real life and we managed the missing data using mean method.

Extracting the Most significant feature:

There were few features in the dataset which was not relevant to the objective. So feature selection is a vital step. By using chi-square test we extracted the most significant features. There were 24 features excluding class feature. After extracting there are 20 features and then we trained the dataset. CNN model is created based on the requirement. Train the model by assigning the training dataset.

Training the model: The model is created based on the requirement. Train the model by assigning the training dataset.

Applying Algorithms:

We used the classification algorithms. The algorithms are: Logistic Regression, Artificial Neural Networks (ANN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision tree Classifier.

Logistic regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data 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. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Types of Logistic Regression:

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those numbers of categories, Logistic regression can be divided into following types −

Binary or Binomial

In such a kind of classification, a dependent variable will have only two possible types either 1 or 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

Multinomial

In such a kind of classification, dependent variable can have 3 or more possible ***unordered*** types or the types having no quantitative significance. For example, these variables may represent "Type A" or "Type B" or "Type C".

 DOI : 10.35291/2454-9150.2021.0332

Ordinal

In such a kind of classification, dependent variable can have 3 or more possible ***ordered*** types or the types having a quantitative significance. For example, these variables may represent "poor" or "good", "very good", "Excellent" and each category can have the scores like 0,1,2,3.

Artificial Neural Network:

Artificial neural network (ANN) is a supervised learning method. It is made of a huge number of simple elements, called perceptrons. A neural system has just three layers of neurons, an input layer that receives the input, one hidden layer and an output layer that produces output. A multilayer perceptron is a gathering of perceptrons, sorted out in various layers that precisely respond tp composite questions. Each and every perceptron emits signals from input layer, and at last output layer.Back propagation algorithm is a group of techniques which is used to instruct ANN.ANN is following a gradient-based optimization algorithm that follows a chain rule. The primary component of back propagation is itsrecurrent,recursive and effective technique for computing the weight updates to improve the system units.

K-Nearest Neighbor:

K-nearest neighbours (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

The following two properties would define KNN well −

Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

K-nearest neighbours (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −

Step 1: For implementing any algorithm, we need dataset. In the first step of KNN, we must load the training as well as test data.
Step 2: We need to choose the value of K i.e. the nearest data points. K can be any integer.
Step 3: For each point in the test data do the following −
3.1 − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
3.2 − Now, based on the distance value, sort them in ascending order.
3.3 − Next, it will choose the top K rows from the sorted array.
3.4 − Now, it will assign a class to the test point based on most frequent class of these rows.
Step 4: End.

Decision Tree Classifier:

In general, Decision tree analysis is a predictive modeling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions tress are the most powerful algorithms that falls under the category of supervised algorithms.They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where we got outcome.

Implementing Decision Tree Algorithm

Gini Index

It is the name of the cost function that is used to evaluate the binary splits in the dataset and works with the categorical target variable "Success" or "Failure".

Higher the value of Gini index, higher the homogeneity. A perfect Gini index value is 0 and worst is 0.5 (for 2 class problem). Gini index for a split can be calculated with the help of following steps −

First, calculate Gini index for sub-nodes by using the formula p^2+q^2 , which is the sum of the square of probability for success and failure.

Next, calculate Gini index for split using weighted Gini score of each node of that split.

Classification and Regression Tree (CART) algorithm uses Gini method to generate binary splits.

Split Creation

A split is basically including an attribute in the dataset and a value. We can create a split in dataset with the help of following three parts −

Part1: Calculating Gini Score − We have just discussed this part in the previous section.

Part2: Splitting a dataset − It may be defined as separating a dataset into two lists of rows having index of an attribute and a split value of that attribute. After getting the two groups - right and left, from the dataset, we can calculate the value of split by using Gini score calculated in first part. Split value will decide in which group the attribute will reside.

Part3: Evaluating all splits − Next part after finding Gini score and splitting dataset is the evaluation of all splits. For this purpose, first, we must check every value associated with each attribute as a candidate split. Then we need to find the best possible split by evaluating the cost of the split. The best split will be used as a node in the decision tree.

IV. RESULTS &DISCUSSION

Our main goal in this project is to predict the kidney disease risk analysis using various machine learning

 DOI : 10.35291/2454-9150.2021.0332

techniques. We predicted using Support Vector Machine(SVM),Logistic Regression,K-Nearest Neighbor(KNN),Decision tree algorithms.

The proposed model with Decision tree technique provides the better result among all techniques like SVM,KNN and

Logistic Regression.

Analysis of results:

Table1. The precision,recall,F1_score and accuracy values of different algorithms

Algorithms	Precision	Recall	F1_score	Accuracy
Logistic Regressor	0.98	0.89	0.93	96
SVM	0.97	0.95	0.96	61.875
KNN	0.96	0.97	0.96	75.625
Decision Tree Classifier	0.97	0.96	0.95	96.875

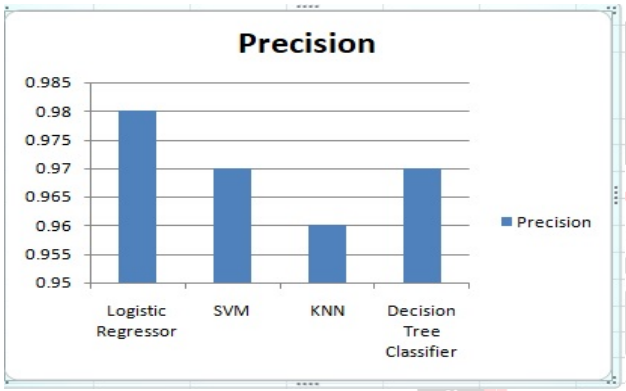

Figure 2: Comparison of Precision values

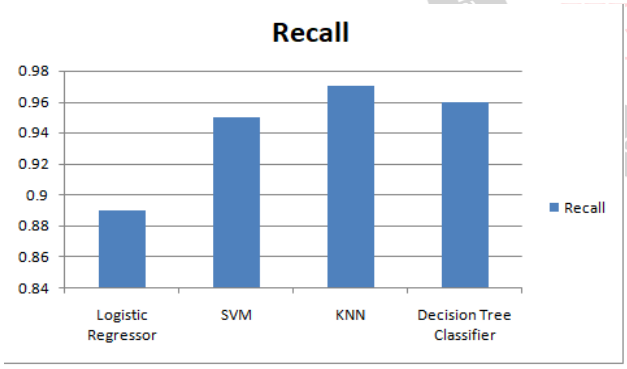

Figure 3: Comparison of Recall

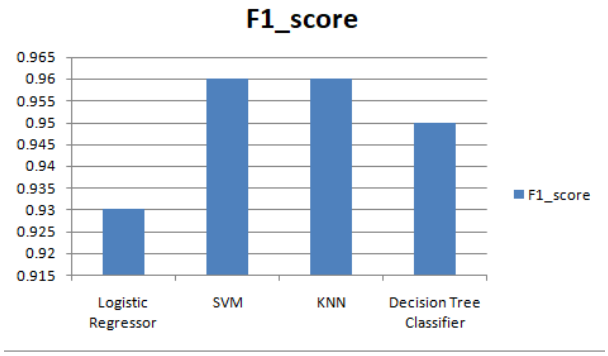

Figure 4: Comparison of F1_Score

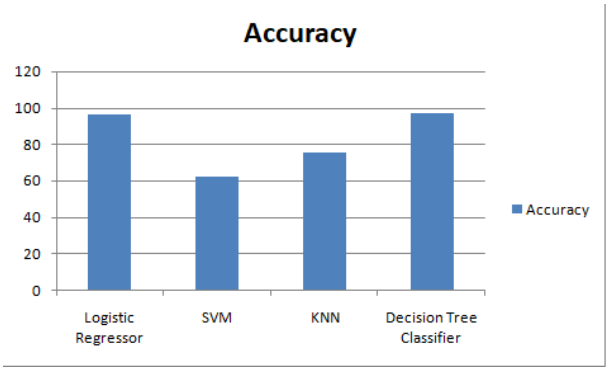

Figure 5: Comparison of Accuracy

So, finally as from the table mentioned we had got the best accuracy for **Decision Tree Classifier** Algorithms that is an accuracy of **96.87%.** This is the execution part of the algorithm.

```
classifier = DecisionTreeClassifier()
classifier.fit(x_train,y_train)
y_pred=classifier.predict(x_test)
accuracy=accuracy_score(y_test,y_pred)
print("Accuracy of the model:", accuracy*100)
```

```
Accuracy of the model: 96.875
```

```
classifier=DecisionTreeClassifier()
classifier.fit(x_train,y_train)
y_pred=classifier.predict([[48.0,80.0,1.0,0.0,0,0,0,0,121.000000,36.0,1.2,4.627244,7800.0,1,1,0,0,0]])
y_pred
```

```
array([1], dtype=int64)
```

```
y_pred=classifier.predict([[17.0,60.0,0.0,0.0,0,0,0,0,114.000000,50.0,1.0,4.900000,7200.0,0,0,0,0,0]])
y_pred
```

```
array([0], dtype=int64)
```

From the above image it is clear that we had first trained the model by using Decision Tree Classifier and we had predicted the output for the newly given inputs. Here 0 represents that the person is not suffering from CKD and 1 represents that the person is suffering from CKD.

V. CONCLUSION

The diagnosis for early stage of kidney disease is necessary in challenges to the medical field. Incorporated data mining algorithms are obtained better accuracy. An important challenge in data mining and machine learning areas is to bulid accurate and computationally efficient classifiers for medical applications.In this paper we have studied different machine learning algorithms. We have analyzed 20 different attributes related to CKD patients and predicted accuracy for different machine learning algorithms like Decision tree classifier, Support vector machine, KNN, ANN, Logistic regression. From the result analysis, it is observed that the decision tree algorithms gives the

 DOI : 10.35291/2454-9150.2021.0332

accuracy of 96.875%, and Logistic regression gives accuracy score of 91.875%, and SVM gives accuracy score of 61%, and KNN gives the accuracy score of 75%. When considering the decision tree algorithm, it builds the tree based on the entire dataset by using all the features of the dataset. The advantage of this system is that the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less period of time.

FUTUREWORK

For future work we propose a way in which we predict the different stages of the Chronic kidney disease in the patient. In clearly, our future work determines that at what stage the patient was suffering with chronic kidney disease like early stage or final stage. Knowing the stages of disease can help the doctors and patients to take treatment according to the stage the disease occur.

REFERENCES

[1] GunarathneW.H.S.D,Perera K.D.M, Kahandawaarachchi K.A.D.C.P, "Performance Evaluation on Machine Learning Classification Techniques for Disease classification and forecasting through Data Analytics for Chronic kidney disease(CKD)",2017 IEEE 17th International Conference on Bioinformatics and Bioengineering.

[2] S.Ramya, "Diagnosis of Chronic Kidney Disease using Machine Learning Algorithms," Proc.International Journal of Innovative Research in Computer and CommunicationEngineering,Vol 4, Issue 1, January 2016.

[3] S. Gopika, and Dr. M.Vanitha, "Machine learning approach of Chronic Kidney Disease Prediction using Clustering

Technique", International journal of Innovative Research in Science, Engineeringand Technology, Vol. 6, no. 7, pp.14488-14496,2017.

[4] AsifSalekin, JhonStankovic, "Detection of Chronic Kidnet Disease and Selecting Important Predictive Attributes", in 2016 IEEE International Conference on Healthcare Informatics (ICHI), Chicago, USA, 2016, pp. 262-270.

[5] Tabassum, MamathaBai B G, JharnaMajumdar, "Analysis and prediction of Chronic Kidney Disease using Data Mining Techniques", International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), vol. 4, no. 9, pp.25-31, 2017.

[6] Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of various Machine Learning Classification Techniques of Chronic Kidney Disease Diagnosis," July18,2016.

[7] SiddeshwarTekale, "Prediction of Chronic Kidney Disease Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering, 2018.

[8] Kai-Cheng Hu, "Multiple Pheromone table based on Ant Colony Optimization for Clustering", Hindawi, Research article, 2015.

[9] VivekanandJha, "Chronic Kidney Disease Global Dimension and Perspectives", Lancet, National Library of Medicine, 2013.

[10] Guneet Kaur, Er.Ajay Sharma , "Predict chronic kidney disease using data mining algorithms in hadoop ", in 2017 Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), 2017, pp. 973-979.

[11] Faisal Aqlan, Ryan Markle, Abdulrahman Shamsan, "Data mining for chronic kidney disease prediction", 67th Annual Conference and Expoof the Institute of Industrial Engineers, United States, 2017, pp.1 1789-1794.

 DOI : 10.35291/2454-9150.2021.0332