A New Emphasis to Predict Chronic Kidney Disease using Machine Learning Algorithms

Ann Mariya George , Shejina N M , Dr G.Kiruthiga

[1] *Student , Department of Computer Science, IES College of Engineering Thrissur, Kerala, India*
[2] *Assistant Professor, Department of Computer Science, IES College of Engineering Thrissur, Kerala, India*

[3]*Associate Professor, Department of Computer Science, IES College of Engineering Thrissur, Kerala, India*

ABSTRACT

According to India's health statistics on Chronic Kidney Disease (CKD) 63,538 cases have been registered. CKD is more conventional among males than females. Agonizingly, India ranks among top 17 countries in CKD since 2015, which is characterized by a gradual loss of excretory organ performance .Machine Learning is used across many sectors around the world mainly in healthcare industry. In the human body, the kidney is instrumental in absorbing and discharging all the toxic and unessential materials, typically wastes, from the body through egesting and excretion process. As per the study ,in India, there are approximately one million cases of Chronic Kidney Disease (CKD) every year. It is dangerous to kidney and it produces gradual loss in kidney functionality. Nevertheless, it is unpredictable because its symptoms grow gradually and are not unique to the disorder, it is important to detect CKD at its early stage. In the early stages of CKD, a few signs or symptoms will be observed. Machine Learning, is making sensible applications in the areas such as analyzing medical science outcomes, sleuthing fraud etc. For the prediction of CKD different machine learning algorithms are used.This making sensible applications in the areas such as analyzing medical science outcomes, sleuthing fraud etc. For the prediction of chronic diseases various machine learning algorithms are implemented.Based on its accuracy differentiating the performance of various machine learning algorithms .In this research work has idolized Rcode to compare their performance. The pivotal purpose of this study is to analyze the Chronic Kidney Disease dataset and conduct CKD and Non CKD classification cases.

Keyword : - *Machine Learning, Chronic Kidney Disease, Classification, Accuracy, Logistic Regression, Support Vector Machine*

1. INTRODUCTION

The technology progressed since, mankind were obsessed with the technology. In these days industries like automation, aerospace, health care, etc., are operating in communication of Machines or interconnection of Internet of Things (IoTs).By clustering, the objects are divided into different clusters that are similar in nature .The objects in other groups are dissimilar .Nowadays, it is applicable in applications like Marketing, World Wide Web (WWW), Earthquake Studies, Aerospace, Biology, Insurance, etc. The numerous applications of classification are speech and handwriting recognition, Identification of biometric, classification of documents, etc. Association Rule Mining (ARM) is: if-then statements facilitate to indicate the relationships between data items among transactional databases.Regression (or linear regression) is employed to seek out the relationship between two continuous

variables. One variable is termed as predictor or independent variable and other is dependent or response variable.These all algorithms mentioned above are a part of Data mining/ Machine Learning/ Computer Vision.
The kidney, in the human body, is instrumental in absorbing and discharging all the toxic and unessential materials, typically wastes, from the body through egesting and excretion process. In India, there are approximately one million cases of Chronic Kidney Disease (CKD) every year. It is dangerous to kidney and it produces gradual loss in kidney functionality. Nevertheless, it is unpredictable because its symptoms grow gradually and are not unique to the disorder, it is important to detect CKD at its early stage. Kidneys filter wastes and excess fluids from the blood that are then excreted in excrement.A few signs or symptoms will be observed in the early stages of CKD. In healthcare organization, Classification is one in all the topmost usually used ways of machine learning and it shows the class of result for each data point.The classifying models are Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes classifier, Neural Networks, and Random Forest. KNN is used to visualize at the relationship between different CKD risk factors, in order to predict the disease at an early stage**.** Machine Learning techniques have proved great success in detection and recognition of many essential diseases in medical science's point of view. Machine learning would thus be helpful for predicting whether the patient has CKD or not in this question. By using old CKD patient data to train predictive model, Machine Learning does so.

1.1 Analysis of Chronic Kidney Disease (CKD)

Health-related quality of life (QOL) is an important measure of how disease affects patients' lives. Dialysis patients have decreased QOL relative to healthy controls.it is assessed that just about 3 million people within the United Kingdom are at risk of CKD. A combination of totally different conditions that usually place as train on the kidneys works on CKD.
Hence, the manuscript is organized as follows: Section II mentions about related work upon this research . Section III, discusses the Proposed System to classify Chronic Kidney Disease (CKD). Section IV describes the information regarding the dataset used and transient introduction about the attributes. Section V contract with the machine learning algorithms, code and its results for variable measures and therefore the corresponding output obtained in each classification algorithm. Further, section VI discusses about an open discussion about current view, results about chronic disease. Section VII finally discusses the conclusion of the research work alongside with the attribute improvement.

2. RELATED WORK

There are many researchers who have worked with the assistance of several different classification algorithms on CKD prediction. All those had their model performances expected. Gunarathne, W.H.S.D. [1] compared the effects of divergent models. Finally, they concluded that the Multiclass Decision forest algorithm provides plentiful precision for the 14-attribute (reduced) data set. S.Dilli Arasu and Dr. R. Thirumalaiselvi [2] worked on missing values in a Chronic Kidney Disease dataset. They deduced that the missing values in the dataset can not only reduce the model's accuracy but also the effects of the prediction. By patterning a numerical method on stages of Chronic Kidney Disease, they found a solution to this issue and by doing so; they stood up with unknown values. They substituted the missing values with those recalculated ones.
In discovering Chronic Kidney Disease using machine learning algorithms, Asif Salekin and John stankovic[3] used novel approach. They got findings on a dataset consisting of 400 records and 25 attributes resulting in a patient prone to CKD or not. In order to achieve results, they used KNN, random forest and Neural Network algorithms. They used wrapper methodology for feature reduction which finds CKD with high accuracy.12 different classification algorithms on various datasets were tested by Sahil Sharma, Vinod Sharma, and Atul Sharma [4], each with 400 records and 24 attributes. They compared their expected outcomes with actual results in order to determine predictive accuracy. They used metrics such as precision, sensitivity, accuracy and specificity for measuring the performance of the classifiers. Note that Chronic Kidney Disease (CKD) is not uncommon. Now, next section will discuss the datasets that are being employed and introduce the table that shows the attributes and description on the same.

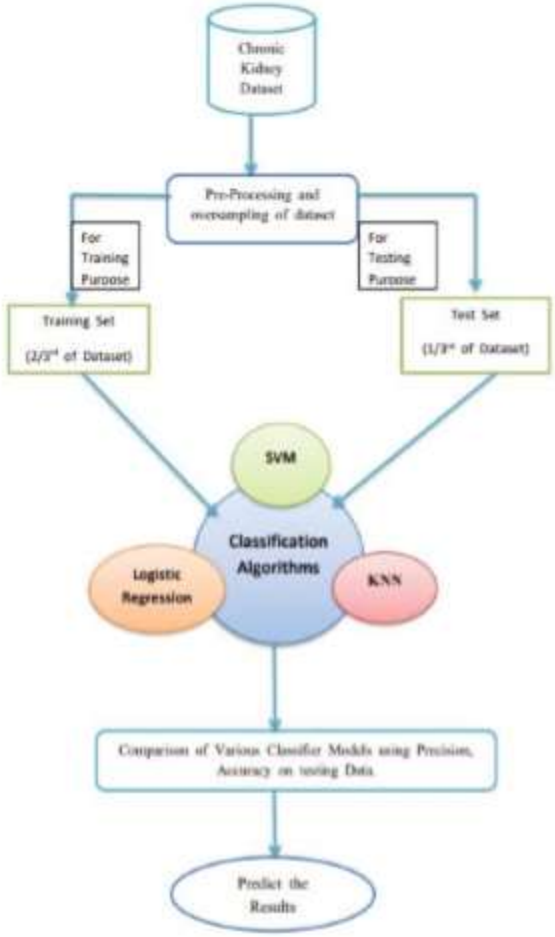
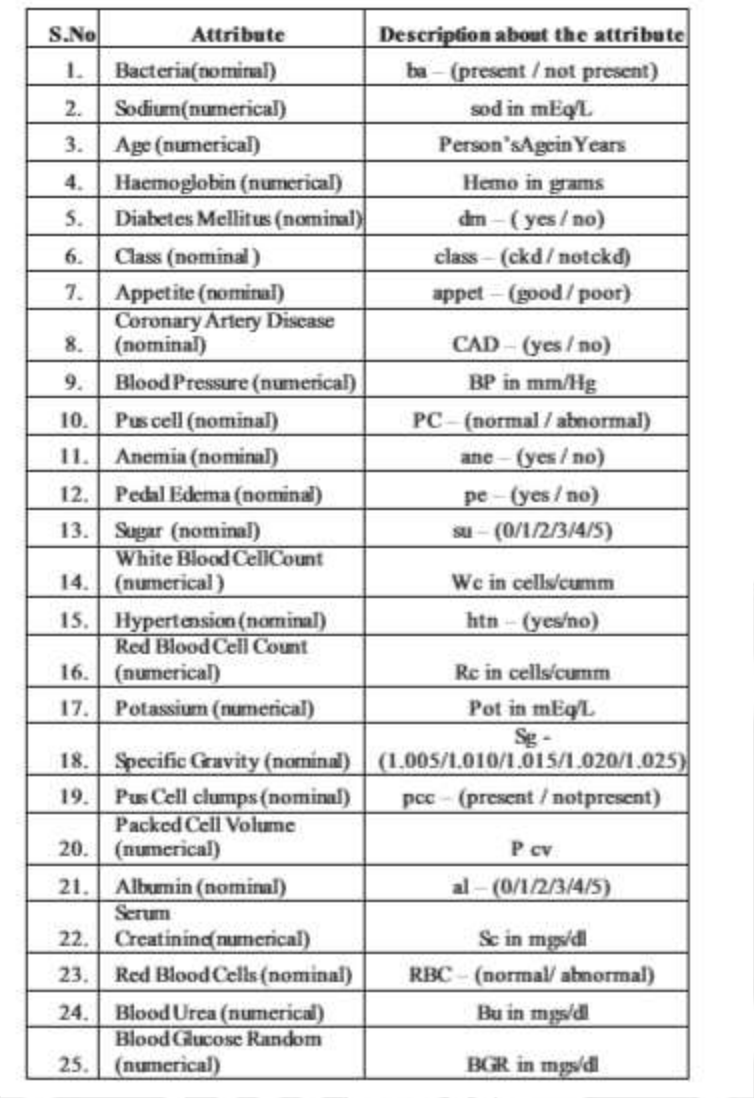

Chronic Kidney Dataset
Pre-Processing and oversampling of dataset
For Training Purpose
For Testing Purpose
Training Set (2/3rd of Dataset)
Test Set (1/3rd of Dataset)
SVM
Classification Algorithms
Logistic Regression
KNN
Comparison of Various Classifier Models using Precision, Accuracy on testing Data
Predict the Results

Fig-1:Proposed Model using Various Classification Algorithms

3. DATASET USED

The proposed system uses the UCI Machine Learning Repository dataset called Chronic Kidney Disease (CKD) that has 25 attributes, out of which, 11 are numerical and 14 are nominal. Entire 400 instances of the dataset are used for training to predict machine learning algorithms. In 400 instances, 250 are labeled as Chronic Kidney Disease (CKD) and 150 are labeled as Non Chronic Kidney disease. The attributes in the data set are bacteria, sodium, age, Hemoglobin, Diabetes Mellitus, Classification, Appetite, Coronary Artery Disease, Blood Pressure, Pus cell, Anemia, Pedal Edema, Sugar, White Blood Cell Count, Hypertension, Red Blood Cell Count, Potassium, Specific Gravity, Pus cell chumps, Packed Cell Volume, Albumin, Serum Creatinine, Red Blood Cells, Blood Urea, and Blood Glucose Random. The dataset that is taken is divided into two groups, one for testing the samples and another for training the samples. The ratio for testing and training data is 30% and 70% respectively. The data set used has been listed in table 1. The readers can refer following URL [16] for collecting data. Now, next section will discuss regarding the machine learning algorithms used to classify CKD. The data was taken over a 2-month period in India with 25 features. The target is the 'classification', which is either 'ckd' or 'notckd' - ckd=chronic kidney disease. Chronic kidney illness has ended up a worldwide wellbeing issue concern with rising predominance. Chorionic kidney disease, also called chronic kidney miscarriage, describes the continual decrease of kidney function Among them Random subspace achieved better accuracy. DT and NB classification techniques were applied to predict CKD for prevention of death rate caused by CKD [13]. Authors implemented these data mining methods using rapid miner tool.



S.No	Attribute	Description about the attribute
1.	Bacteria(nominal)	ba – (present / not present)
2.	Sodium(numerical)	sod in mEq/L
3.	Age (numerical)	Person'sAgeinYears
4.	Haemoglobin (numerical)	Hemo in grams
5.	Diabetes Mellitus (nominal)	dm – (yes / no)
6.	Class (nominal)	class – (ckd / notckd)
7.	Appetite (nominal)	appet – (good / poor)
8.	Coronary Artery Disease (nominal)	CAD – (yes / no)
9.	Blood Pressure (numerical)	BP in mm/Hg
10.	Pus cell (nominal)	PC – (normal / abnormal)
11.	Anemia (nominal)	ane – (yes / no)
12.	Pedal Edema (nominal)	pe – (yes / no)
13.	Sugar (nominal)	su – (0/1/2/3/4/5)
14.	White Blood CellCount (numerical)	Wc in cells/cumm
15.	Hypertension (nominal)	htn – (yes/no)
16.	Red Blood Cell Count (numerical)	Rc in cells/cumm
17.	Potassium (numerical)	Pot in mEq/L
18.	Specific Gravity (nominal)	Sg - (1.005/1.010/1.015/1.020/1.025)
19.	Pus Cell clumps (nominal)	pcc – (present / notpresent)
20.	Packed Cell Volume (numerical)	P cv
21.	Albumin (nominal)	al – (0/1/2/3/4/5)
22.	Serum Creatinine(numerical)	Sc in mgs/dl
23.	Red Blood Cells (nominal)	RBC – (normal/ abnormal)
24.	Blood Urea (numerical)	Bu in mgs/dl
25.	Blood Glucose Random (numerical)	BGR in mgs/dl

TABLE I:DATASET USED

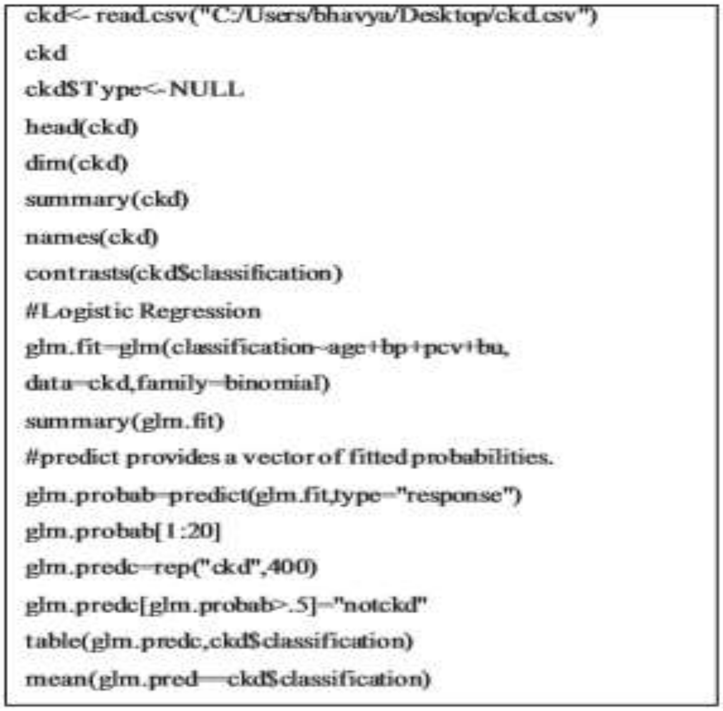
4 SIMULATION RESULTS

This section describes about the simulation results that are being used in the paper.

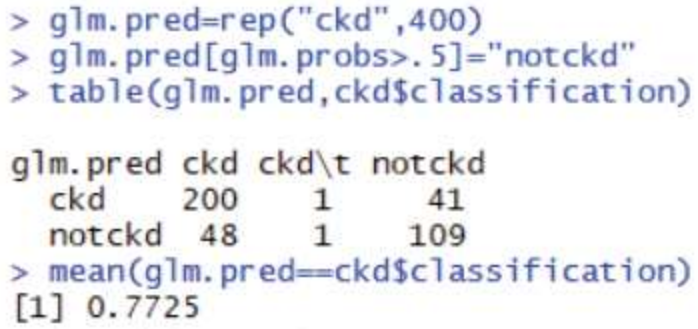
4.1 Logistic Regression

Logistic Regression is a calculation for order. As a result , the logic is 1/0, Yes/No, True/False. It can be employed to access a paired answer. There is a tendency to utilize the likelihood log as an impoverished variable. Logistic Regression is used for the classification problems in Machine Learning Algorithms. It is a prophetic analysis algorithm and it is based on the concept of probability. It means that, given a certain factor, logistic regression is used to predict an outcome that has two values. The source code is exemplified in Table I and the output in Fig.2. From them, it is deduced that the accuracy of Logistic Regression is 0.7725.

TABLE II : RCODE FOR LOGISTIC REGRESSION



```
ckd<- read.csv("C:/Users/bhavya/Desktop/ckd.csv")
ckd
ckd$Type<-NULL
head(ckd)
dim(ckd)
summary(ckd)
names(ckd)
contrasts(ckd$classification)
#Logistic Regression
glm.fit=glm(classification~age+bp+pcv+bu,
data=ckd,family=binomial)
summary(glm.fit)
#predict provides a vector of fitted probabilities.
glm.probab=predict(glm.fit,type="response")
glm.probab[1:20]
glm.predc=rep("ckd",400)
glm.predc[glm.probab>.5]="notckd"
table(glm.predc,ckd$classification)
mean(glm.pred==ckd$classification)
```



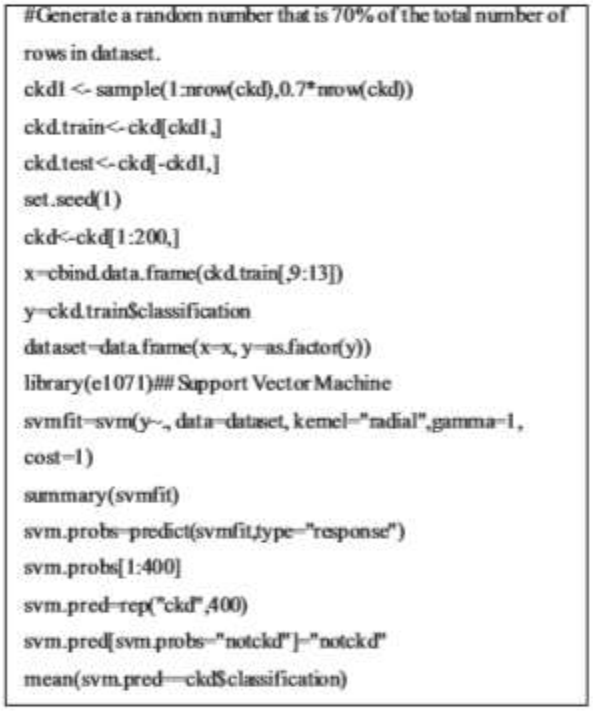
```
> glm.pred=rep("ckd",400)
> glm.pred[glm.probs>.5]="notckd"
> table(glm.pred,ckd$classification)

glm.pred ckd ckd\t notckd
  ckd    200    1     41
  notckd  48    1    109
> mean(glm.pred==ckd$classification)
[1] 0.7725
```

Fig -2: Output for Logistic Regression

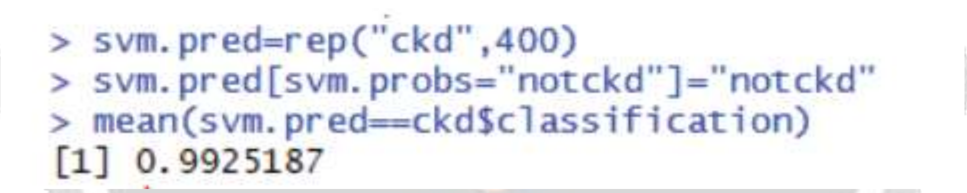
4.2 Support Vector Machines (SVM)

In ML, SVM support vector systems are supervised models compatible with learning. Support Vector Machine (SVM) offers platform for regression and classification. This can be used to solve both linear problems and non-linear ones. This algorithm uses a hyper plane to categorize the data points. Within this SVM algorithm, each data point will be plotted as a point in n dimensional space, with a value of each attribute being the value of a given coordinate. Classification can be accomplished by searching for the right hyper-plane which basically distinguishes between the two CKD and not CKD groups. Table III presents the code behind SVM and from the results in Fig.3, it can be witnessed that the accuracy of SVM = 0.9925187.

TABLE III : RCODE FOR SVM



```
#Generate a random number that is 70% of the total number of
rows in dataset.
ckd1 <- sample(1:nrow(ckd),0.7*nrow(ckd))
ckd.train<-ckd[ckd1,]
ckd.test<-ckd[-ckd1,]
set.seed(1)
ckd<-ckd[1:200,]
x=cbind.data.frame(ckd.train[,9:13])
y=ckd.train$classification
dataset=data.frame(x=x, y=as.factor(y))
library(e1071)## Support Vector Machine
svmfit=svm(y~., data=dataset, kernel="radial",gamma=1,
cost=1)
summary(svmfit)
svm.probs=predict(svmfit,type="response")
svm.probs[1:400]
svm.pred=rep("ckd",400)
svm.pred[svm.probs=="notckd"]="notckd"
mean(svm.pred==ckd$classification)
```



```
> svm.pred=rep("ckd",400)
> svm.pred[svm.probs="notckd"]="notckd"
> mean(svm.pred==ckd$classification)
[1] 0.9925187
```

Fig- 3. Output for SVM

4.3 K-Nearest Neighbors Classification

The performance of the K nearest neighbor classifier algorithm is to predict the target variable by capturing the nearest neighbor class. The nearest class will be known as the target variable using the distance measures like Euclidean distance.

Algorithm:

1. Initialize the parameter K.
2. Calculate the distance between the test sample and all the training samples
3. Sort the distance in the ascending order.
4. Take K-nearest neighbors.
5. Gather the class of the nearest neighbor.
6. Here as observed, the accuracy in KNN = 0.7875

From the algorithm mentioned above, it is evident that the results are better in Support Vector Machine. The result is provided with an accuracy of 0.9925187. Now, the subsequent section will provide a conclusion regarding this work in brief adding some future enhancement possibilities with this work.

4.4 Random Forest Classification

Random Forest or Random Decision Forests is used for both regression as well as classification.For classification tasks, the output of the random forest is the class selected by the most trees. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. A 98.0 per cent F1-measure was obtained with better efficiency relative to study using RF and five apps.

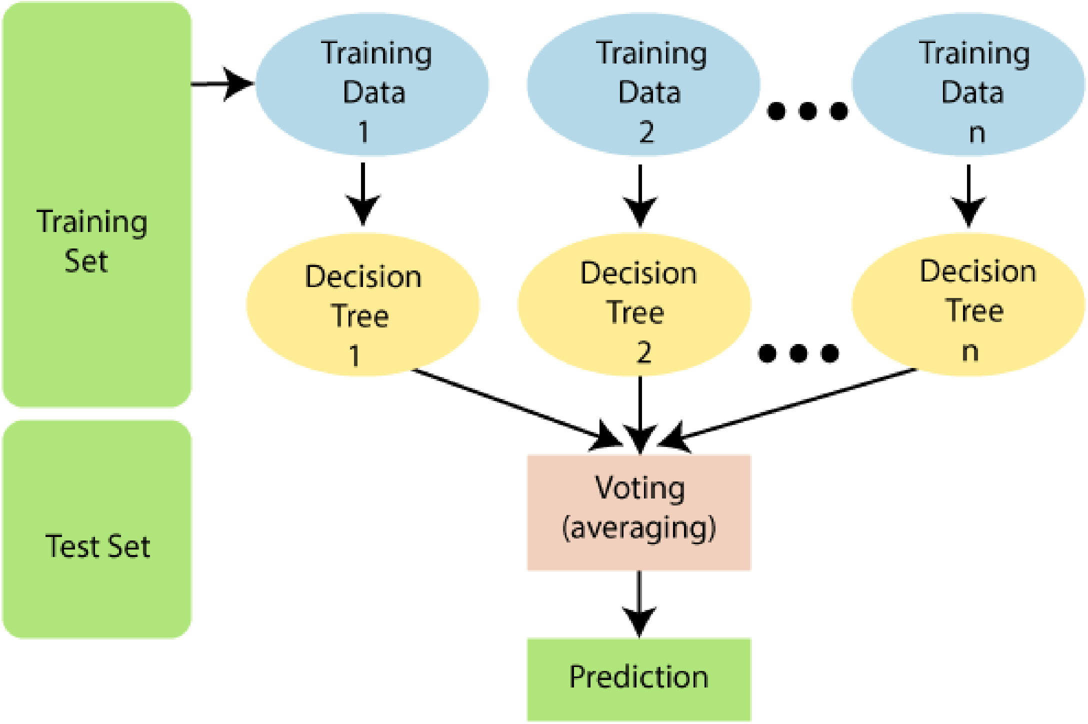


Fig- 4. Random Forest

4. CONCLUSIONS

The main aim is to diagnose Chronic Kidney Disease (CKD) at an earlier stage, this manuscript introduced a variety of machine learning algorithms. This publication offered a range of machine learning methods with the main goal of earlier diagnosis of Chronic Kidney Disease (CKD). The aforementioned input characteristics are used to authenticate and train the systems that were collected from CKD patients. To examine CKD, Support Vector Machine, Logistic Regression, knn, and Random Forest are analysed. Precision largely affected how well the algorithms performed. Within the constrained parameters of this medical scenario, our results demonstrated that the Random Forest Machine method predicts Chronic Kidney Disease more accurately than Logistic Regression and K-Nearest Neighbors. The advantage of this strategy is that the prediction procedure requires significantly less time, allowing doctors to treat CKD patients as soon as possible and classifying a broader population.

5. REFERENCES

[1].] L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,”2015. [Online]. Available:http://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.
[2]. Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selectiing Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.
[3] Q. Zhang and D. Rothenbacher, "Prevalence of chronic kidney disease in population-based studies: systematic review," BMC Public Health, vol. 8, (1), pp. 117, 2008.

[4] K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," Indian Journal of Science and Technology, vol. 9, (29), 2016.

[5] J. Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, (1), pp. 119, 2019.

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

[7] 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 17thInternational Conference on Bioinformatics and Bioengineering.

[8] S.Ramya, Dr.N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

[9] S. A. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: a review," IJET, vol. 7, no. 3, pp. 1019– 1023, 2018.

[10] A.J. Aljaaf et al, "Early prediction of chronic renal disorder mistreatment machine learning supported by prognosticative analytics," in 2018 IEEE Congress on organic process Computation (CEC), 2018.

[11] J.Xiao et al, "Comparison and development of machine learning tools in the prediction of chronic renal disorder progression," Journal of Translational drugs, vol. 17, (1), pp. 119, 2019.

[12] P. Yang et al, "A review of ensemble strategies in bioinformatics, "Current Bioinformatics, vol. 5, (4), pp. 296-308, 2010.

[13] L.Deng et al, "Prediction of protein-protein interaction sites mistreatment associateReference