



# Predictive Analysis of Chronic Kidney Disease (CKD) based on Machine Learning Classification Algorithm

Dillip Narayan Sahu<sup>1</sup>, Vijay Pal Singh<sup>2\*</sup>

<sup>1</sup>Lecturer, Department of MCA, School of Computer Science, Gangadhar Meher University (GMU), Odisha, India

<sup>2</sup>Associate Professor, Department of Computer Science, OPJS University, Rajasthan, India

\*Corresponding Author: Dr. Vijay Pal Singh

**Abstract:** As per as the World Health Organization report is concern, about 10% of the world population is affected by chronic kidney disease (CKD), and millions die only because of inappropriate and non-affordable treatment. Kidney disease is a worldwide health crisis in the present scenario. This disease can be curable with early diagnosis and proper treatment. The purpose of this paper is to establish some predictive models using Machine Learning algorithms by taking a real time CKD dataset. In this paper, we have shown some real-time experiments and observations with the help of some Machine Learning algorithms, and also shown a clear picture on the predictive analysis on medical diagnosis of the chronic kidney disease (CKD) using Machine Learning algorithms using which patients may get accurate data so as to diagnose better for their early treatment.

**Keywords:** Algorithm, Chronic Kidney Disease, Classifier, Machine Learning, Prediction.

## I.INTRODUCTION

Computer program is set to learn from experience E with respect to some class of task T and performance measure P, if its performance for task T as measured by P improves with experience E[1][2]. In this paper we are predicting chronic kidney disease (CKD) based on different machine learning algorithm by taking a dataset which includes some real-time cases/data of kidney patients[3].

## II.EXPERIMENTS AND OBSERVATIONS1

We have taken Weka as a Machine Learning tool to predict and classify the disease based on the real-time kidney patient dataset[4].

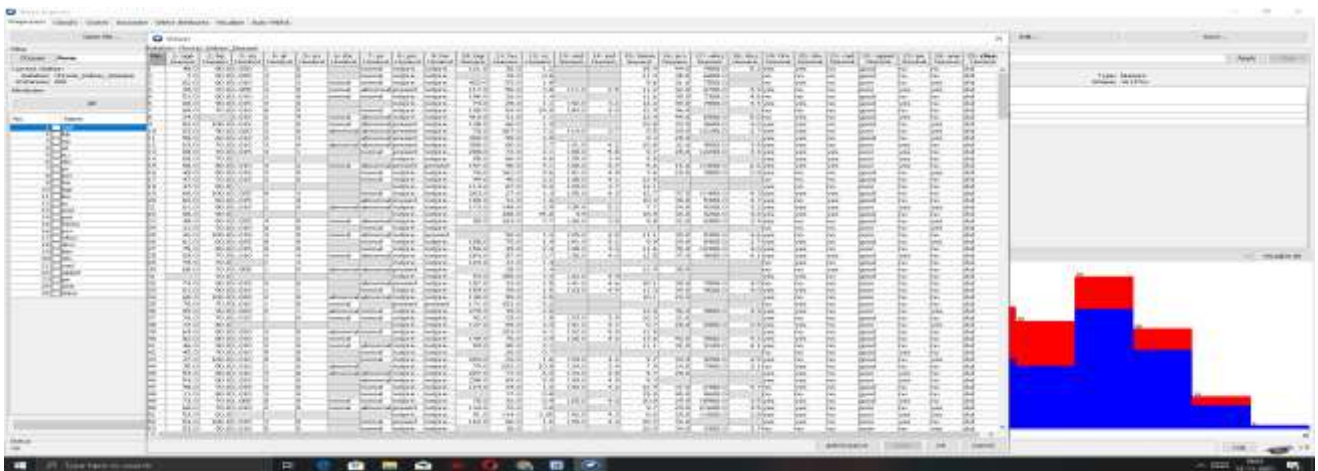


Fig.1 CKD Dataset

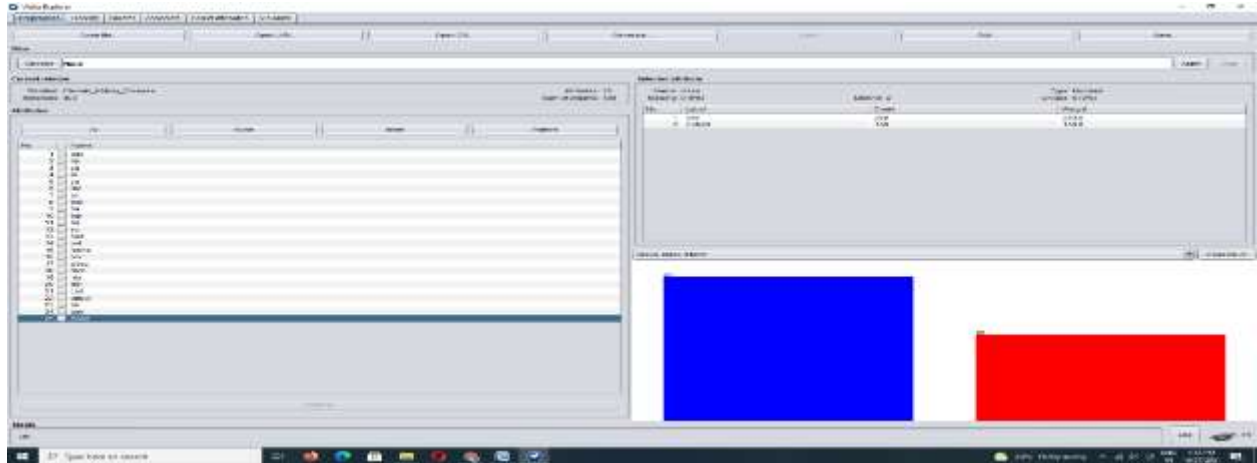


Fig.2 Preprocess of the CKD Dataset

NAME=weka.classifiers.rules.ZeroR Classifier

SYNOPSIS=Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class)[5].

CAPABILITIES includes:

Class -- Binary class, Date class, Missing class values, Nominal class, Numeric class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Relational attributes, String attributes, Unary attributes

**Classifier Output**==== Run information ====

Scheme: weka.classifiers.rules.ZeroR

Relation: Chronic\_Kidney\_Disease

Instances: 400

Attributes: 25

age	bp	sg	al	su	rbc	pc	pcc	ba
bgr	bu	sc	sod	pot	hemo	pcv	wbcc	
rbcc	htn	dm	cad	appet	pe	ane	class	

Test mode: 10-fold cross-validation

==== Classifier model (full training set) ====

ZeroR predicts class value: ckd

Time taken to build model: 0 seconds

==== Stratified cross-validation ===== Summary ====

Correctly Classified Instances 250 62.5 %

Incorrectly Classified Instances 150 37.5 %

Kappa statistic 0

Mean absolute error 0.4689

Root mean squared error 0.4841

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 400

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
---------	---------	-----------	--------	-----------	-----	----------	----------	-------

1.000	1.000	0.625	1.000	0.769	?	0.500	0.625	ckd
-------	-------	-------	-------	-------	---	-------	-------	-----

0.000	0.000	?	0.000	?	?	0.500	0.375	notckd
-------	-------	---	-------	---	---	-------	-------	--------

Weighted Avg.	0.625	0.625	?	0.625	?	?	0.500	0.531
---------------	-------	-------	---	-------	---	---	-------	-------

==== Confusion Matrix ====

a b <-- classified as

250 0 | a = ckd

150 0 | b = notckd

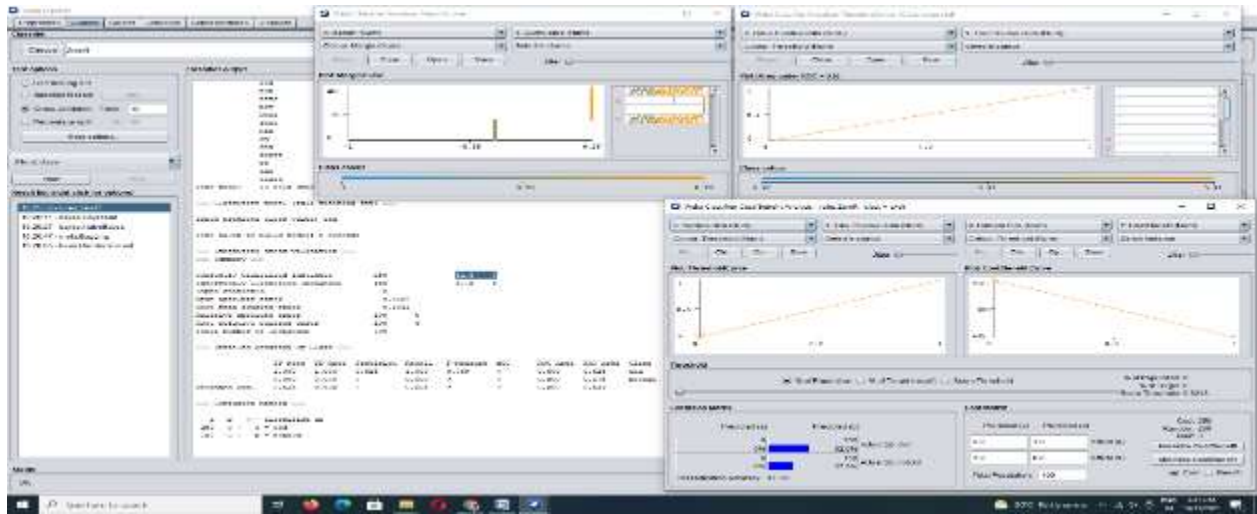


Fig.3 Classifier Output with Visualize Margin Curve, Threshold Curve, Cost/Benefit Analysis

III.EXPERIMENTS AND OBSERVATIONS-2

NAME=weka.classifiers.bayes.BayesNet

SYNOPSIS=Bayes Network learning using various search algorithms and quality measures[6].

Base class for a Bayes Network classifier. Provides datastructures (network structure, conditional probability distributions, etc.) and facilities common to Bayes Network learning algorithms like K2 and B.

CAPABILITIES

Class -- Binary class, Missing class values, Nominal class

Attributes -- Binary attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Classifier Output==== Run information ===

Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Relation: Chronic\_Kidney\_Disease

Instances: 400

Attributes: 25

age	bp	sg	al	su	rbc	pc	pcc	ba
bgr	bu	sc	sod	pot	hemo	pcv	wbcc	
rbcc	htn	dm	cad	appet	pe	ane	class	

Test mode: 10-fold cross-validation

==== Classifier model (full training set) ===

Bayes Network Classifier not using ADTree

#attributes=25 #classindex=24

Network structure (nodes followed by parents)

age(3): class	bp(2): class	sg(5): class	al(6): class	su(6): class	rbc(2): class
pc(2): class	pcc(2): class	ba(2): class	bgr(2): class	bu(2): class	sc(2): class
sod(3): class	pot(4): class	hemo(2): class	pcv(2): class	wbcc(2): class	rbcc(2): class
htn(2): class	dm(2): class	cad(2): class	appet(2): class	pe(2): class	ane(2): class

class(2): LogScore Bayes: -4816.950237024905 LogScore BDeu: -4930.853594314016

LogScore MDL: -4998.805083491287 LogScore ENTROPY: -4762.14223388052

LogScore AIC: -4841.1422338805205

Time taken to build model: 0.01 seconds

==== Stratified cross-validation ===== Summary ===

Correctly Classified Instances	395	98.75 %
Incorrectly Classified Instances	5	1.25 %
Kappa statistic	0.9735	
Mean absolute error	0.013	
Root mean squared error	0.1035	
Relative absolute error	2.7761 %	



Root relative squared error 21.3829 %  
 Total Number of Instances 400

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.980	0.000	1.000	0.980	0.990	0.974	1.000	1.000	ckd
1.000	0.020	0.968	1.000	0.984	0.974	1.000	0.999	notckd

Weighted Avg. 0.988 0.008 0.988 0.988 0.988 0.974 1.000 1.000

==== Confusion Matrix ====

```
a b <-- classified as
245 5 | a = ckd
0 150 | b = notckd
```

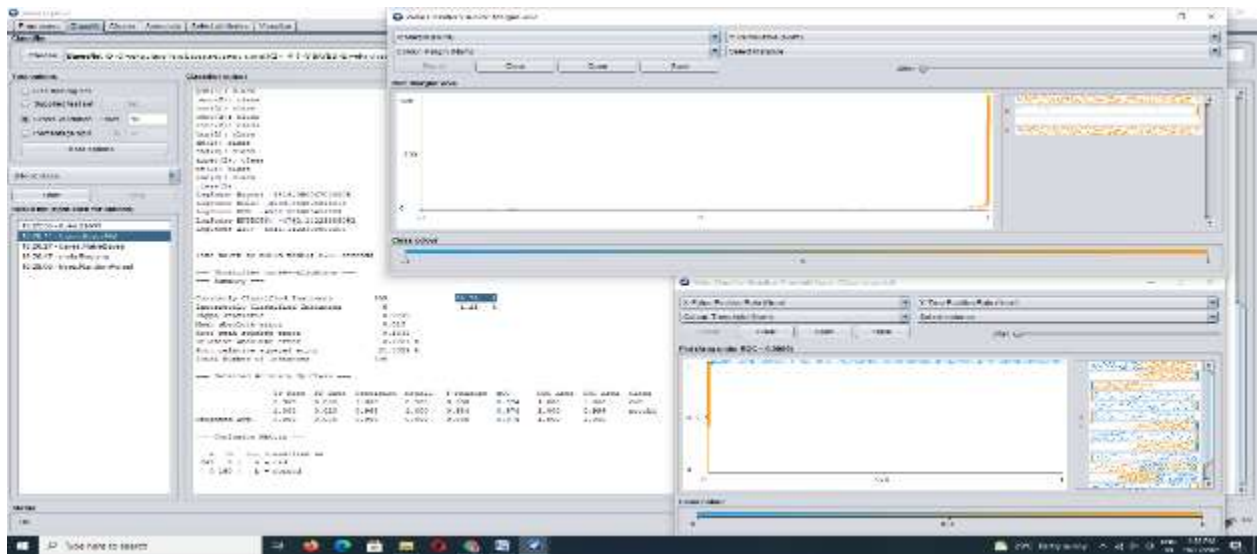


Fig.4 Classifier Output with Visualize Margin Curve, Threshold Curve

IV.EXPERIMENTS AND OBSERVATIONS-3

NAME=weka.classifiers.bayes.NaiveBayes

SYNOPSIS

Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data[7][8]. For this reason, the classifier is not an UpdateableClassifier (which in typical usage are initialized with zero training instances) -- if you need the UpdateableClassifier functionality, use the NaiveBayesUpdateable classifier[9]. The NaiveBayesUpdateable classifier will use a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances.

CAPABILITIES

Class -- Binary class, Missing class values, Nominal class

Attributes -- Binary attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Classifier Output==== Run information ====

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: Chronic\_Kidney\_Disease

Instances: 400

Attributes: 25

Age	bp	sg	al	su	rbc	pc
Pcc	ba	bgr	bu	sc	sod	pot
Hemo	pcv	wbcc	rbcc	htn	dm	cad
Appet	pe	ane	class			

Test mode: 10-fold cross-validation

==== Classifier model (full training set) ====

Naive Bayes Classifier: Class

Attribute ckd notckd



(0.62) (0.38)

==== Summary ====

```

Correctly Classified Instances    380    95 %
Incorrectly Classified Instances  20     5 %
Kappa statistic                   0.8961
Mean absolute error               0.0479
Root mean squared error           0.2046
Relative absolute error           10.2125 %
Root relative squared error       42.2526 %
Total Number of Instances        400
    
```

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.920	0.000	1.000	0.920	0.958	0.901	1.000	1.000	ckd
	1.000	0.080	0.882	1.000	0.938	0.901	1.000	1.000	notckd
Weighted Avg.	0.950	0.030	0.956	0.950	0.951	0.901	1.000	1.000	

==== Confusion Matrix ====

```

a  b  <-- classified as
230 20 | a = ckd
 0 150 | b = notckd
    
```

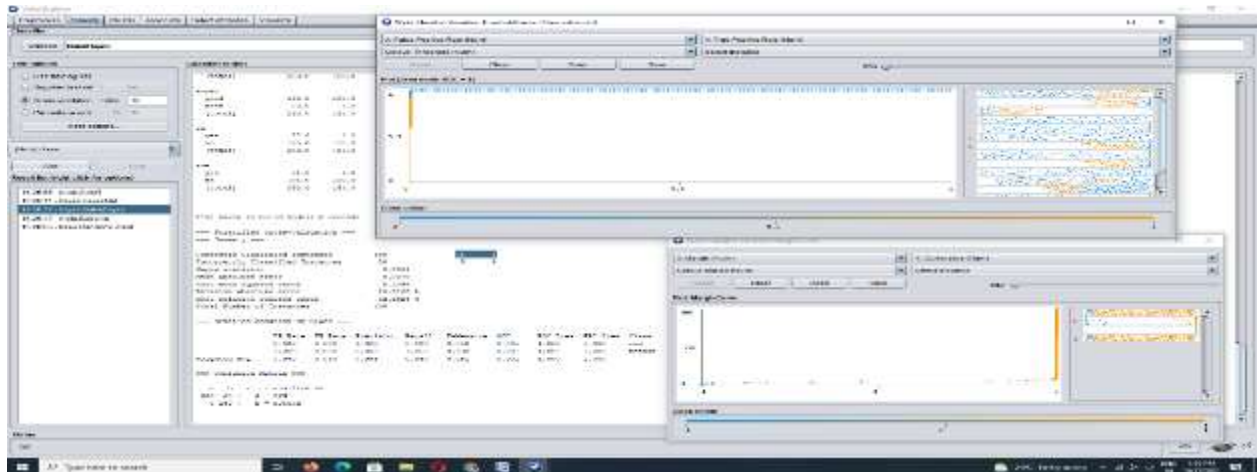


Fig.5 NaiveBayes Classifier Output with Visualize Curve

V.EXPERIMENTS AND OBSERVATIONS-4

NAME=weka.classifiers.meta.Bagging

SYNOPSIS=Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner[10].

CAPABILITIES

Class -- Binary class, Date class, Missing class values, Nominal class, Numeric class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Classifier Output----- Run information ----

Scheme: weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

Relation: Chronic\_Kidney\_Disease

Instances: 400 Attributes: 25

== Classifier model (full training set) ==

Bagging with 10 iterations and base learner

weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0

Time taken to build model: 0.05 seconds

==== Stratified cross-validation ===== Summary ====

```

Correctly Classified Instances    395    98.75 %
    
```



Incorrectly Classified Instances 5 1.25 %  
 Kappa statistic 0.9734  
 Mean absolute error 0.0545  
 Root mean squared error 0.1204  
 Relative absolute error 11.6266 %  
 Root relative squared error 24.8732 %  
 Total Number of Instances 400

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.984	0.007	0.996	0.984	0.990	0.974	0.999	0.999	ckd
0.993	0.016	0.974	0.993	0.983	0.974	0.999	0.998	notckd

Weighted Avg. 0.988 0.010 0.988 0.988 0.988 0.974 0.999 0.999

=== Confusion Matrix ===

a b <-- classified as  
 246 4 | a = ckd  
 1 149 | b = notckd

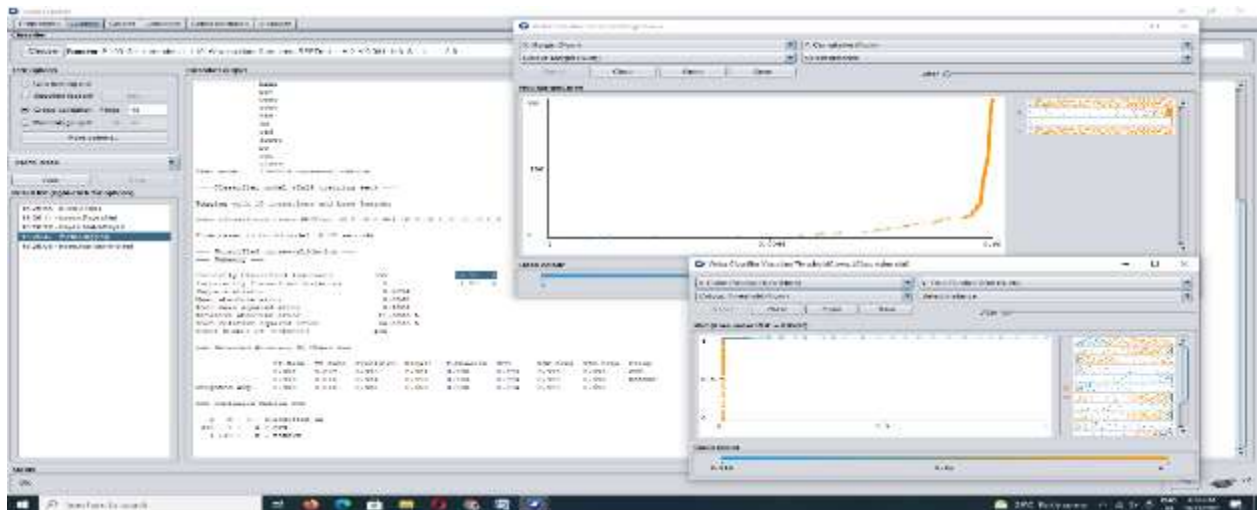


Fig.6 Bagging Classifier Output with Visualize Curve

VLEXPERIMENTS AND OBSERVATIONS-5

NAME=weka.classifiers.trees.RandomForest

SYNOPSIS= Class for constructing a forest of random trees.

CAPABILITIES

Class -- Binary class, Missing class values, Nominal class, Numeric class

Attributes -- Binary attributes, Date attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes

Classifier Output==== Run information ===

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: Chronic\_Kidney\_Disease

Instances: 400 Attributes: 25

Test mode: 10-fold cross-validation

=== Classifier model (full training set) === RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===== Summary ===

Correctly Classified Instances	400	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0414	



Root mean squared error 0.0844  
 Relative absolute error 8.8189 %  
 Root relative squared error 17.4439 %  
 Total Number of Instances 400

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	ckd
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	notckd

Weighted Avg. 1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000

=== Confusion Matrix ===

a b <-- classified as  
 250 0 | a = ckd  
 0 150 | b = notckd

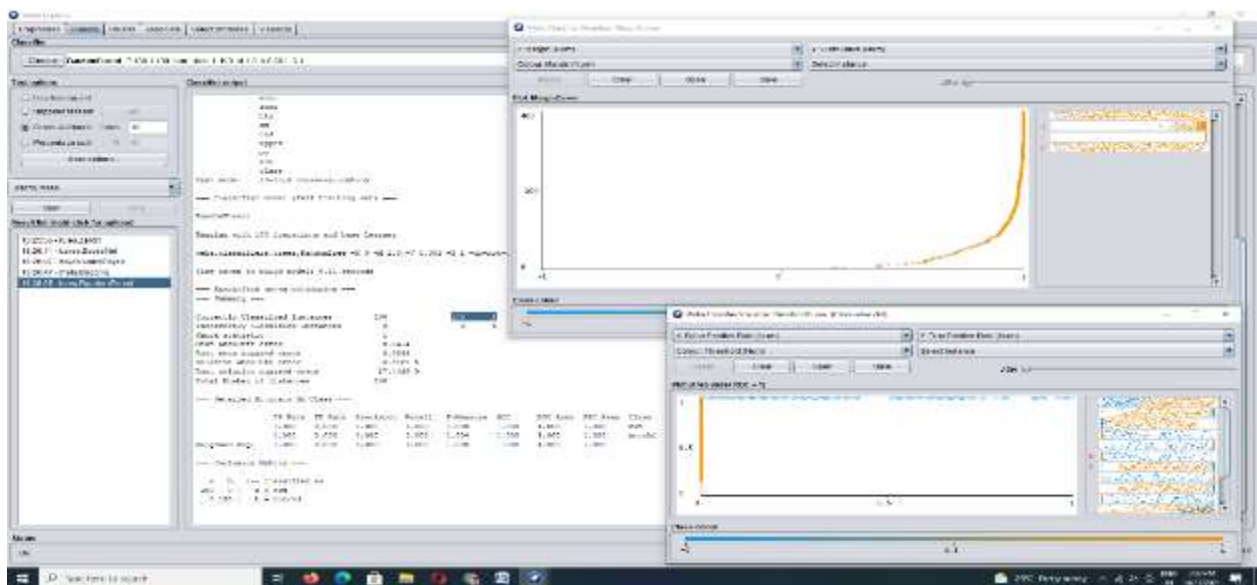


Fig.7 RandomForest Classifier Output with Visualize Curve

VII.CONCLUSION

We have taken 5 different machine learning classifier algorithms and with the observations to decide the acceptability of a particular domain in the machine learning model. In the study of the above real time medical dataset implementation and in different observations, it is found that the accuracy level using the machine learning classification model Bagging is very much satisfactory, having good accuracy rate of 98.75% and so will be a good option in the field of medical sciences to predict early diagnosis of chronic heart diseases.

REFERENCES

[1] Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B. & Yang, C. W. (2013). Chronic kidney disease: global dimension and perspectives. *The Lancet*, 382(9888), 260-272.  
 [2] Ali, S., Dave, N., Virani, S. S., & Navaneethan, S. D. (2019). Primary and secondary prevention of cardiovascular disease in patients with chronic kidney disease. *Current Atherosclerosis Reports*, 21(9), 1-9.  
 [3] Levey, A. S., Coresh, J., Bolton, K., Culleton, B., Harvey, K. S., Ikizler, T. A. & Briggs, J. (2002). K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases*, 39(2 SUPPL. 1), i-ii+.  
 [4] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.  
 [5] Gunarathne, W. H. S. D., Perera, K. D. M., & Kahandawaarachchi, K. A. D. C. P. (2017, October). Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD). In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 291-296). IEEE.  
 [6] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).  
 [7] Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology*, 24(12), 1565-1567.  
 [8] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9(4), 705-724.  
 [9] Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology*, 7(1), 242.  
 [10] Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. *International Journal of Computing and Business Research (IJCBR)*, 6(2), 1-12.  
 [11] Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. *International Journal of Engineering Research and Technology*, 4(7), 206-210.