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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

Comparative Analysis for the Prediction of major Diseases using Supervised and Hybrid Machine Learning Algorithms

Dillip Narayan Sahu[1], Vijay Pal Singh[2*]

[1]Research Scholar, Department of Computer Science, OPJS University, Rajasthan, India

[2]Associate Professor, Department of Computer Science, OPJS University, Rajasthan, India

Abstract: In the few recent years, common major diseases have emerged together of the foremost common causes of deaths worldwide. As the changing in lifestyle, food habits, working cultures etc, has significantly contributed to the present issues related to health across world-wide including the developed, underdeveloped and developing countries, a challenge to the medical science to overcome from this situation[1]. As per as WHO (World Health Organization) Global Health Estimates report is concerned, an estimated 74% of all deaths were noncommunicable diseases globally, 3 out of 10 major diseases are communicable. In this paper, we have taken 5 major diseases (Ischaemic Heart Disease (IHD) with Stroke, Chronic Kidney Disease (CKD), Diabetes Mellitus (DM) including BP, Chronic Liver, and Cancer) among the top 10 deadliest diseases[2]. All these major diseases can be curable with proper diagnosis and early detection. The purpose of this paper is to establish some Machine Learning supervised algorithms with some hybrid approach for better comparative analysis and predict for the particular disease at an early stage with a greater accuracy level. The outcome of this paper also justify that the hybrid algorithm model has better processing, performance with more accuracy outputs so as to help the medical and healthcare sector in the early stage disease prediction.

Keywords: Algorithm, Classifier, Machine Learning, Predictive Analysis.

INTRODUCTION

The use of machine learning with deep learning in medical and health care sector is increasing day by day. The Machine Learning is a powerful field of Artificial Intelligence where we can do a better analysis, processing, visualizing, cleaning, classification, clustering, etc. with the help of different machine learning algorithms which are exist as well as some hybrid methods to predict and to obtain better results [3][4]. In this paper, we have taken 5 different dataset based on 5 different major diseases as per the report of World Health Organization (WHO). We investigated a total of 5 different major diseases based on the WHO report on the major diseases with different experiments and observations by the help of Weka data analysis for machine learning tool.

People of many developing countries are still living in poorness with dysfunctional health care systems and very restricted access to basic treatment [5]. The information needed for the analysis of health state of affairs and health issues comprise of mortality, morbidity, demographic condition, socioeconomic factors etc [6].

Following are the number of the main diseases in today's world that cause health issues and deaths in several countries in the world [7]. The five major diseases are- Cardio Vascular Disease (Heart), Chronic Kidney Disease, Chronic Liver Disease, Cancer Disease, Diabetes Mellitus.

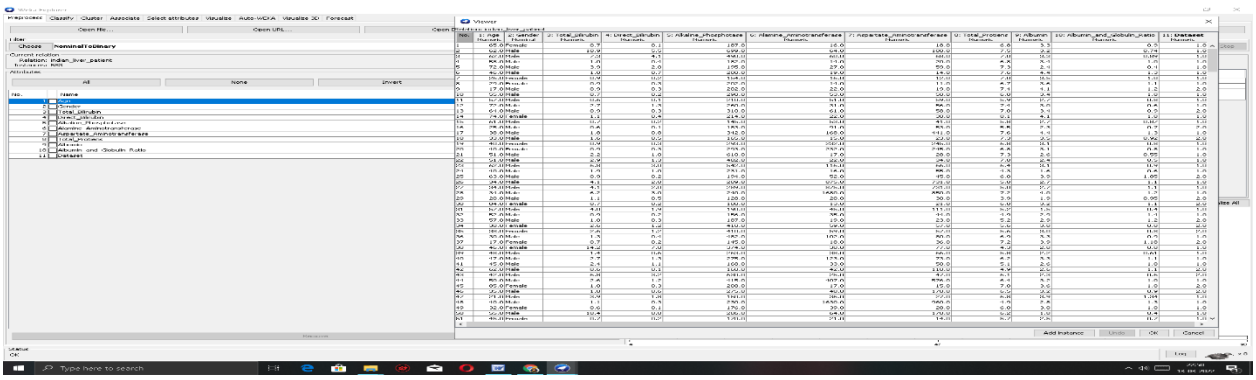

Fig.1 Chronic Liver Dataset

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

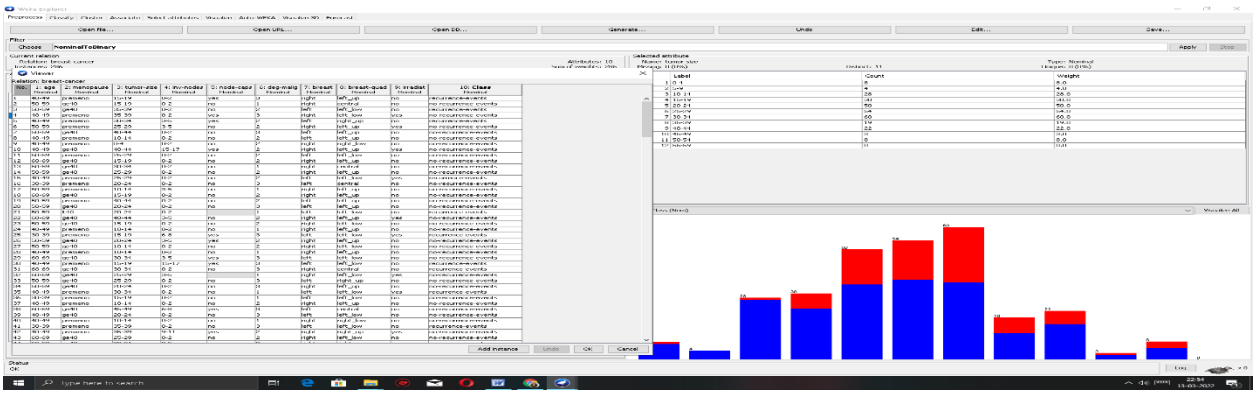

Fig.2 Cancer Dataset

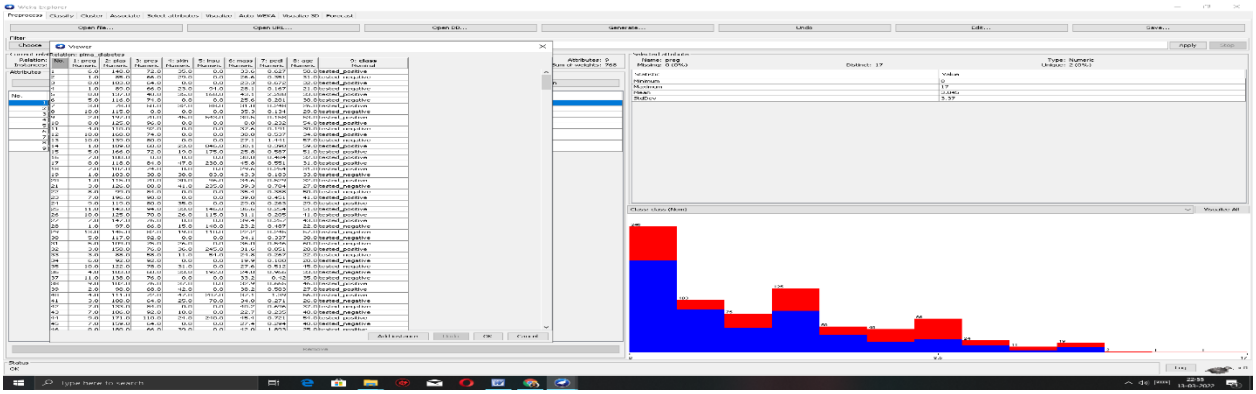

Fig.3 Diabetes Mellitus Dataset

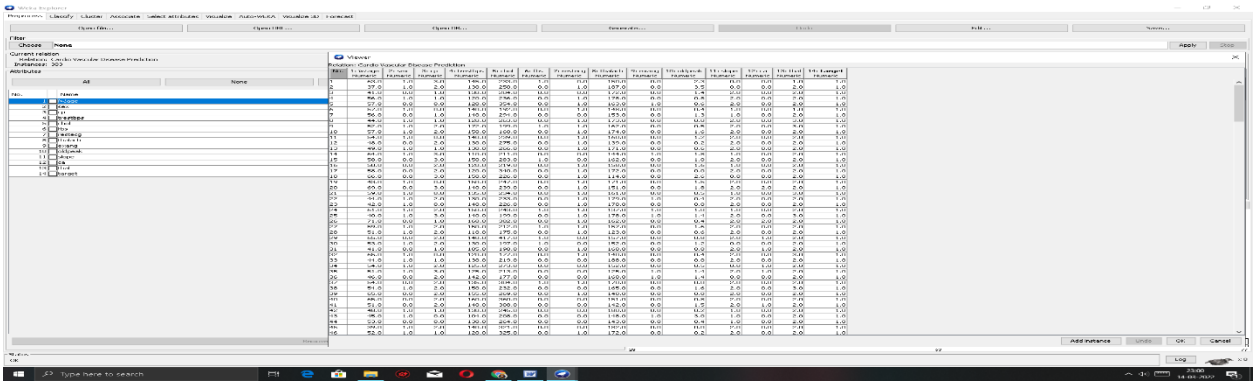

Fig.4 Cardio Vascular Disease Dataset

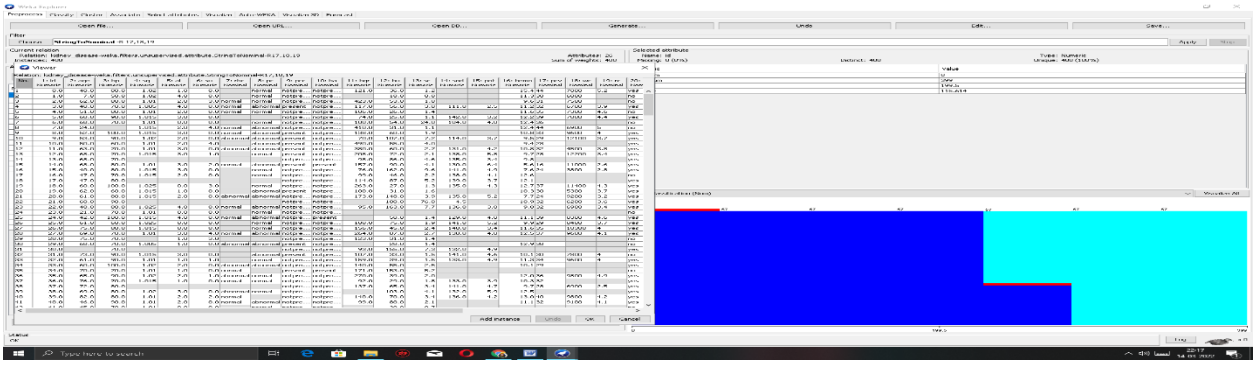

Fig.5 Chronic Kidney Diesase Dataset

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

Experiments and Observations-1

Classifier Algorithm=BayesNet on CKD dataset
10 fold cross validation
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: kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19
Instances: 400 Attributes: 26
Id age bp sg al su rbc pc pcc ba bgr bu sc
Sod pot hemo pcv wc rc htn dm cad appet pe ane
classification
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Bayes Network Classifier
#attributes=26 #classindex=25
=== Stratified cross-validation ====== Summary ===

Correctly Classified Instances	391	97.75 %
Incorrectly Classified Instances	9	2.25 %
Kappa statistic	0.9534	
Mean absolute error	0.0167	
Root mean squared error	0.1039	
Relative absolute error	5.2518 %	
Root relative squared error	26.1069 %	
Total Number of Instances	400	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.972	0.013	0.992	0.972	0.982	0.953	0.990	0.987	ckd
	0.000	0.018	0.000	0.000	0.000	-0.009	0.536	0.008	ckd
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	notckd
Weighted Avg.	0.978	0.008	0.990	0.978	0.984	0.966	0.991	0.987	

=== Confusion Matrix ===

a	b	c	<-- classified as
241	7	0	a = ckd
2	0	0	b = ckd
0	0	150	c = notckd

Experiments and Observtions-2

Classifier Algorithm=NaiveBayes on CKD dataset
10 fold cross validation
Classifier Output- === Run information ===
Scheme: weka.classifiers.bayes.NaiveBayes
Relation: kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19
Instances: 400
Attributes: 26
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier

	Class		
Attribute	ckd	ckd	notckd
	(0.62)	(0.01)	(0.37)

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

Correctly Classified Instances	392	98 %
Incorrectly Classified Instances	8	2 %
Kappa statistic	0.958	
Mean absolute error	0.0153	
Root mean squared error	0.1131	
Relative absolute error	4.8193 %	
Root relative squared error	28.4297 %	

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

Total Number of Instances 400
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.976	0.013	0.992	0.976	0.984	0.958	0.992	0.990	ckd
0.000	0.003	0.000	0.000	0.000	-0.004	0.099	0.005	ckd
1.000	0.020	0.968	1.000	0.984	0.974	1.000	1.000	notckd
Weighted Avg. 0.980	0.016	0.978	0.980	0.979	0.959	0.990	0.989	

=== Confusion Matrix ===

a	b	c	<-- classified as
242	1	5	a = ckd
2	0	0	b = ckd
0	0	150	c = notckd

Experiments and Observations-3

Classifier Algorithm=meta.Bagging on CKD dataset
10 fold cross validation
=== 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: kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19
Instances: 400 Attributes: 26
Test mode: 10-fold cross-validation
=== 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
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances 397 99.25 %
Incorrectly Classified Instances 3 0.75 %
Kappa statistic 0.9841
Mean absolute error 0.009
Root mean squared error 0.0663
Relative absolute error 2.8347 %
Root relative squared error 16.6627 %
Total Number of Instances 400
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.996	0.013	0.992	0.996	0.994	0.984	0.988	0.980	ckd
0.000	0.000	?	0.000	?	?	0.440	0.006	ckd
1.000	0.004	0.993	1.000	0.997	0.995	1.000	1.000	notckd
Weighted Avg. 0.993	0.010	?	0.993	?	?	0.990	0.983	

=== Confusion Matrix ===

a	b	c	<-- classified as
247	0	1	a = ckd
2	0	0	b = ckd
0	0	150	c = notckd

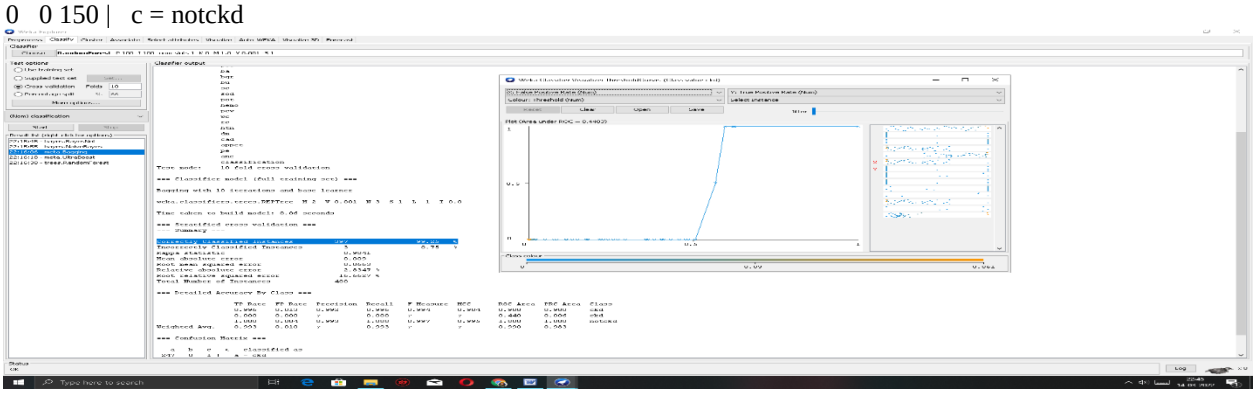

Fig.6 Bagging Classifier with accuracy of 99.25%

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

Experiments and Observations-4

Classifier Algorithm=meta.Ultraboost on CKD dataset
10 fold cross validation
=== Run information ===
Scheme: weka.classifiers.meta.UltraBoost -S 1 -B "weka.classifiers.meta.FilteredClassifier -F \"weka.filters.unsupervised.attribute.RemoveType -V -T nominal\" -S 1 -W weka.classifiers.bayes.NaiveBayes" -B "weka.classifiers.meta.FilteredClassifier -F \"weka.filters.unsupervised.attribute.RemoveType -V -T numeric\" -S 1 -W weka.classifiers.functions.Logistic -- -R 1.0E-8 -M -1 -num-decimal-places 4"
Relation: kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19
Instances: 400
Attributes: 26
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
UltraBoost Base classifiers
FilteredClassifier using weka.classifiers.bayes.NaiveBayes on data filtered through weka.filters.unsupervised.attribute.RemoveType -V -T nominal
Filtered Header
@relation 'kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19-weka.filters.unsupervised.attribute.RemoveType-V-Tnominal'
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances 397 99.25 %
Incorrectly Classified Instances 3 0.75 %
Kappa statistic 0.9842
Mean absolute error 0.044
Root mean squared error 0.1245
Relative absolute error 13.8347 %
Root relative squared error 31.2791 %
Total Number of Instances 400
=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.996	0.013	0.992	0.996	0.994	0.984	0.988	0.981	ckd
0.000	0.003	0.000	0.000	0.000	-0.004	0.606	0.009	ckd
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	notckd
Weighted Avg. 0.993	0.008	0.990	0.993	0.991	0.985	0.991	0.983	

=== Confusion Matrix ===

a	b	c	<-- classified as
247	1	0	a = ckd
2	0	0	b = ckd
0	0	150	c = notckd

Experiments and Observations-5

Classifier Algorithm=RandomForest on CKD dataset
10 fold cross validation
=== 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: kidney_disease-weka.filters.unsupervised.attribute.StringToNominal-R17,18,19
Instances: 400 Attributes: 26
=== Stratified cross-validation ====== Summary ===
Correctly Classified Instances 398 99.5 %
Incorrectly Classified Instances 2 0.5 %
Kappa statistic 0.9894
Mean absolute error 0.04
Root mean squared error 0.0813
Relative absolute error 12.5748 %
Root relative squared error 20.4263 %
Total Number of Instances 400
=== Detailed Accuracy By Class ===

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

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

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11237

	1.000	0.013	0.992	1.000	0.996	0.989	0.989	0.983	ckd
	0.000	0.000	?	0.000	?	?	0.298	0.005	ckd
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	notckd
Weighted Avg.	0.995	0.008	?	0.995	?	?	0.989	0.985	

=== Confusion Matrix ===

a	b	c	← classified as
248	0	0	a = ckd
2	0	0	b = ckd
0	0	150	c = notckd

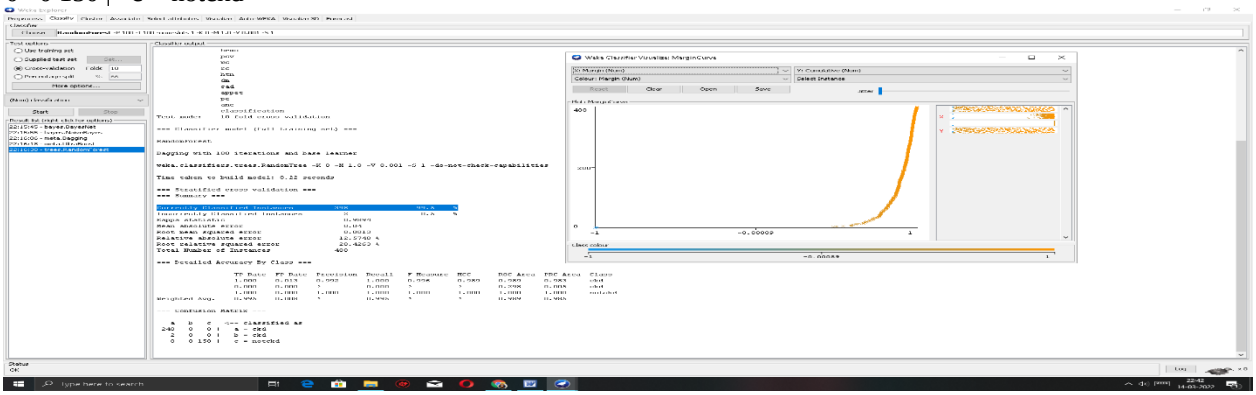

Fig.7 Random Forest Classifier with accuracy of 99.5%

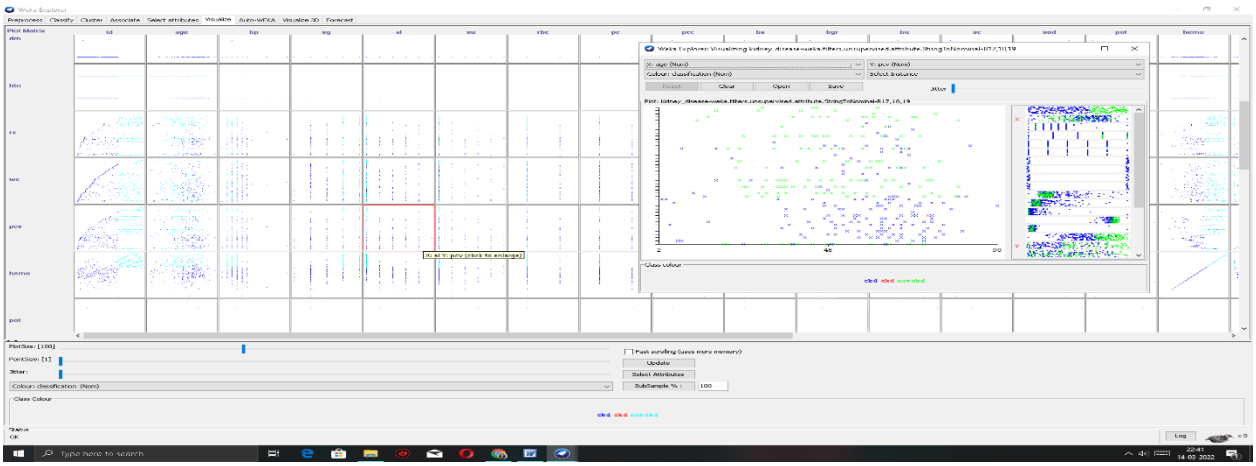

Fig.8 Visualize Curve for Classification

CONCLUSION

Though it is not an easy task to clearly classify errors and make a predictive analysis from the real time disease dataset from the medical and healthcare domain, but it is very important to apply the Artificial Intelligence - Machine Learning model to clearly analyze, visualize, preprocess, predict, classify based on different symptoms of the patients for a better early case diagnosis with a good accuracy level which in turn a necessity for the medical and health care sector and for the health of most people in this world who suffered from these kind of major diseases. As we can see, we have taken 5 major disease dataset as input and process for a better result by applying different existing as well as some hybrid methods for an early case diagnosis of the disease. Hence Machine Learning hybrid classification algorithms will be an option in the medical and healthcare field which will be helpful for early case diagnosis of the major diseases based on different cases.

REFERENCES

[1] Y Khourdifi, M Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization", Int. J. Intell. Eng. Syst, volume 12, issue 1, 2019

[2] S Mohan, C Thirumalai, G Srivastava, "Effective heart disease prediction using hybrid machine learning techniques",IEEE Access, volume 7, 2019

DOI: 10.17148/IJARCCE.2022.11237

[3] M Chen, Y Hao, K Hwang, L Wang, L Wang, "Disease prediction by machine learning over big data from healthcare communities", Ieee Access, volume 5, 2017

[4] S Chae, S Kwon, D Lee, "Predicting infectious disease using deep learning and big data", International journal of environmental research and public health, volume 15, issue 8, 2018

[5] A U Haq, J P Li, M H Memon, S Nazir, R Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms", Mobile Information Systems,2018

[6] M Maniruzzaman, M J Rahman, B Ahammed, M M Abedin, "Classification and prediction of diabetes disease using machine learning paradigm", Health Information Science and Systems, volume 8, issue 1, 2020

[7] Uddin, S., Khan, A., Hossain, M. et al., "Comparing different supervised machine learning algorithms for disease prediction", BMC Med Inform Decis Mak 19, 281, 2019

[8] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach", 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, 2019

[9] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", IEEE Access, vol. 5, no. 1, pp. 8869-8879, 2017.

[10] Ajinkya Kunjir, Harshal Sawant and Nuzhat F. Shaikh, "Data Mining and Visualization for prediction of Multiple Diseases in Healthcare", IEEE big data analytics and computational intelligence, pp. 2325, Oct 2017.

[11] B. Nithya and V. Ilango, "Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 2017.

[12] S. Leoni Sharmila, C. Dharuman and P. Venkatesan, "Disease Classification Using Machine Learning Algorithms - A Comparative Study", International Journal of Pure and Applied Mathematics, vol. 114, no. 6, pp. 1-10, 2017.

[13] Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mohan Ba and H Sarojadevi, "Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology (IJIET), vol. 10, no. 2, May 2018.

[14] Shraddha Subhash Shirsath, "Disease Prediction Using Machine Learning Over Big Data", International Journal of Innovative Research in Science, vol. 7, no. 6, June 2018.

[15] Deepthi, Y., Kalyan, K.P., Vyas, M., Radhika, K., Babu, D.K., Krishna Rao, N.V., "Disease Prediction Based on Symptoms Using Machine Learning.", Energy Systems, Drives and Automations. Lecture Notes in Electrical Engineering, vol 664. Springer, Singapore, 2018

[16] Mir A, Dhage SN, "Diabetes disease prediction using machine learning on big data of healthcare", Fourth international conference on computing communication control and automation (ICCUBEA), 2018

[17] Ray S, "A quick review of machine learning algorithms", International conference on machine learning, big data, cloud and parallel computing (Com-IT-Con), India, 14th–16th Feb 2019

[18] Shetty D, Rit K, Shaikh S, Patil N, "Diabetes disease prediction using data mining", International conference on innovations in information, embedded and communication systems (ICIIECS), 2017

[19] Thirunavukkarasu K, Singh AS, Irfan M, Chowdhury A, "Prediction of liver disease using classification algorithms", 4th international conference on computing communication and automation (ICCCA), 2018

[20] SanthanaKrishnan J, Geetha S, "Prediction of heart disease using machine learning algorithms", 1st international conference on innovations in information and communication technology (ICIICT), 2019