International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Chronic Disease Prediction using Effective Feature Selection

Nikitha Saurabh, Tanzila Nargis

Abstract*: Healthcare is a major sector where there is demand for predictive analytics using machine learning. Healthcare will be largely benefited when useful knowledge can be transferred into timely action to manage hazardous situations in medical sector. Chronic kidney disease is a life threatening disease which can be prevented with timely right predictions and appropriate precautionary measures. In this paper, various machine learning classifiers are applied on the medical dataset to develop a prediction model to tell if a person's present medical condition can lead to the chronic stage of the disease in future. The higher prediction accuracy and decreased build time is obtained with reduced feature set attributes by applying Best First and Greedy stepwise algorithm combined with different classification techniques like Naive Bayes ,Support vector machine (SVM), J48, Random Forest, and K Nearest Neighbor(KNN).

***Index Terms*: Chronic Kidney Disease(CKD), Prediction, Classification, Machine Learning, Feature Selection**

I. INTRODUCTION

The gradual loss of functioning of kidney that can occur over a prolonged duration is known to be Chronic Kidney Disease or Chronic renal failure. As the functioning of kidney degrades, harmful amounts of fluid, electrolytes and unwanted wastes can pile up in the body. CKD often goes undetected until the disease reaches the final stage and results in a kidney failure and may lead to the death of the patient. Early and timely prediction of such fatal illness can help the doctors to take appropriate treatment to save the patients.

In healthcare ,machine learning and predictive analytics are some of the trending topics in today's world. As machine learning has already been successful in many areas ,it's use in predictive analytics can prove to be beneficial for improving and enhancing healthcare sector. However, if the prediction cannot serve any purpose in real time then it is a mere waste. In healthcare ,prediction can serve the purpose when knowledge acquired can be transformed into right preventive and precautionary measures. Machine learning is trains the computers to learn from past experiences. Computational methods are used by machine learning algorithms to gain knowledge directly from data without depending on a predefined equation. Machine learning is categorised into: supervised learning ,in which future output is predicted on the basis of given labelled input and output data, and unsupervised learning, which trains the model only on unlabelled input data. Classification of data is a form of supervised learning which is used to categorise data into desired and distinct classes. In this paper machine learning classification techniques namely SVM, KNN, Random Forest , Naive Bayes and J48 have been applied on dataset obtained from UCI repository to develop a prediction model to predict the CKD with higher accuracy. [10]The Greedy Stepwise and Best First algorithms are applied to use reduce feature set used in building the prediction model. The prediction result of the classifier models are assessed in terms of performance parameters such as root mean squared error, kappa statistics accuracy, receiver operating characteristic area ,mean absolute error, and model build time.

II. RELATED WORKS

Narandar Kumar et. al[1], has conducted the experiment on weka tool using five classifier models like J48,K-nearest neighbor, ,Random Forest, Support vector machine and Naive Bayes on 24 attribute dataset.The experimental outcome showed that Random Forest achieved higher prediction accuracy compared to other models. Haya Alasker et. al[2], evaluated in total six classifiers based on the how they classified the instances in the dataset. Decision Table ,Naive Bayes, J48, Multilayer Perceptron, KNN, One rule and were used to predict CKD. The classifiers were compared based on sensitivity ,accuracy, MSE , ROC Area ,time taken to test the model, specificity, and RMSE. Results of the tests show that Naive Bayes classification was good compared to others and also its prediction accuracy was higher in comparison to other models. Nuzrat Tazin et. al[3], has compared four classification techniques namely Decision Tree , Naive Bayes, K-nearest neighbor and Support Vector Machine and also used ranking algorithm to use reduced feature set to build the prediction model. The outcomes of experiment suggest that Decision tree performs better with higher prediction accuracy. N. A Gunarathne W.H.S.D et. al[4],has compared four multiclass classifier models namely Logistic Regression, Decision forest, Neural Network and Decision jungle on reduced feature set of input data using Microsoft Azure ML Studio. It is observed from the analysis of the outcomes that Multiclass decision Forest provides achieved higher accuracy.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Nikitha Saurabh, Department of Information Science and Engineering, NMAMIT, Nitte, Karkala, India.
Tanzila Nargis, Department of Information Science and Engineering, NMAMIT, Nitte, Karkala, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Charleonnan et al. [5] have analyzed ML classifiers like K-nearest neighbours, logistic regression, decision tree and support vector machine models to predict CKD. As the outcome of experiment, it is seen that SVM classifier achieve maximum accuracy and maximum sensitivity after training.

Sahil Sharma, Atul Sharma and Vinod Sharma [6], have assessed 400 instance dataset of 24 attributes with 12 classifiers. Models have been compared on the basis of predictive accuracy, sensitivity, precision and specificity. Experimental results show that, the decision tree achieved accuracy of 98.6%, 0.9720 sensitivity, precision and specificity of 1.

S. Vijayarani and Mr. S. Dhayanand [7], have assessed the models like SVM and ANN in terms of accuracy and build time. From the test observations, it is evident that ANN performed better than SVM.

K. R. Lakshmi et. al.[8], has used classifiers like Logical Regression ,Artificial Neural Network and Decision tree for analysis of kidney disease dataset. Computational outcomes show that Artificial Neural Network has achieved higher accuracy.

Panduranga Vital and P Swathi Baby [9] predicted kidney disease using classification models like Random Forest AD Trees, KStar ,J48, Naïve Bayes.

III. METHODOLOGY

A. *System Design*

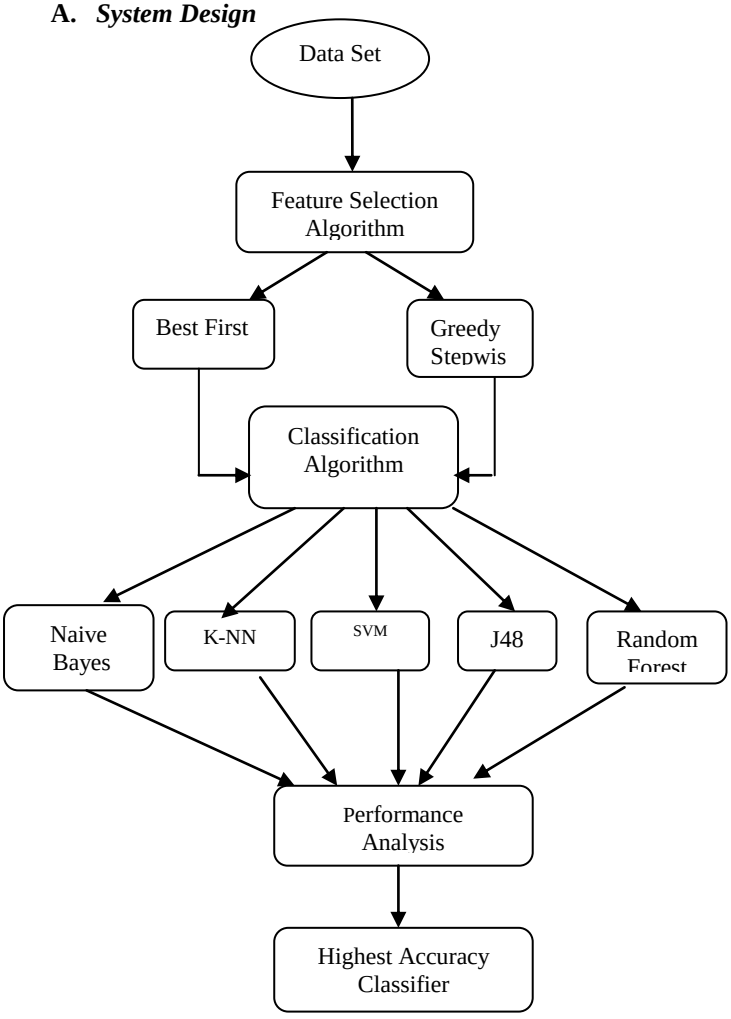

Fig. 1: System Design

The first phase of the experiment is to collect data. Next feature selection algorithm Best First and Greedy stepwise are applied which results in reduced feature set which is then used by various classification algorithms to build the CKD prediction model. Finally results of various classifiers are compared based on some performance parameters which yields best classifier with highest rate of accuracy.

B. *Data Set*

The data set used for analysis is obtained from UCI ML Repository. The classifier models in the weka platform is applied on the dataset obtained. CKD dataset has 400 instances in total, in which CKD instances are 250 and NOTCKD instances are 150.There are 24 attributes and a target attribute. Table I represents list of attributes that define the dataset.

Table I: Dataset Feature Description

Sl NO.	Attributes	Description
Numerical Attributes		
1	bu	Urea level in blood
2	rc	Count of red blood cells
3	Pcv	Volume of packed cell
4	sc	Serum Creatinine
5	hemo	Hemoglobin
6	Sod	Sodium
7	bgr	Blood Glucose Random
8	Pot	Potassium
9	bp	Blood pressure
10	wc	Count of white blood cells
11	age	Patient's age
Nominal Attributes		
12	Sg	Specific Gravity
13	rbc	Red Blood Cells
14	su	Sugar
15	Appetite	Appet
16	pc	Pus Cell

17	al	Albumin
18	htn	Hypertension
19	ba	Bacteria
20	pcc	Pus Cell Clumps
21	cad	Coronary Artery Disease
22	dm	Diabetes Mellitus
23	pe	Pedal Edema
24	Ane	Anemia
25	Class	Class

C. *Feature Selection*

The Best First feature selection algorithm uses greedy hill climbing along with backtracking to search the attribute space.. It may start looking in the forward direction starting from an empty atrribute set, or it may look backward using full attribute set or begin at any point and look in both forward and backward directions. On applying Best First to the dataset it selected 17 attributes out of 24 attributes. The selected attributes are bp , rbc ,sg, , sod, al, sc , bgr , bu , hemo, pot, pcv , appet , wc , ane, htn, pe, dm. In Greedy Stepwise feature selection proceeds in a greedy manner either in forward or backward direction through the attribute space. It may begin with empty or full attribute set or any point in the attribute space .Reduction in evaluation by addition or deletion of attributes results in stopping the feature selection process. Greedy stepwise algorithm selected 16 attributes out of 24 attributes out of 24 attributes from the dataset namely bp, ane, sg, htn ,al, rbc, dm, bgr, sc, hemo, pcv, pot , sod, wc, appet, pe.

D. *Classification Algorithms*

Bayes theorem forms the basis of Naive Bayes[12] Algorithm. Given set of classes generate a hypothesis. The values of the attribute are selected which are independent to each other based on the target. Conditional probability is an intrinsic part of Naive Bayes classifier that is used in calculation of past and present frequency values. J48 is Decision Tree[11] is a classifier that has a structure that represents a tree. It performs non linear classification. In decision tree classification category represented by leaf nodes and attributes by internal nodes and root node. To arrive at a category at the leaf node, internal nodes have to be examined to choose the right direction in the tree traversal. Random Forest[14] is a category of ensemble classification algorithm. It has a group of decision trees that work as an ensemble .Each tree in random forest gives class prediction and the class with most votes becomes the prediction model. K Nearest Neighbour(KNN)[15] classifier works on similarity measure and performs classification based data point distance value. It is a supervised model used for both classifying tasks and regression tasks. Support Vector machine[13] is a discriminative supervised classifier. Data points are plotted in a n-dimensional plane. SVM outputs optimal hyperplane that classifies new data instances.

E. *Performance Measurement Parameters*

Accuracy: It is he ratio of sum of right predictions by sum of instances to be predicted, is given by the formula:

Accuracy=(Tpositive+Tnegative)/(Tpostive+Tnegative+Fpositive+Fnegative)

Kappa Statistics: The Kappa statistics is a measure of how accurate your model is in comparison to any random model.

Kappa=K(A)-K(E)/(1-K(E))

K(A)=Percentage of agreement

K(E)=Chance of Agreement

MAE (Mean Absolute Error): It tells how much a prediction is close to possible outcome. It is the average of variation between the original values and predicted values, calculated by the formula:

$$MAE = 1/n \sum_{i=1}^{n} |fi - yi|$$

fi=prediction

yi=true value

Root Mean Squared Error(RMSE): This is the average amount of error made on the test set in the units of the output variable. The RMSE is calculated by finding the square root of square of inaccuracy between the target and predicted values.

$$Ei = \sqrt{\left(\frac{1}{n}\right)\sum_{j=1}^{n}\left(\frac{P_{(i,j)} - T_j}{T_j}\right)^2}$$

P_i is predicted output

i = fit instance

= target value of fit instance

Receiver Operating Characteristic(ROC) Area: ROC is used to represent performance graphically. It is shows how true positive rate of different classifiers are plotted against their false positive rates.

IV. EXPERIMENTAL RESULT ANALYSIS

WEKA tool with version 3.8.2 is used for analyzing the classifiers. CKD dataset from UCI repository is tested with five classifiers like J48, K-Nearest Neighbor,Naive Bayes,SVM and Random Forest in combination with Best First and Greedy Stepwise feature selection techniques. The output is assesessed based on different parameters to foretell the better performing classifier model to predict the chronic disease. The performance is compared by using RMSE, accuracy ,MAE, Kappa statistics ,ROC Area and time taken to build the prediction model.

Table II: Classification Results before Feature Selection(24 attributes)

Algorithm	Accuracy	Kappa Statistics	ROC	MAE	RMSE	Time (s)
J48	99	0.9786	0.999	0.0225	0.0807	0.03
KNN	95.75	0.9113	0.966	0.045	0.2056	0
Naive Bayes	95	0.8961	1.000	0.0479	0.2046	0
SVM	97.75	0.9526	0.982	0.0225	0.15	0.03
Random Forest	100	1	1	0.0414	0.0844	0.41

The result of classification usimg different classifiers and 24 attributes is shown in Table II.It is evident that Random Forest has the highest accuracy but more time is taken to build the prdiction model.

Table III: Classification Results Using Best First Algorithm (17 attributes)

Algorithm	Accuracy	Kappa Statistics	ROC	MAE	RMSE	Time (s)
J48	99	0.9786	0.999	0.0226	0.0806	0.01
KNN	95.75	0.9526	0.982	0.0251	0.1496	0
Naive Bayes	95.5	0.9062	1.000	0.0438	0.1887	0
SVM	98.25	0.963	0.986	0.0175	0.1323	0.01
Random Forest	99.75	0.9947	1.000	0.0379	0.0858	0.13

Table III shows the results obtained after applying Best First feature selection which yields 17 attributes. Accuracy of J48,KNN,Naive Bayes remained constant, SVM showed increase in accuracy rate, Random Forest shows a slight decrease in accuracy. After feature selection time to build all the classifier models have been reduced.

Table IV: Classification Results Using Greedy Stepwise Algorithm(16 attributes)

Algorithm	Accuracy	Kappa Statistics	ROC	MAE	RMSE	Time (s)
J48	99	0.9786	0.999	0.0225	0.0805	0
KNN	98	0.9578	0.984	0.0227	0.1411	0
Naive Bayes	97	0.937	1.000	0.0334	0.0334	0.02
SVM	98.25	0.963	0.986	0.01175	0.1323	0
Random Forest	99.75	0.9947	1.000	0.0364	0.0857	0.11

Table IV shows the result after the application of Greedy stepwise feature selection which yields 16 attributes. KNN and Naive Bayes perform better with this feature selection. Random Forest has again performed better than other classifiers with less model building time.

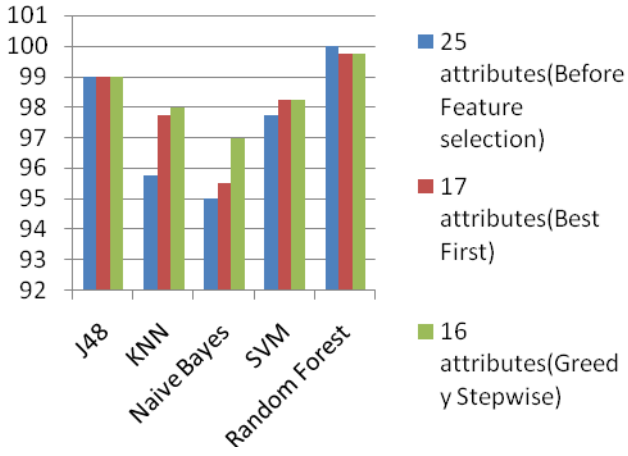

Fig. 2: Comparison of Accuracy of classifiers

From Fig. 2 it is observed that accuracy of all classifiers have improved after feature selection. Random Forest is observed to be having the highest accuracy compared to other classifiers.

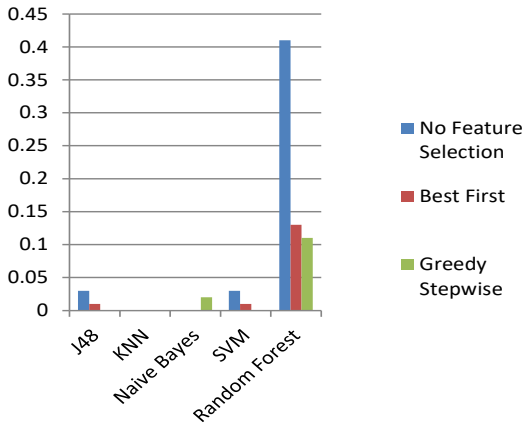

Fig. 3: Comparison of Model Build Time

From Fig. 3 it is observed that time taken to build the prediction model have improved after feature selection, especially after applying Greedy Stepwise algorithm. Most Significantly the time taken by Random Forest to build the model has reduced to a great extent after the application of feature selection algorithm.

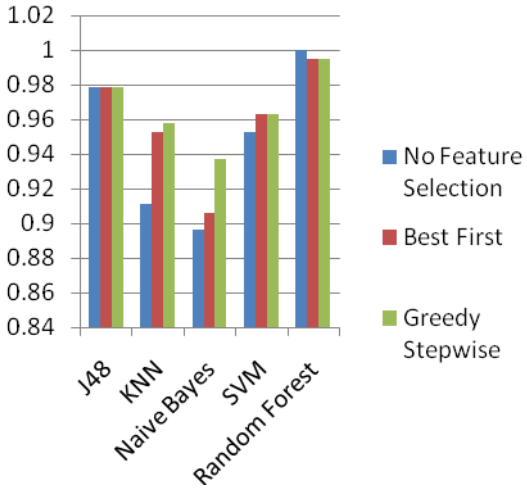

Fig. 4: Comparison of Kappa Statistics

From Fig. 4 it is observed that kappa statistics result is better for Random Forest compared to other classifiers.

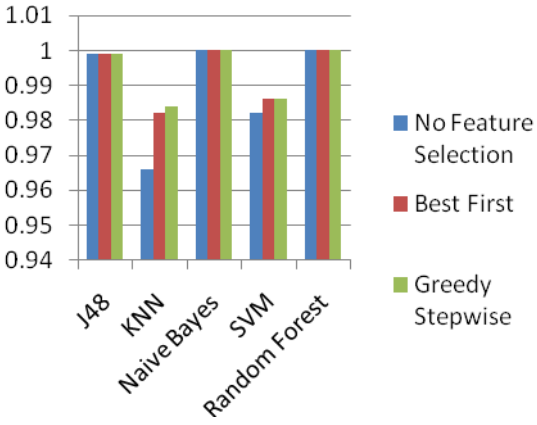

Fig. 5: Comparison of ROC

From Fig. 5 it is observed that J48,Naive Bayes and Random Forest have similar ROC values.

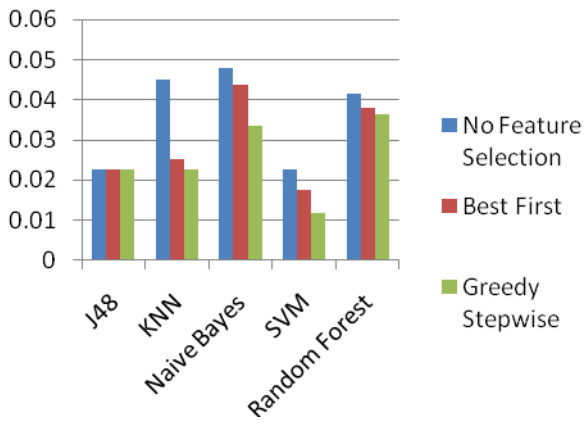

Fig. 6: Comparison of MAE

From Fig. 6 it is observed that Mean absolute error of classifiers are reduced after the applying feature selection algorithms especially after Greedy Stepwise feature selection.

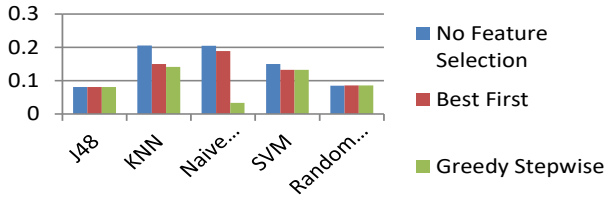

Fig. 7: Comparison of RMSE

From Fig. 7 it is observed that RMSE is reduced after applying feature selection for KNN and Naive Bayes.J48 and Random Forest have better RMSE results.

V. CONCLUSION

A Prediction model to predict whether a patient may suffer from a chronic disease of the kidney is implemented using Weka machine learning tool and medical dataset related to chronic kidney disease obtained from UCI repository. Classification models like Naive Bayes, J48, Random Forest, SVM and KNN were used in the experiment. Performance of these classifiers were analysed using various performance parameters. From the analysis of results obtained after the application of feature selection algorithm like Best first and Greedy stepwise it is evident that Greedy stepwise yielded 16 features out of 24 which proved to advance the accuracy of the classifiers and also shrink the model build time. Experimental observations also show that Random Forest algorithm provides the peak accuracy of 99.75%. In future, Experiments can be carried out to yield a better feature set with minimal number of most significant features which can be used to build a highly accurate and cost effective prediction model.

REFERENCES

1. Narander Kumar and Sabita Khatri "Implementing WEKA for medical data classification and early disease prediction" , 3rd IEEE International Conference on "Computational Intelligence and CommunicationTechnology" (IEEE-CICT 2017), Electronic ISBN: 978-1-5090-6218-8.
2. Haya Alasker, Shatha Alharkan, Wejdan Alharkan, Amal Zaki , Lala Septem Riza "Detection of Kidney Disease Using Various Intelligent Classifiers ",IEEE 2017 3rd International Conference on Science in Information Technology (ICSITech).
3. Nusrat Tazin, Shahed Anzarus Sabab and Muhammed Tawfiq Chowdhury,"Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique",2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec),Electronic ISBN: 978-1-5090-5421-3.
4. Gunarathne W.H.S.D, Perera K.D.M and 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.
5. Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, "Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques," Proc. Management and Innovation Technology International Conference (MITiCON-2016) , IEEE, Oct. 2016, doi:10.1109/MITICON.2016.8025242.
6. Sahil Sharma, Vinod Sharma, Atul Sharma, "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis," July18, 2016.
7. Dr. S. Vijayarani and Mr. S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS," International Journal of Computing and Business Research (IJCBR) , vol. 6, no. 2, 2015.
8. K. R. Lakshmi, Y. Nagesh,M. V. Krishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering & Technology (IJAET), vol. 7, no. 1, pp. 242-254, 2014.
9. P. Swathi baby, T. Panduranga Vital, "Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms", International Journal of Engineering Research and Technology, pp. 206210, Vol. 4, Issue 07, ISSN : 2278-0181, 2015.
10. Naganna Chetty, Kunwar Singh Vaisla, Sithu D Sudarsan, "Role of Attributes Selection in Classification of Chronic Kidney Disease Patients," Proc. IEEE International Conference on Computing, Communication and Security(ICCCS), IEEE, Dec. 2015, doi:10.1109/CCCS.2015.7374193.
11. S.Dilli Arasu, Dr. R. Thirumalaiselvi, "A Novel Imputation Method For Effective Prediction of Coronary Kidney Disease," Proc. 2nd IEEE International Conference on Computing and Communications Technologies (ICCCT), IEEE, Feb. 2017, doi: 10.1109/ICCCT2.2017.7972256.
12. S.Umadevi , Dr.K.S.Jeen Marseline , "A Survey on Data Mining Classification Algorithms" ,IEEE International Conference on Signal Processing and Communication (ICSPC'17) –28th& 29thJuly 2017.
13. Poonam Pandey, Radhika Prabhakar "An Analysis of Machine Learning Techniques (J48 & AdaBoost) -for Classification ",IEEE 2016 1st India International Conference on Information Processing (IICIP) ,12-14 August,2016.
14. Sarika Pachange , Bela Joglekar, Dr Parag Kulkarni "An Ensemble classifier approach for Disease Diagnosis using Random Forest ",2015 Annual IEEE India Conference (INDICON) ,17-20 December,2015.
15. Halil Yigit ,"A weighting approach for KNN classifier",IEEE,2013 International Conference on Electronics, Computer and Computation (ICECCO) ,7-9 Nov,2013.

AUTHORS PROFILE

Nikitha Saurabh Currently working as an Assistant Professor in the Department of Information Science and Engineering at NMAMIT, Nitte, Karkala, Karnataka ,India. She has 9 years of teaching and 1 year of industry experience. Her research interests include Cyber Security, Machine Learning and Artificial Intelligence.

Tanzila Nargis Currently working as an Assistant Professor in the Department of Information Science and Engineering at NMAMIT, Nitte, Karkala, Karnataka ,India. She has 2 years of teaching experience. Her research interests include Cyber Security, Network Security,Wireless Sensor Networks and Machine Learning.