IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Comparative Study of Chronic Kidney Disease Prediction using Machine Learning Techniques

Sayali Jadhav[1*], Priya Chandran[2], Suhasini Vijaykumar[3]

[1,2,3] MCA Department, BVIMIT, Mumbai University, Mumbai, India

**Corresponding Author sayalijadhav14.sj@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i6.501506 | Available online at: www.ijcseonline.org

Accepted: 13/Jun/2019, Published: 30/Jun/2019

Abstract— The healthcare industry is producing massive amount of data which need to be mine to discover hidden information for effective prediction, exploration, diagnosis and decision making. Chronic kidney disease (CKD), also known as chronic renal disease involves conditions that damage your kidneys and decrease their ability to keep you healthy. Early detection and treatment can often keep chronic kidney disease from getting worse. Machine learning techniques are commonly used to predict this situation. This research work mainly focused on finding the best classification algorithm based on different evaluation criteria like performance accuracy and root mean square error. We have performed a comparative study of the performance of machine learning algorithms J48, Support Vector Machine and Multilayer perceptron. The results show that MLP is giving minimum root mean square error value compared to J48 and SVM.

***Keywords*—** Data Mining, Neural Network, machine Learning, Kidney Disease Prediction, MLP, J48, SVM

I. Introduction

Data mining is an area of research to identify the necessary hidden information from a dataset that is being used. To identify a block of data or decision making knowledge in the database this approach uses an intermixture of technique and eradicating these data in such a way that they can be used in decision support, forecasting, and estimation. The data mining techniques of classification, clustering and association helps in extracting knowledge from large amount of data.

In the medical domain, medical data mining has the elevation potential for extracting the hidden patterns in the dataset. These patterns are used for medical diagnosis and prognosis. The medical data are globally scattered, heterogeneous in nature. The data should be concerted together. A major problem in health science or bioinformatics exploration is that managing the correct diagnosis of certain important information [1]. Various data mining classification approaches and machine learning algorithms are applied for prediction of chronic diseases. Chronic kidney disease (CKD), also known as chronic renal disease, is an abnormal function of kidney or progressive failure of renal function over a period of months or years. Often, CKD is diagnosed as a result of screening of people known to be at risk of kidney problems, such as those with high blood pressure or diabetes and those with a blood relative with CKD. In India, as per survey conducted, it has been identified that a 60% death rate is due to chronic disease. As per the survey results of 201, 17% of urban Indians have kidney diseases.[3] The Global Burden of Disease (GBD) study in 2015, identifies that chronic kidney disease is ranked 17th among the causes of deaths globally and as per world health ranking in 2016 India is on 24th rank and having high death rate as 21.56 % [3]. Hence the proactive stand for early-stage disease detection and proper treatment has become a need. This work predominantly focused on, prediction of chronic kidney disease.

Data mining and Machine learning and techniques together have proved successful in the prediction and diagnosis of various critical diseases [5][9][10][11]. Machine Learning is a rising field concerned with the study of huge and multiple variable data. There are various applications for Machine Learning, the most vital of which is data mining. Machine learning along with data mining can often be effectively applied to such problems, as they improve the efficiency of the systems and their designs [2]. The same set of features is used for the representation of every instance, in any dataset used by Machine learning algorithms. Chronic Kidney disease will be predicted using classification techniques of data mining. The classifiers used here are, Support Vector Machine (SVM) and J48 classifier and Multi Layer Perceptron (MLP). Kidney disease prediction is done using classification algorithms, applied on WEKA tool.

II. Related Work

Many classifiers of data mining are used by different researchers to predict and detect chronic kidney disease problem. Giovanni Caocci et.al [1] interpreted discrimination between an Artificial Neural Network and Logistic Regression in order to predict long term kidney transplantation Outcome. Based on the Sensitivity and specificity of Logistic Regression the comparison has been done. Dr. Vijayarani and Mr. Dhayanand [3] have considered six different attributes of renal affected disease, among those GFR i.e. Glomerular Filtration Rate, is a measurement attribute for prediction of kidney disease. They have implemented and compared two classification techniques naïve Bayes and SVM (Support Vector Machine). Their experimental results show that SVM is more accurate than Naïve Bayes. In October 2014, Abeer & Ahmad [8] have implemented two data mining classifiers SVM and Logistic Regression (LR) Their results showed that SVM has more accuracy than other techniques with 93.14 percent. Joshi et al. [5] have done diagnosis and prognosis of breast cancer using classification rules. By comparing classification rules such as Bayes Net, Logistic Model Tree (LMT), Multilayer Perceptron, Stochastic Gradient Descent (SGD), Simple Logistic, Sequential Minimal Optimization (SMO), AdaBoostM1, Attribute Selected, Classification via Regression, Filtered Classifier, Multiclass Classifier and J48, they suggested that LMT Classifier gives more accurate diagnosis i.e. 76 % healthy and 24 % sick patients. Ashfaq et.al,[1] have presented a work using machine learning techniques, namely, Support Vector Machine and Random Forest (RF). These were used to study, classify and compare liver, cancer, and heart disease data sets with varying kernels and kernel parameters. Results of RF and SVM were compared for different data sets such as liver disease dataset and heart disease dataset etc. The results with different kernels were tuned with proper parameter selection and analysed to establish better learning techniques for predictions. It is concluded that varying results were observed with SVM classification technique with different kernel functions.

III. Research Methodology

This section describes the proposed methodology for data mining from CKD dataset. The kidney disease dataset has been used for the analysis of kidney disease. This dataset contains four hundred instances and twenty-five attributes are used in this comparative analysis. Classification techniques are applied to all features and selected features. In order to carry out experiments and implementations, WEKA is used as the data mining tool to classify the accuracy on the basis of datasets by applying different algorithmic approaches. In this work, we have used different machine learning algorithms namely SVM and J48 classifier and Multilayer perceptron to predict the survivability of Chronic-Kidney disease through classification algorithms.

A) J48 Algorithm

J48 classifier is a simple C4.5 decision tree for classification. It is supervised method of classification. It creates a small binary tree. It is univariate decision tree. It is an extension of ID3 algorithm [3]. In this classifier divide and conquer approach is used to classify the data [5]. There are 2 approaches univarient and multivarient. Univarient decision tree, in this technique, splitting is performed by using one attribute at internal nodes. It divides the data into range based on the attribute value for that value that are found in training sample. As this approach is range based and univarient, it does not perform better than multivarient approach.

Some basic steps are given below to construct tree:-

- First, check whether all cases belongs to same class, then the tree is a leaf and is labelled with that class.
- Then, for each attribute, calculate the information and information gain.
- And then last, find the best splitting attribute (depending upon current selection criterion)

Entropy is used in this process. Entropy is a measure of disorder of data. Entropy is measured in bits, nats or bans. This is also called measurement of uncertainty in any random variable. Entropy for any P can be calculated as:

$$Entropy\,(p) = -\sum_{\mathrm{j}}{}^{\mathrm{n}} = 1\,\frac{|pj|}{|p|}\,log\,\frac{|pj|}{|p|} \qquad (1)$$

The conditional entropy is:-

$$Entropy(j|p) = \frac{|pj|}{|p|}\,log\,\frac{|pj|}{|p|} \qquad (2)$$

Information Gain is used for measuring association between inputs and outputs. It is a state to state change in information entropy. Finally information gain can be calculated as:-

$$Gain\,(p,j) = Entropy(p\; - \; Entropy(j|p)) \qquad (3)$$

As this is decision tree approach it is very much useful in predicting the values. J48 accuracy of correctly classified instance is much more than that of the other algorithms which are univarient in nature.

B) Support Vector Machine

SVM is a powerful, based on linear and nonlinear regression. Support Vector Machines, a method for the classification of both linear and nonlinear data. [2] In a casing, a SVM is an algorithm that works as follows. It uses a nonlinear mapping to renovate the unique training data into a higher dimension [1].SVMs area unit largely used for learning classification, regression or ranking operate. It is based on the concept decision boundaries. A decision plane is one that separates between a set of objects having different class memberships [5]. Surrounded by this new dimension, it examines for the linear optimal separating hyperplane. With a suitable nonlinear mapping to a necessarily high dimension, data from two classes can always be separated by a hyperplane. Hyperplane can be taken as a line that linearly separates a set of data. SVM is that the most strong and precise classification technique, there are a unit several issues [6]. More the margin between our data point, the more we are accurately classified. The training time of even the fastest SVMs can be exceedingly slow, they are extremely accurate to their ability to model complex nonlinear decision boundaries [16]. They are much less prone to over fitting than other methods. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study [5]. The kernel function is expressed as follows:

$$K(\bar{x},\bar{x}_{\mathrm{i}}) = exp((-||\vec{x} - \vec{x}_{\mathrm{i}}||_2)/2a_2) \quad (4)$$

In the above equation, the kernel width parameters control the amplitude of the Gaussian function reflecting the generalization ability of SVM. The regularization parameter C is censurable for inhibiting transaction between maximizing the margin and minimizing the training error. SVMs have frequently been found to provide maximum classification accuracies than other widely used pattern recognition techniques, such as the maximum likelihood and the multilayer perceptron neural network classifiers.

C) Multilayer Perceptron

This classification algorithm belongs to feed forward artificial neural network. The Multilayer Perceptron is one of most important class of neural networks, consisting of an input layer, one or more hidden layers, and the output layer [5]. MLP distinguish data that are not linearly separable. MLPs have been applied successfully to solve difficult and diverse problems, by training them in a supervised manner using a well-known. This is based on the error correction learning rule. As such, it may be viewed as a generalization of an adaptive filtering algorithm

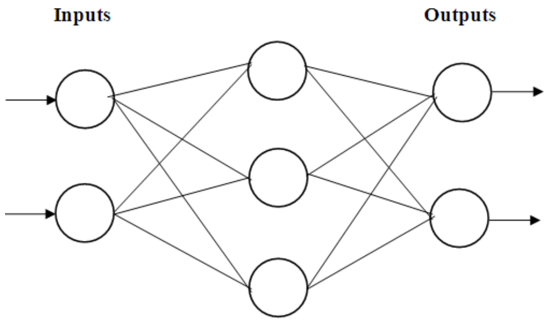


Figure 3.1: Structure of Multilayer Perceptron

Figure 3.1 shows the basic structure of multilayer perceptron model. It consists of an input layer, an arbitrary hidden layer and an output layer. The input layer receives the signal, the hidden layer act as a computational engine capable of approximating any condition and thus giving the results to the output layer. It is a form of supervised learning and they train the input set and learn from the correlation predicted from the input and output and use them for prediction on test data. Multilayer perceptron model consist of three or more than three layers and all the nodes in the layers are fully interconnected. A side from the input nodes, every node could be a nerve cell (or process element) with a nonlinear activation perform. MLP classification urinary organ dataset utilizes a supervised learning technique known as back propagation for coaching urinary organ the network. MLP could be a modification of the quality linear perceptron and might distinguish knowledge that isn't linearly dissociable urinary organ dataset method [6].

IV. Results and Discussion

The input dataset were classified using different methods of data mining. Performance of various algorithms was studied. These methods include J48, SVM, and Multilayer perceptron. Classification is a data mining algorithm which finds out the output of a new data instance. In this paper the experimental study is conducted on various classification algorithms and best algorithm is identified for chronic kidney dieses.

A confusion matrix is a table that is usually used to describe the performance of a classification model on a set of test data for which the true values are known In general, Positive = identified and negative = rejected [3]. Therefore:
True positive (TP) = correctly identified
False positive (FP) = incorrectly identified
True negative (TN) = correctly rejected
False negative (FN) = incorrectly rejected

Sensitivity and specificity are statistical measures of the performance of a binary classification test. Sensitivity (TPR) measures the proportion of positives that are correctly identified. Specificity (TNR) measures the proportion of negatives that are correctly identified.
The sensitivity and specificity is calculated by following formulas:

$$Sensitivity\ (TPR) = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity\ (TNR) = \frac{TN}{N} = \frac{TN}{FP + TN} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

The number of real positive cases in the data is denoted by P. The number of real negative cases in the data is denoted by N.

A) Dataset Analysis

For prediction of chronic kidney disease we have used Chronic-Kidney-Disease dataset for prediction and classification [7]. The dataset used for our experiment contains 25 (24 + class = 25 (11 numeric, 14 nominal)) attributes and 400 instances. In order to obtain better accuracy 10 fold cross validation was performed. For each classification we selected training and testing sample randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The thrust classifications and accuracy used by are:

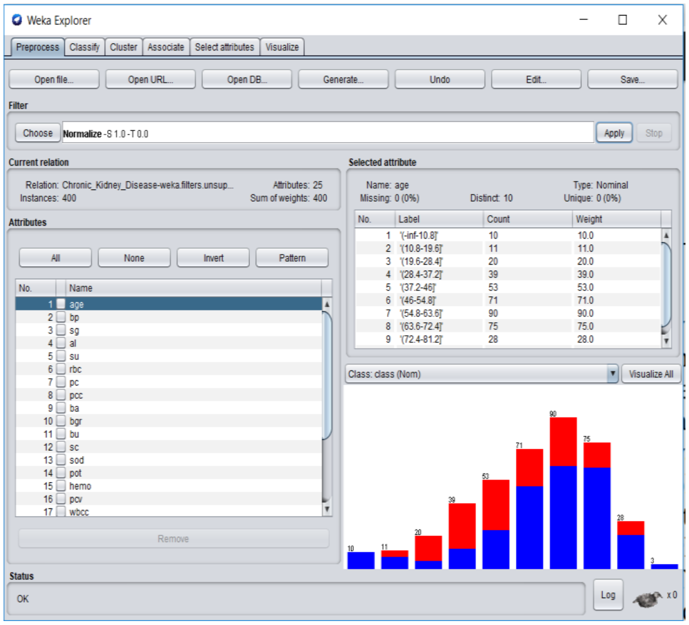
- Correctly Classified Accuracy: It shows the accuracy percentage of test that is correctly classified.
- Incorrectly Classified Accuracy: It shows the accuracy percentage of test that is incorrectly classified.
- Mean Absolute Error: It shows the number of errors to analyze algorithm classification accuracy.
- Kappa statistics: It measures inter-rater agreement for qualitative items.

Table 3.1 shows the description of the attributes in Chronic-Kidney-Disease dataset.

Table 3.1 Attributes of CKD Dataset

Attribute symbols	Attribute description	Attribute type
age	age in years	numerical
bp	Blood Pressure (bp in mm/Hg)	numerical
sg	Specific Gravity (1.005,1.010,1.015,1.020,1.025)	nominal
al	Albumin (0,1,2,3,4,5)	nominal
su	Sugar (0,1,2,3,4,5)	nominal
rbc	Red Blood Cells (normal, abnormal)	nominal
pc	Pus Cell (normal, abnormal)	nominal
pcc	Pus Cell clumps (present, not present)	nominal
ba	Bacteria (present, not present)	nominal
bgr	Blood Glucose Random (bgr in mgs/dl)	numerical
bu	Blood Urea (bu in mgs/dl)	numerical
sc	Serum Creatinine (sc in mgs/dl)	numerical
sod	Sodium (sod in mEq/L)	numerical
pot	Potassium (pot in mEq/L)	numerical
hemo	Hemoglobin (hemo in gms)	numerical
pcv	Packed Cell Volume	numerical
wc	White Blood Cell Count (wc in cells/cumm)	numerical
rc	Red Blood Cell Count (rc in millions/cmm)	numerical
htn	Hypertension (yes, no)	nominal
dm	Diabetes Mellitus (yes, no)	nominal
cad	Coronary Artery Disease (yes, no)	nominal
appet	Appetite (good, poor)	nominal
pe	Pedal Edema (yes, no)	nominal
ane	Anemia (yes, no)	nominal
class	Class (ckd, notckd)	nominal

The below Figure 3.1 shows Normalized Dataset. Then one by one the classification algorithms J48, SVM and MLP are applied on the filtered dataset. The attributes and their distributions is shown in figure 3.2



Filter

Normalize -S 1.0 -T 0.0

Current relation

Relation: Chronic_Kidney_Disease-weka.filters.unsup... Attributes: 25
Instances: 400 Sum of weights: 400

Attributes

No.		Name
1	□	age
2	□	bp
3	□	sg
4	□	al
5	□	su
6	□	rbc
7	□	pc
8	□	pcc
9	□	ba
10	□	bgr
11	□	bu
12	□	sc
13	□	sod
14	□	pot
15	□	hemo
16	□	pcv
17	□	wbcc

Selected attribute

Name: age Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-10.8]'	10	10.0
2	'(10.8-19.6]'	11	11.0
3	'(19.6-28.4]'	20	20.0
4	'(28.4-37.2]'	39	39.0
5	'(37.2-46]'	53	53.0
6	'(46-54.8]'	71	71.0
7	'(54.8-63.6]'	90	90.0
8	'(63.6-72.4]'	75	75.0
9	'(72.4-81.2]'	28	28.0

Class: class (Nom)

Status

OK

Figure 3.1: Normalized Dataset

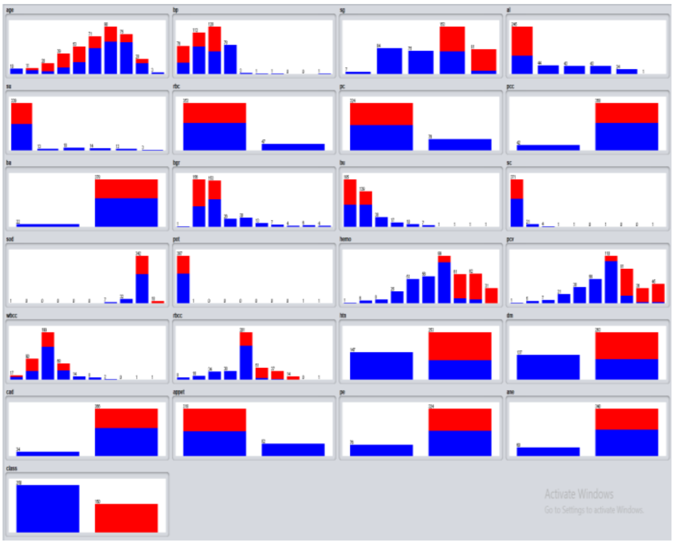


Figure 3.2: Attribute Distribution

Confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice versa).The matrix is represented in the form of nxn, where n is the number of classes. The accuracy of each classification algorithms can be calculated from that. In our experiment we have two classes, and therefore we have a 2x2 confusion matrix, which is shown in table 3.2. The confusion matrix obtained for J48, SVM and MLP are given in table 3.3, 3.4 and 3.5 respectively.

Table 3.2: Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	TP	FN
b (NOT CKD)	FP	TN

Table 3.3: J48 Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	241	9
b (NOT CKD)	0	150

Table 3.4: SVM Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	246	4
b (NOT CKD)	1	149

Table 3.5: MLP Confusion Matrix

	a (CKD)	**b (NOT CKD)**
a (CKD)	249	1
b (NOT CKD)	0	150

B) Predictive Performance of classifier

Evaluation of performance is compared using predictive Accuracy, Mean absolute error, Root mean squared error and Receiver Operating Characteristic (ROC) Area and Kappa statistics.

Table 3.6: Predictive Performance of classifier

Evaluation Criteria	**Classifier**		
	J48	**SVM**	**MLP**
Timing to Build Model (sec)	0.4	0.07	5.29
Correctly classified Instance	97.75	98.75	99.75
Incorrectly Classified Instance	2.25	1.25	0.25
Predictive Accuracy	97.75	98.75	99.75
Kappa statistics	0.95	0.97	0.99
Mean Absolute error	0.0362	0.0125	0.0085
Root mean square	0.145	0.112	0.062
Relative absolute error	7.72	2.67	1.81
Root relative squared error	30.02	23.1	12.86

Table 3.6 depicts the performance of each algorithm based on different evaluation criteria. Here the prediction accuracy is 97.75, 98.75, and 99.75 for J48, SVM, and MLP respectively and MLP is having the minimum root mean square error with a value 0.062.

C) Perform Analysis

In figure 3.3 and 3.4, the performance analysis is identified with the help of a graph. The fig 4.4 shows the correctly classified instances, the predictive accuracy for these three techniques are 97.75, 98.75 and 99.75 for J48, SVM and MLP respectively.

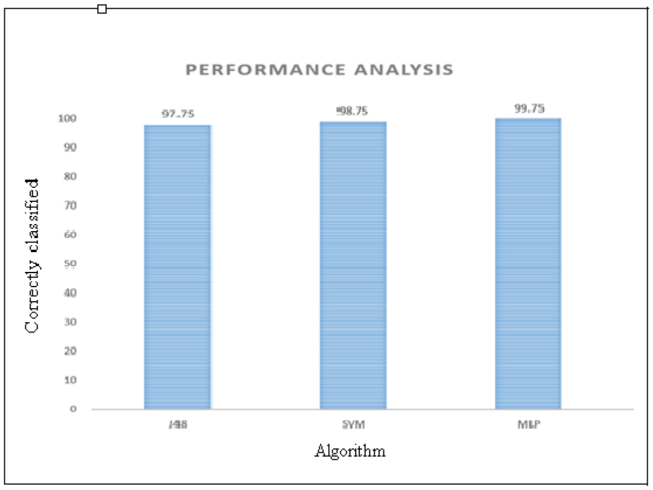


Figure 3.3: Correctly Classified Instances

Figure 3.4 shows, mean absolute error for these techniques. It concludes that J48 have highest error compare to other two algorithms, SVM and MLP. MLP identified the minimum error, 0.0085, which means that it has highest predictive accuracy.

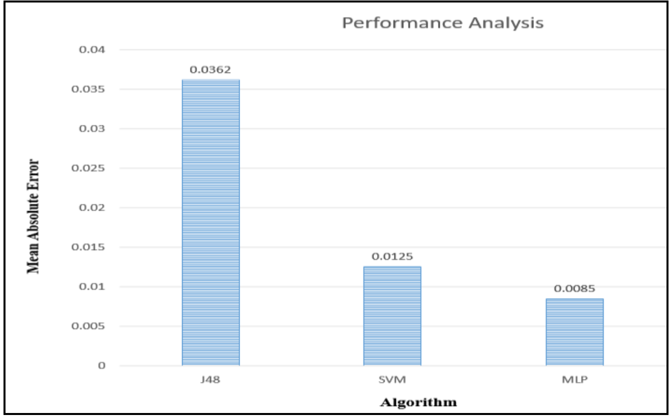


Figure 3.4: Error Classification

V. CONCLUSION

The main objective of this research is to predict chronic kidney disease using machine learning algorithms. We have used Multilayer Perceptron, SVM and J48 in our experiments. The performances of the algorithms have been compared with classification accuracy to each other on the basis of correctly classified instances. From the experimental result, MLP performs best in classifying process than the SVM algorithm with root mean square error value 0.062. In this experiment, MLP gives better classification accuracy and prediction performance to predict chronic kidney disease (CKD) using relevant dataset available than J48 and SVM algorithm.

REFERENCES

[1]. Dr. S. Vijayarani1 , Mr.S.Dhayanand2 Assistant Professor1 , M.Phil Research Scholar2 “Kidney Disease Prediction Using SVM And ANN Algorithms”in 2015 international Journal of Computing and Business Research (IJCBR) Volume 6 Issue 2 March 2015

[2]. Parul Sinha, Poonam Sinha “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM” in 2015 International Journal of Engineering Research & Technology (IJERT) Vol. 4 Issue 12, December-2015

[3]. Harshali Patil, Manisha Divate “Kidney Disease Detection In Indian Patients In An Early Stage Using Weka Tool” in 2018 Proceedings of International Conference on Advances in Computer Technology and Management (ICACTM) In Association with Novateur Publications IJRPET-ISSN No: 2454-7875 ISBN No. 978-81-921768-9- 5 February, 23rd and 24th, 2018

[4]. N. Afhami “Prediction of Diabetic Chronic Kidney Disease Progression Using Data Mining Techniques”in 2018 International Journal of Computer Science Engineering (IJCSE), Vol. 7 No.02 Mar-Apr 2018

[5]. Lambodar Jena, Narendra Ku. Kamila “Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease”, International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-11) Research Article November 2015

[6]. S.S. Senthil priya1, P. Anitha “ Comparison Of Feature Selection Methods For Chronic Kidney Data Set Using Data Mining Classification Analytical Model”,International Research Journal Of Engineering And Technology (Irjet), Volume: 06 Issue: 2 | Feb 2019

[7]. https://archive.ics.uci.edu/ml/datasets.php

[8]. El-Houssainy A.RadyaAyman S.Anwarb “Prediction of kidney disease stages using data mining algorithms”**,** Informatics in Medicine Unlocked 15 (2019) 100178

[9]. Pushpa M. Patil “Review On Prediction Of Chronic Kidney Disease Using Data Mining Techniques”, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 5, May 2016

[10]. Sujata Drall, Gurdeep Singh Drall, Sugandha Singh, Bharat Bhushan Naib, “Chronic Kidney Disease Prediction: A Review”, International Journal of Management, Technology And Engineering, ISSN No : 2249-7455, Volume 8, Issue V, May/2018

[11]. Dr. S. Sasikala1, Dr. S. Jansi2, Ms. S. Saranya3,Ms. P. Deepika4, Ms. A. Kiruthika “Anticipating the Chronic Kidney Disorder (CKD) using Performance Optimization in AdaBoost and Multilayer Perceptron”, Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-2, 2017

[12]. International Journal of Scientific Research in Network Security and Communication (ISSN: 2321-3256)