Prediction of Chronic Kidney Disease Statistics Using Data Mining Techniques

Rajesh S. Walse
School of Cmputational Sciences
Swami Ramanand Teerth Marathwada University,
Nanded, India,
rajeshwalse@gmail.com

Aniket A. Muley
School of Mathematical Sciences
Swami Ramanand Teerth Marathwada University,
Nanded, India
aniket.muley@gmail.com

Gajanan D. Kurundkar
Department of Computer Science
Shri Gurubuddhi Swami Mahavidyalaya, Purna Dist.
Parbhani, S.R.T.M. University, Nanded, India
gajanan.kurundkar@gmail.com

Parag Bhalchandra
School of Computational Sciences
Swami Ramanand Teerth Marathwada University
Nanded, India
srtmun.parag@gmail.com

ABSTRACT

In this research, the Apriori associator in the WEKA data mining tool used for preprocessing, exploring, and analyzing the chronic kidney-related data. The minimum matrix or confidence value in the association rule mining supporter, confidence number of cycles performed the role of preparation of Rules. This research is carried out by formatting and found ten best rules. The rules create x belongs to y attributes; the constant output of Apriori is to set the best standards by using its value and over caste, and its production shows the rules in the form of the model. At the time of execution, minimum support is 0.2 (80 instances), and the minimum metric <confidence> level is more significant than 0.9. The association is always on affected data. Moreover, Apriori is used to Nominal data or binary data. Knowledge extraction from the chronic kidney disease database becomes one of the vital processes for their research development issues. It plays a crucial role in the medical field as well as it plays an essential role in medical Industries for future planning and further prediction issues. The significant association rules obtained from this study can be useful for Nephrologists doctors and Medical Industries.

Keywords— Data Mining; Classification; preprocess; Association; Apriori; WEKA; CKD.

I. INTRODUCTION

Numerous Computer Science techniques of data mining and machine learning is to be used to learning the power of various parameters and to make predictions of the based on different types of data base. The DM techniques are the method of finding the hidden patterns from the big and tedious data. It may provide a crucial role in decision making for complex not only agriculture but also health-related problems NCBI [17].

A. **Significance of the study:** Jnephrol [18] with increasing life expectancy and the prevalence of lifestyle disease, the US has seen a 30% considerable growth in the prevalence of chronic kidney disease (CKD) in the last decade. Unfortunately, from India, there is no longitudinal study and limited data on the incidence of CKD.

B. **Data Analysis techniques:** This research paper experiment was carried out on chronic kidney disease patient, Today's chronic kidney disease patient in India is increased day today because of their lifestyle, eating habits, earlier chronic or acute kidney disease patient is very less, and kidney fails growth rate is very less, i.e., based on different health problem maximum kidney problem was increased based on Hypertension, and Blood sugar is the kidney patients. Therefore, nowadays lots of Nephrologists doctors and Medicine Industries is to be used to Predict kidney problems of patients on the basis of different Machine Learning algorithms; many researchers are working on kidney patients database, once the doctor has predicted the leading cause of kidney damage based on Machine Learning tools by applying Clustering and Classification techniques, once doctor predicts the cause. Definitely, this will help kidney patients and control further losses of the kidney. Data mining plays an active role in predicting future kidney-related health problems. The main intention of the Experiment is using the WEKA tool first

preprocess the data based on Current relation, attributes, selected attributes, class, and visualize every attribute using their Missing value, distinct, type, and unique. Similarly, the researcher uses the association Apriori algorithm to discover new rules; therefore, this type of formed rules and preprocess will help to propose the best model with higher accuracy for Clustering and Classification of data.

Data Fusion sampling multi-resolution analysis -De-noising Feature-Extraction Normalization - Dimension reduction - Classification Clustering - Visualization Validation, Data cleaning: The removal of noise and inconsistent data with data integration of the combination of multiple sources of data. During the data selection, the data relevant for analysis is retrieved from the database also the data mining use of ingenious methods to excerpt patterns from data with pattern evaluation and to show the result in the form of visualization technique which will be helpful for knowledge presentation is used to present the extracted knowledge to the end-user.

Saggar et al. [13] studied and optimized association rule mining using improved Genetic Algorithms. Gandhi et al. [7] forecasted rice crop yield of a tropical wet and dry climatic zone of India using data mining techniques. Bharara et al. [3] reviewed knowledge extraction for business operations using Data Mining. Sujatha and Isakki Devi [14] focused on crop yield forecasting using classification techniques. Ariff et al. [2] studied RFID based systematic livestock health management system. Jinyin [9] performed a novel cluster center fast determination clustering algorithm. Dilli Arasu and Thirumalaiselvi [1] dealt with novel imputation methods for the effective prediction of coronary kidney disease. ZouChuan et al. [4] performed an applied study of Guangdong provincial hospital of traditional Chinese medicine, Guangzhou and explore clustering analysis for syndrome evolution law of peritoneal dialysis patients. Kunwar et al. [11] studied and analyzed Chronic Kidney Disease using data mining classification techniques. AnhLuong [5] applied K-Means Approach to Clustering disease Progressions. Formaggio et al. [6] performed applied Object-Based Image Analysis (OBIA) and Data Mining (DM) in the Landsat time series in intensive agricultural regions. Sabri [8] used data mining techniques for segmenting customers' information. Kumar and Lhatri [12] used WEKA for medical data classification and early disease prediction. Khanna [10], NCBI [17] performed a study on the economics of Dialysis in India. J Nephrol [18] studied the prevalence of chronic kidney disease in India, and where are we heading? Ubolthaм et al. [15] performed a diagnostic study of acute kidney injury using the KDIGO guideline approach.

II. RESEARCH METHODOLOGY

In the proposed study, Fig. 1 explores the flow diagram of our study. It is the systematic way, which explores the scientific representation of our study.

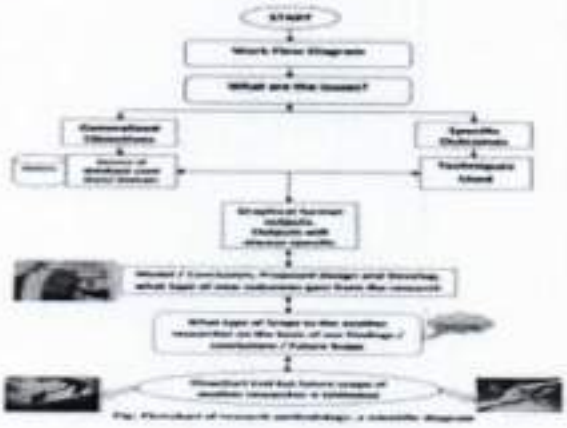

Fig: 1 The research work flow

III. EXPERIMENT SETUP

The Chronic kidney failure disease dataset has collected from standard UCI machine learning repository, and for further database analyzed using WEKA. On that basis, we are getting the exact graphical

output based on their classification of all 25 attributes of 400 instances. These generated rules are given the direction for further research as well as the best rule for new clustering and classification of data for prediction means how to apply relation to chronic kidney disease patients based on setting rules and as well as based on probability. If the maximum probability, choose that rule, if the Minimum metric <confidence> level is more significant if it is greater than 0.9.

- **Data set:** The clinical data of 400 records considered for analysis has taken from the UCI Machine Learning Repository. The data obtained after cleaning and removing missing values. The data has implemented using Rapid Miner Tool. There are 25 attributes in the dataset. Attributes play an essential role; on that basis, the researcher set further prediction on different instances based on standard lab values. The numerical attributes include age, blood pressure, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packaged cell volume, WBC count, RBC count. The nominal attributes include specific gravity, albumin, sugar, RBC, Pus cell, Pus cell clumps, bacteria, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia and class (CKD and Non-CKD)Number of instances: 400, number of attributes: 25 class (CKD, Non-CKD) Missing attribute value: yes, Class distribution (63% for CKD and 37% for not CKD)

Fig: 2 Data with missing observation

Fig: 3 Clean data set

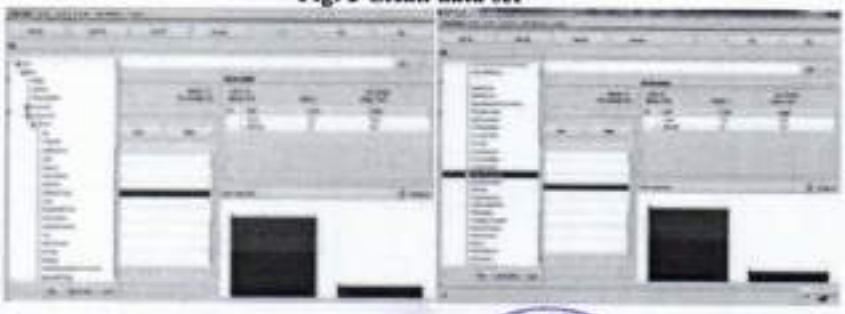

Fig: 4 Nominal type

Figure 2 shows the instances with missing values do not have to be removed; the researcher replaces the missing values with proper average value and is called imputing missing values. It is common to assign missing values with the mean of the exponential distribution, process this research technique efficiently in WEKA by using the Replace Missing Values filter (Fig. 3). Apriori apply on only nominal data, and binary data sets only this two types of data Association of Apriori algorithm executes the data to find best rules, and before providing data to apriori algorithm convert some numerical attributes to Nominal by using choose filter unsupervised attributes by turning attributes numerical to binary type (Fig. 4).

IV. MATERIAL AND METHOD

Here, we are getting detail to preprocess steps of all 25 attributes for 400 instances, i.e., for records.

1. **Relevant Information: (Total No of 25 Attributes)**

Table: 1 Attributes and its units

Sr. No.	Attributes		Type	Type of attributes and its unit
1	age	Age	numerical	Age in years
2	bp	Blood pressure	numerical	Bp in mm/Hg
3	sg	Specific gravity	nominal	1.005,1.010,1.015,1.020,1.025
4	al	Albumin	nominal	0,1,2,3,4,5
5	su	Sugar	nominal	0,1,2,3,4,5
6	rbc	Red blood cells	nominal	normal, abnormal
7	pc	Pus cell	nominal	normal, abnormal
8	pcc	Pus cell clumps	nominal	present, notpresent
9	ba	Bacteria	nominal	present, notpresent
10	bgr	Blood glucose random	numerical	mgs/dl
11	bu	Blood urea	numerical	mgs/dl
12	sc	Serum creatinine	numerical	mgs/dl
13	sod	Sodium	numerical	mEqL
14	pot	Potassium	numerical	mEqL
15	Hemo	Hemoglobin	numerical	gms
16	pcv	Packed cell volume	numerical	Numerical
17	wc	White blood cell count	numerical	cells/cumm
18	rc	Red Blood Cell count	numerical	millions/cmm
19	htn	Hypertension	nominal	yes, no
20	dm	Diabetes mellitus	nominal	yes, no
21	cad	Coronary artery disease	nominal	yes, no
22	appet	appetite	nominal	good, poor
23	pe	Pedal edema	nominal	yes, no
24	ane	Anemia	nominal	yes, no
25	class	Class	nominal	ckd, notckd

2. **Number of Instances:** 400 (250 CKD, 150 notckd)
3. **Number of Attributes:** 24 + class = 25 (11 numeric,14 nominal)
4. **Missing Attribute Values:** Yes (Denoted by "?")
5. **Class Distribution:** (2 classes)

Class	Number of instances
ckd	250
notckd	150

IV. RESULTS AND DISCUSSION

The available primary data of chronic kidney disease and some necessary information on kidney-related information are to be collected from UCI standard repository. This is mainly required to design and

develop the standard predictive model in terms of algorithm, which will be helpful for any chronic kidney disease patients of further prediction neither nephrologists nor medical industry. The research in terms of prediction on CKD patient database is to be required for new prediction, to control further damage of kidney, day by day CKD patient is increased in India. However, further, it helps doctors and patients to control new diseases and prevent further misery. Processing the attribute relation file format (ARFF), the list of all CKD attributes, its statistics, and another parameter can be utilized and replace all missing values as shown in below Figure:

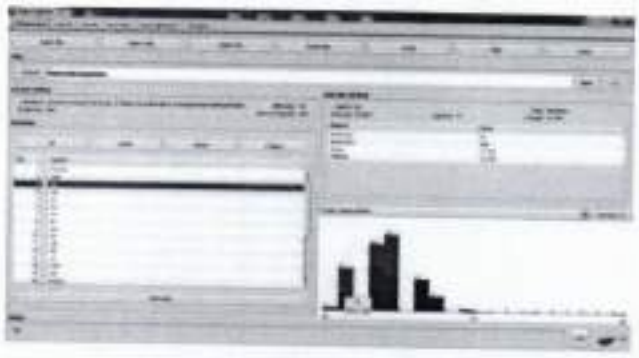

Fig: 5 Processed ARFF file

In preprocess filter in unsupervised attributes and then convert all Numeric to Nominal before applying the Apriori algorithm for the generation of the best rule.

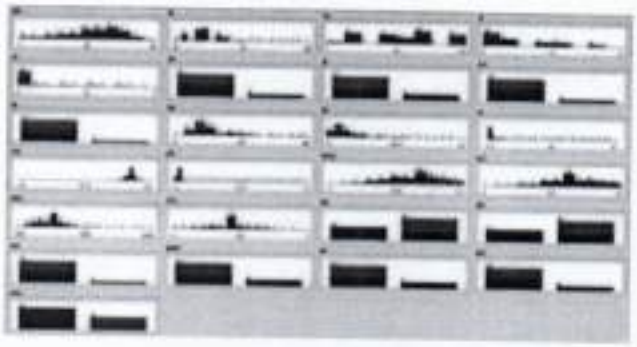

Fig: 6 Graphical visualization of processed CKD attributes

The processed data using Weka can be analyzed using different data mining techniques like Classification, Clustering, Association rule mining and Visualization algorithms. Figure 6 shows 25 attributes are shown in Table 2 processed. Further, these attribute 1 to 25 are visualized into a two-dimensional graphical format respectively.

A. **Preprocess for Classification:**

Table: 2 Relation: -Chronic_kideny_disease Attributes: 25, Instances: 401 and sum of weights: 400

Co-ordinator
IQAC
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn) Dist. Parbhani - 431511 (M.S.)

PRINCIPAL
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn.) Dist. Parbhani

Sr. No.	Attribute	Details of selected attributes			
		Missing	Distinct	Type	Unique
1	Age Missing	0	77	Nominal	16(4%)
2	Blood Pressure	0	11	Nominal	3(1%)
3	Specific Gravity	0	6	Nominal	1
4	Albumin	0	7	Nominal	1
5	Sugar	0	7	Nominal	0
6	Red blood cell	0	2	Nominal	0
7	Pus Cell	0	2	Nominal	1
8	Pus Cell clumps	0	2	Nominal	0
9	Bacteria	0	2	Nominal	0
10	Blood Glucose Random	0	147	Nominal	65(16%)
11	Blood Urea	0	119	Nominal	55(14%)
12	Serum Creatinine	0	85	Nominal	41(10%)
13	Sodium	0	35	Nominal	7(2%)
14	Potassium	0	41	Nominal	8(2%)
15	Hemoglobin	0	116	Nominal	28(7%)
16	Packed Cell Volume	0	43	Nominal	8(2%)
17	White Blood Cell Count	0	90	Nominal	51(8%)
18	Red Blood Cell Count	0	46	Nominal	3(1%)
19	Hypertension	0	2	Nominal	0
20	Diabetes Mellitus	0	2	Nominal	2
21	Coronary Artery Disease	0	2	Nominal	1
22	Appetite	0	2	Nominal	1
23	Pedal Edema	0	2	Nominal	1
24	Anemia	0	2	Nominal	1
25	Class	0	2	Nominal	1

Table Shows the performance criteria values. In WEKA, preprocess is one of the options; this is useful for getting detail information or statistics based on the data set. In this research, the kidney dataset processed with different attributes (25), which contains 400 rows, i.e., instances and 25 attributes, means columns. The researcher has select every attributes to displays type of attributes, the type means Nominal, how many missing values present in the data set for each attribute viz. instances, how many distinct values are present in the dataset, distinct means different values, if we select attribute is shown Nom- in front of attribute-nom means nominal type. If data is numeric and select as an attribute, it will show the statistics report in the form of min, max, the mean and standard deviation of Statistics and Value and it is to be selected as the attribute Class, then it shows label count and weight in the form of true or false or yes or no.

Fig: 6 The graphical result (X: Age and Y: Class CKD and not-CKD)

Fig: 7 The graphical result (X: Blood Pressure and Class Y: CKD and not-CKD)

Fig: 8 The graphical result (X: Serum Creatinine and Class Y: CKD and not-CKD)

Figure 6-8 explore Plot the matrix and visualize in all the above three graphs are shown to Classify and to form a Cluster, which are based on chronic kidney disease dataset and to classify the kidney patients based on their different age groups (i.e., X: Age and Y: Class CKD and not-CKD) similar Figure shows the result on the relation of (Blood Pressure and Class (i.e., X: Blood Pressure and Class Y: CKD and not-CKD) and last one Figure shows the graphical result on the relation on (X: Serum Creatinine and Class Y: CKD and not-CKD).

The Prediction based on Clustering and Classification is significant for nephrologists doctors and medical drug industries for further treatment of CKD patients; further, how many are the CKD patient anyhow many are a not-CKD. The same technique is to be applied to different Attributes and Instances.

A. **Analysis:**

Summary of Associate for discovering association rule: In WEKA preprocess, in this research, we are explaining how to explore and analyze the Chronic Kidney related data in the Apriori associator. In this algorithm, the minimum matrix or confidence value in Association rule mining supporter confidence number of cycle performing the role of preparing Rules. The rules create x belongs to y attributes; the constant output of Apriori is to set the best rules by using its value and over caste, and its output shows the rules in the form of the model to describe every rule. The association is always on applied data. Moreover, Apriori always applied to Nominal data or Binary Data.

For research, get some components, the central part of the apriori algorithm is delta, metric type, primary metric, number rules, and upper bound minimum support. Apriori algorithm runs between two bounds upper bound minimum support and lowers bound minimum support, which we have given us in 0.2 and 0.1. Apriori algorithm runs between them a frequency of delta 0.05. The primary use of metric types is how to use rank our rules and associations more the confidence to list the attributes need 10 rules.

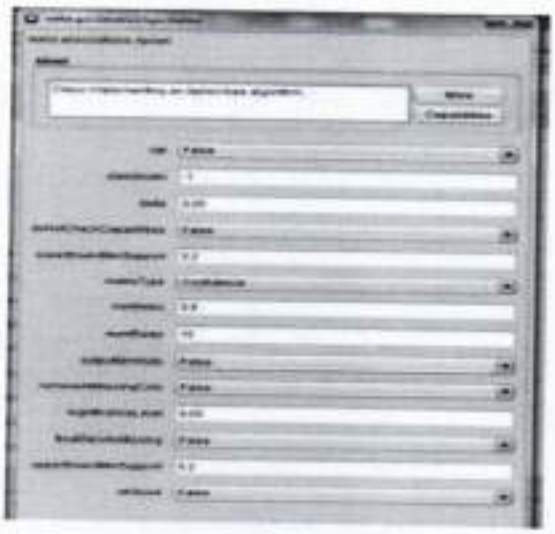

Fig: 9 Minimum Support == 0.2, Confidence metric == 0.90, Significance level = 0.05

Fig: 10 Minimum Support == 0.2, Confidence metric == 0.90, Significance level = 0.05

Fig: 11: Minimum Support == 0.2, Confidence metric == 0.90, Significance level = 0.05

Co-ordinator
IQAC
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn) Dist. Parbhani - 431511 (M.S.)

PRINCIPAL
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn.) Dist.Parbhani

Fig: 12: Minimum Support == 0.2, Confidence metric == 0.90, Significance level = 0.05

It is assumed that support is 0.2 at 5% of significance and it has been decreased up to 10%, it has been observed that there is no significant variation has been found importance among the generated rule.

Table: 3 10 Best generated rules

Sr.No.	Best Rule
1	Specific gravity=1.02 Class=notckd 80 ==> Albumin=0 80 <conf:(1)> lift:(1.64) lev:(0.08) [31] conv:(31.12)
2	Specific gravity=1.02 Class=notckd 80 ==> Sugar=0 80 <conf:(1)> lift:(1.18) lev:(0.03) [12] conv:(12.37)
3	Specific gravity=1.02 Class=notckd 80 ==> Red Blood cells=normal 80 <conf:(1)>lift:(1.14) lev:(0.02) [9] conv:(9.58)
4	Specific gravity=1.02 Class=notckd 80 ==> Pus Cell=normal 80 <conf:(1)> lift:(1.24) lev:(0.04) [15] conv:(15.36)
5	Specific gravity=1.02 Class=notckd 80 ==> Pus Cell Clumps=notpresent 80 <conf:(1)> lift:(1.12) lev:(0.02) [8] conv:(8.58)
6	Specific gravity=1.02 Class=notckd 80 ==> Bacteria=notpresent 80 <conf:(1)> lift:(1.06) lev:(0.01) [4] conv:(4.59)
7	Specific gravity=1.02 Class=notckd 80 ==> Hypertension=no 80 <conf:(1)> lift:(1.58) lev:(0.07) [29] conv:(29.53)
8	Specific gravity=1.02 Class=notckd 80 ==> Diabetes Mellitus=no 80 <conf:(1)> lift:(1.52) lev:(0.07) [27] conv:(27.53)
9	Specific gravity=1.02 Class=notckd 80 ==> Coronary Artery Disease=no 80 <conf:(1)> lift:(1.1) lev:(0.02) [6] conv:(6.98)
10	Specific gravity=1.02 Class=notckd 80 ==> Appetite=good 80 <conf:(1)> lift:(1.26) lev:(0.04) [16] conv:(16.56)

V. CONCLUSION

This paper shows one of the small importances of Weka to utilization and analysis for census Classification issues and knowledge evolution. The Chronic kidney disease data has predicted and diagnosed further damage of kidney based on prediction of Data mining algorithms: In our research work, some of the other factors were considered Red Blood Cell count, Hypertension, Diabetes Mellitus, Coronary Artery Disease, Appetite, Pedal Edema, Anemia, near about 25 factors. The rules create x belongs to y attributes; the primary output of Apriori is to set the best rules by using its value and over caste, and its output shows the rules in the form of the model. Based on rules, describe every rule — the class implementing an Apriori type algorithm. Based on the generated rules, we observed that, if Specific gravity=1.02 Class=not ckd 80 condition is on left hand side then this implies various factors viz. Albumin=0 80, Sugar=0 80, Red Blood cells=normal 80, Pus Cell Clumps=not present 80, Bacteria=not present 80, Hypertension=no 80, Diabetes

Mellitus=no 80, Coronary Artery Disease=no 80 and Appetite=good 80. In a nutshell, based on the data, it has been observing that as specific gravity and class matter the respective presence of activities in the patients. It is for Nephrologists doctors and Medical Industries. Knowledge extraction from the Chronic Kidney Disease database becomes one of the vital processes of each research worker for their research development issues. It is not only crucial in the Medical field but also plays a vital role in Medical Industries for future planning and further prediction issues.

VII. ACKNOWLEDGEMENT

Authors are grateful to the UCI Machine learning repository for providing all the necessary data and WEKA for providing such a reliable tool to extract and analyze knowledge from the database.

References

[1] Arasu, S. Dilli, and R. Thirumalaiselvi. "A novel imputation method for effective prediction of coronary Kidney disease." In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), pp. 127-136. IEEE, 2017.

[2] Ariff, M. H., I. Ismarani, and N. Shamsuddin. "RFID based systematic livestock health management system." In 2014 IEEE Conference on Systems, Process and Control (ICSPC 2014), pp. 111-116. IEEE, 2014.

[3] Bharara, Sanyam, A. Sai Sabitha, and Abhay Bansal. "A review on knowledge extraction for business operations using data mining." In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 512-518. IEEE, 2017.

[4] Chuan, Zou, Tang Ying, Bai Li, Zeng Yuqun, and Lu Fuhua. "Application of clustering analysis to explore syndrome evolution law of peritoneal dialysis patients." In 2013 IEEE International Conference on Bioinformatics and Biomedicine, pp. 23-26. IEEE, 2013.

[5] Duc Thanh AnhLuong, Dept of the computer. Sci&Engi, uni, at Buffalo Buffalo, NY, USAA K- Means Approach to Clustering

a. disease Progressions IEEE Keywords, Sep 2017

[6] Formaggio, Antonio Roberto, Matheus Alves Vieira, and Camilo Daleles Rennó. "Object Based Image Analysis (OBIA) and Data Mining (DM) in Landsat time series for mapping soybean in intensive agricultural regions." In 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 2257-2260. IEEE, 2012.

[7] Gandhi, Niketa, and Leisa J. Armstrong. "Rice crop yield forecasting of tropical wet and dry climatic zone of India using data mining techniques." In 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 357-363. IEEE, 2016.

[8] Güllüoğlu, Sabri Serkan. "Segmenting customers with data mining techniques." In 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), pp. 154-159. IEEE, 2015.

[9] Jinyin, Chen, et al. "A novel cluster center fast determination clustering algorithm." Applied Soft Computing 57 (2017): 539-555.

[10] Khanna, Umesh. "The economics of dialysis in India." Indian journal of nephrology 19, no. 1 (2009): 1.

[11] Kunwar, Veenita, et al. "Chronic Kidney Disease analysis using data mining classification techniques." 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence). IEEE, 2016.

[12] Narander Kumar, SabitaKLhatri, Department of Computer, 3rd IEEE International Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017) " Implemeating WEKA for medical data classification and early disease prediction 978-1-50, 2017

[13] Saggar, Manish, Ashish K. Agrawal, and Abhimanyu Lad. "Optimization of association rule mining using improved genetic algorithms." 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583). Vol. 4. IEEE, 2004.

[14] Sujatha, R., and P. Isakki. "A study on crop yield forecasting using classification techniques." In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), pp. 1-4. IEEE, 2016.

[15] Ubolthum, Issariya, Nakornthip Prompoon, and Wirichada Pan-Ngum. "AKIHelper: Acute kidney injury diagnostic tool using KDIGO guideline approach." In 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp. 1-6. IEEE, 2016.

[16] Varma, P. P. "Prevalence of chronic kidney disease in India-Where are we heading?." Indian journal of nephrology 25, no. 3 (2015): 133.

[17] Khanna, Umesh. "The economics of dialysis in India." Indian journal of nephrology 19, no. 1 (2009): 1.

[18] Varma, P. P. "Prevalence of chronic kidney disease in India-Where are we heading?." Indian journal of nephrology 25, no. 3 (2015): 133.

Co-ordinator
IQAC
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn) Dist. Parbhani - 431511 (M.S.)

PRINCIPAL
Shri Guru Buddhiswami Mahavidyalaya
Purna (Jn.) Dist.Parbhani