International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952331

Data Mining Techniques to Predict Chronic Kidney Disease

Golam Murshid*, Thakor Parvez, Nagani Fezal, Lakhani Azaz, Mohammad Asif

B.E. Computer Jamia Institute of Engineering and Management Studies, Akkalkuwa, Maharashtra India

ABSTRACT

Chronic Kidney Disease incorporates the state where the kidneys fail to function and reduce the potential to keep a person suffering from the disease healthy. When the condition of the kidneys gets worse, the wastes in the blood are formed in high level. Data mining has been a present pattern for accomplishing analytic outcomes. Colossal measure of un-mined data is gathered by the human services industry so as to find concealed data for powerful analysis and basic leadership. Data mining is the way towards extricating concealed data from gigantic datasets. The goal of our paper is to anticipate CKD utilizing the classification strategy Naïve Bayes. The phases of CKD are anticipated in the light of Glomerular Filtration Rate (GFR).

Chronic Kidney Disease (CKD) is one of the most widespread illnesses in the United States. Recent statistics show that twenty-six million adults in the United States have CKD and million others are at increased risk. Clinical diagnosis of CKD is based on blood and urine tests as well as removing a sample of kidney tissue for testing. Early diagnosis and detection of kidney disease is important to help stop the progression to kidney failure. Data mining and analytics techniques can be used for predicting CKD by utilizing historical patient's data and diagnosis records. In this research, predictive analytics techniques such as Decision Trees, Logistic Regression, Naive Bayes, and Artificial Neural Networks are used for predicting CKD. Pre-processing of the data is performed to impute any missing data and identify the variables that should be considered in the prediction models. The different predictive analytics models are assessed and compared based on accuracy of prediction. The study provides a decision support tool that can help in the diagnosis of CKD.

Keywords : Glomerular Filtration Rate, Chronic Kidney Disease, KNN, SVM, ESRD, USRDS

I. INTRODUCTION

The kidneys' functions are to filter the blood. All the blood in our bodies passes through the kidneys several times a day. The kidneys remove wastes, control the body's fluid balance, and regulate the balance of electrolytes. As the kidneys filter blood, they create urine, which collects in the kidneys' pelvis - funnel-shaped structures that drain down tubes called ureters to the bladder. Each kidney contains around a million units called nephrons, each of which is a microscopic filter for blood. Chronic Kidney Disease (CKD) is the gradual loss of kidney function over time. CKD, also called chronic kidney failure or chronic or renal disease, can be caused by several factors including high blood pressure, diabetes, and other disorders. According to the National Kidney Foundation (www.kidney.org),

CSEIT1952331 | Received : 10 April 2019 | Accepted : 28 April 2019 | March-April -2019 [5 (2) : 1220-1226]

there are twenty-six million adults in the United States who have CKD and million others are at increased risk. Kidneys filter the excess fluids and wastes from the blood and as CKD progresses, these wastes and fluids can build in the body and can cause heart and blood vessel disease. Patients who have CKD suffer from symptoms such as lack of energy, fatigue, drowsiness, pain, and purities [1]. The factors that increase the risk of Kidney disease include Diabetes, Hypertension, Smoking, Obesity, Heart Disease, Family History of Kidney Disease, Alcohol Intake, Drug Abuse/Drug Overdose, Age, Race/Ethnicity, and Male Sex [2]. CKD has five different stages of development. Each stage increases in severity as one progress from Stage 1 to Stage 5. In Stage 1, a person can develop below normal kidney functions and even experience a slight loss in kidney function. During Stage 2, a person can experience slight to moderate loss in kidney function. Stage 3 further intensifies, with a person experiencing moderate to severe loss in kidney function. In Stage 4, a person will experience a severe loss in kidney function. In Stage 5, a person will experience complete kidney failure. According to [3], there are no predictive instruments that are commonly accepted for CKD. This lack of a common predictive instrument is becoming a bigger issue as the amount of CKD patients continues to grow. It is also stated that the health burden due to CKD is likely to continue to rise with the aging population and world-wide increase in Type 2 Diabetes [4]. CKD is a disease that affects everyone differently and is progressive in some, but not all patients [5]. Even though different approaches for preventing, reducing, halting, and reversing CKD have been described in the medical writings, all related factors have not been identified comprehensively [6]. Data mining techniques are used to investigate renal disease and to analyse the differences among various administrative areas. Data mining methods used in the literature include Adaptive Neuro-Fuzzy Inference, Support Vector Machines, Artificial Neural Networks, etc. In this study, we will develop predictive analytics models to study and analyse chronic kidney disease. The data set we use has some missing data and we will adopt methods to impute the missing data.

Data mining may have many approaches such as Naive Bayes, J48 , KNN , SVM , Supervised or Unsupervised learning and many more approaches. Data mining here, comes up with a number of techniques that gives knowledge to the healthcare industry which when applied to the processed data.

PROBLEM DEFINITION

The scope of this problem is to classify our dataset using different machine-learning algorithm, which includes training and testing the model. We will try to explore the correlation between the dataset attributes to find out there dependency on each other in the development of chronic kidney disease.

In India an automated diagnosis system would reduce the lengthy process in health care. With an improved symptom analyzing algorithm, the system can suggest diagnostic test to the users hence reducing time and cost in big hospitals.

II. LITERATURE SURVEY

McClellan WM. has proposed Epidemiology and risk factors for chronic kidney disease" End-stage renal disease (ESRD) occurs when kidney function is insufficient to sustain life and haemodialysis, peritoneal dialysis, or kidney transplantation is substituted for native kidney function. There are multiple causes of kidney injury that lead to the final common pathway of ESRD, and this syndrome is characterized by hypertension, anomie, renal bone disease, nutritional impairment, neuropathy, impaired quality of life, and reduced life expectancy. The description and study of the epidemiology of ESRD in the United States population has been greatly enriched by the United States Renal Data System (USRDS), a surveillance system that collects information about the occurrence and outcomes of care on all incident patients receiving treatment for ESRD in the United States. Chronic kidney disease (CKD) is defined by the presence of sustained abnormalities of renal function and results from different causes of renal injury. CKD can lead to progressive loss of renal function and may terminate in ESRD after a variable period of time following the initiating injury [1].

Cristóbal Romero Sebastia´n Ventura, Enrique Garcı´a proposed on "Data mining in course management systems: Moodle case study and tutorial" Educational data mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational context. This work is a survey of the specific application of data mining in learning management systems and a case study tutorial with the Moodle system. Our objective is to introduce it both theoretically and practically to all users interested in this new research area, and in particular to online instructors and e-learning administrators. We describe the full process for mining e-learning data step by step as well as how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering and association rule mining of Moodle data. We have used free data mining tools so that any user can immediately begin to apply data mining without having to purchase a commercial tool or program a specific personalized tool [2].

Hippisley-Cox, J., and Coupland, C, presented "Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales" Chronic Kidney Disease is a major cause of morbidity and interventions now exist which can reduce risk. We sought to develop and validate two new risk algorithms for estimating (a) the individual 5 year risk of moderate-severe CKD and (b) the individual 5 year risk of developing End Stage Kidney Failure in a primary care population. Our final model for moderate-severe CKD included: age, ethnicity, deprivation, smoking, BMI, systolic blood pressure, diabetes, rheumatoid arthritis, cardiovascular disease, treated hypertension, congestive cardiac failure; peripheral vascular disease, NSAID use and family history of kidney disease. In addition, it included SLE and kidney stones in women. The final model for End Stage Kidney Failure was similar except it did not include NSAID use [3].

Navdeep Tangri, Lesley A. Stevens, , John Griffith, PhD, Hocine Tighiouart, MS Ognjenka Djurdjev, David Naimark, Adeera Levin, Andrew S. Levey, have developed a system to "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure" have developed a system to An Estimated 23 Million people in the United States (11.5% of the adult population) have chronic kidney disease (CKD) and are at increased risk for cardiovascular events and progression to kidney failure. Similar estimates of burden of disease have been reported around the world.6 Although there are proven therapies to improve outcomes in patients with progressive kidney disease, these therapies may also cause harm and add cost. Clinical decision making for CKD is challenging due to the heterogeneity of kidney diseases, variability in rates of disease progression, and the competing risk of cardiovascular mortality. Accurate prediction of risk could facilitate individualized decision making, enabling early and appropriate patient care [4].

S. H. Liao, P. H. Chu, and P. Y. Hsiao, have analyzed "Data mining techniques and applications - A decade review from 2000 to 2011," In order to determine how data mining techniques (DMT) and their applications have developed, during he past decade, this paper reviews data mining techniques and their applications and development, through a survey of literature and the classification of articles, from 2000 to 2011. Keyword indices and article abstracts were used to identify 216 articles concerning DMT applications, from 159 academic journals (retrieved from five online databases), this paper surveys and classifies DMT, with respect to the following three areas: knowledge types, analysis types, and architecture types, together with their applications in different research and practical domains. A discussion deals with the direction of any future developments in DMT methodologies and applications: DMT is finding increasing applications in expertise orientation and the development of applications for DMT is a problem-oriented domain. It is suggested that different social science methodologies, such as psychology, cognitive science and human behavior might implement DMT, as an alternative to the methodologies already on offer. The ability to continually change and acquire new understanding is a driving force for the application of DMT and this will allow many new future applications [5].

III. IMPLEMENTATION

Classification – it maps data into predefined groups or classes. In classification the classes are indomitable before examining the data thus it is often mentioned as supervised learning . Classification is the process which classifies the collection of objects, data or ideas into groups, the members of which have one or more characteristic in common. In this research work Naïve Bayes, SVM, ANN and proposed algorithm namely ANFIS are used to classify different stages of Chronic Kidney Failure disease from the dataset [23].

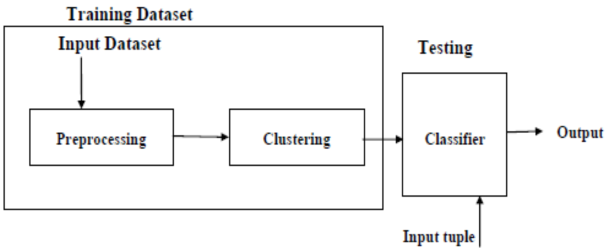
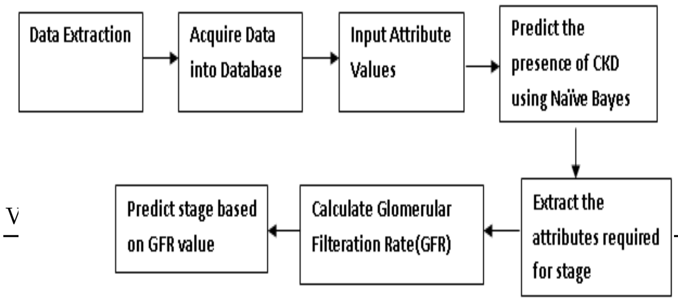
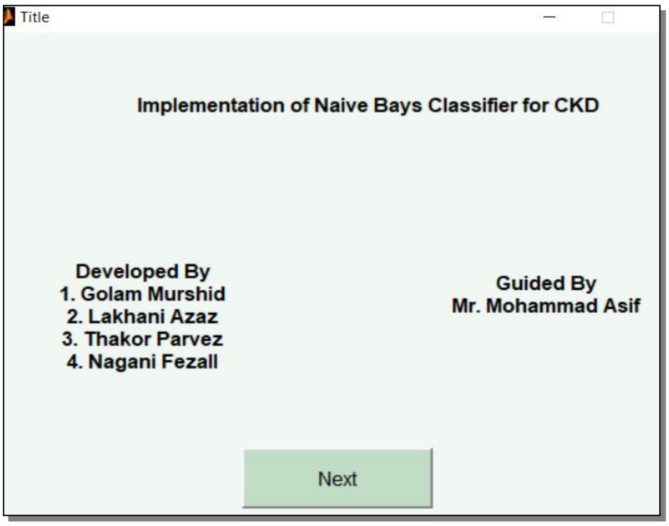


Figure 3.1: Disease Detection

System Architecture


Data Extraction
Acquire Data into Database
Input Attribute Values
Predict the presence of CKD using Naïve Bayes
Extract the attributes required for stage
Calculate Glomerular Filteration Rate(GFR)
Predict stage based on GFR value



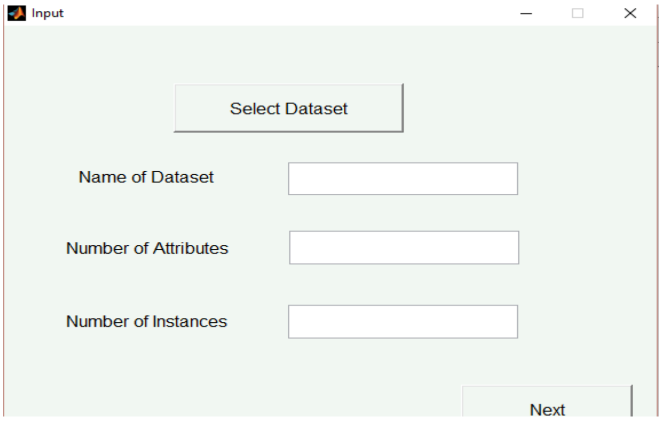
Implementation of Naive Bays Classifier for CKD

Developed By
1. Golam Murshid
2. Lakhani Azaz
3. Thakor Parvez
4. Nagani Fezall

Guided By
Mr. Mohammad Asif

Figure.3.2 : Architecture Diagram of System

Figure 3.3 : First page after run

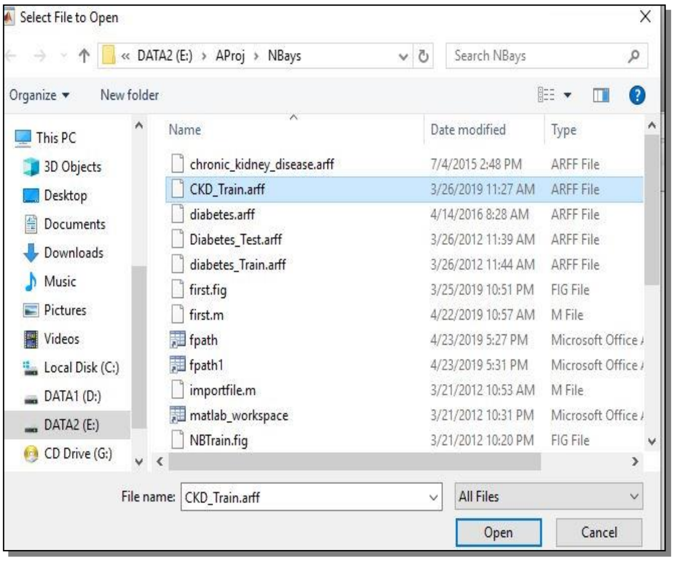


Name of Dataset

Number of Attributes

Number of Instances

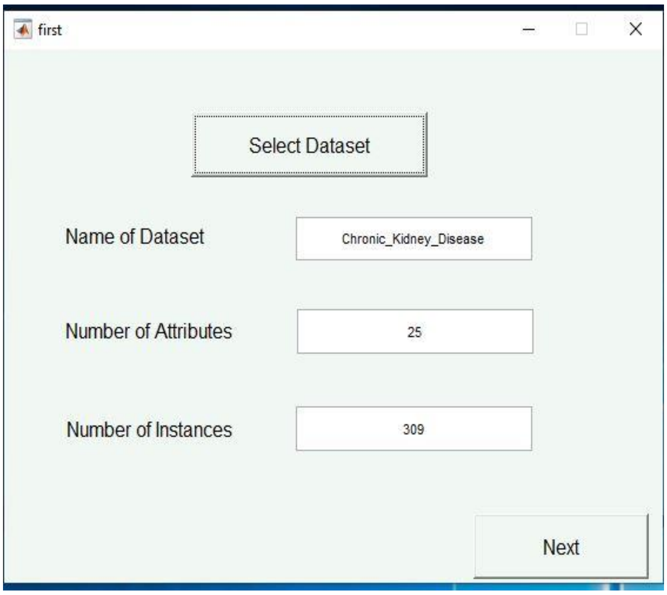
Figure 3.11 : Before selecting the dataset



Name	Date modified	Type
chronic_kidney_disease.arff	7/4/2015 2:48 PM	ARFF File
CKD_Train.arff	3/26/2019 11:27 AM	ARFF File
diabetes.arff	4/14/2016 8:28 AM	ARFF File
Diabetes_Test.arff	3/26/2012 11:39 AM	ARFF File
diabetes_Train.arff	3/26/2012 11:44 AM	ARFF File
first.fig	3/25/2019 10:51 PM	FIG File
first.m	4/22/2019 10:57 AM	M File
fpath	4/23/2019 5:27 PM	Microsoft Office
fpath1	4/23/2019 5:31 PM	Microsoft Office
importfile.m	3/21/2012 10:53 AM	M File
matlab_workspace	3/21/2012 10:31 PM	Microsoft Office
NBTrain.fig	3/21/2012 10:20 PM	FIG File

File name: CKD_Train.arff

Figure 3.12 : Selecting file for training dataset

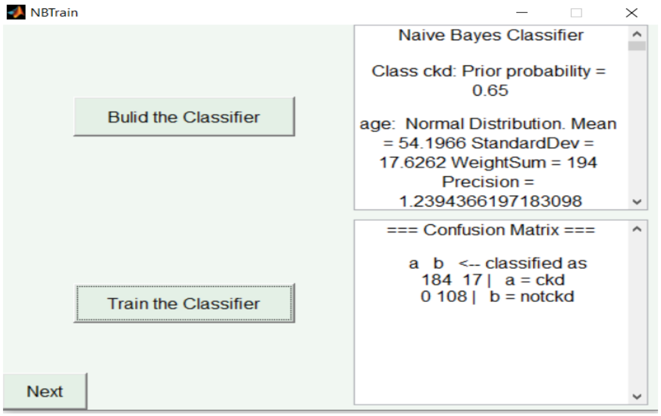


Name of Dataset: Chronic_Kidney_Disease

Number of Attributes: 25

Number of Instances: 309

Figure 3.13 : After selecting the dataset



```
Naive Bayes Classifier

Class ckd: Prior probability =
0.65

age:  Normal Distribution. Mean
= 54.1966 StandardDev =
17.6262 WeightSum = 194
Precision =
1.2394366197183098
```

```
=== Confusion Matrix ===

  a   b   <-- classified as
 184  17 |   a = ckd
  0 108 |   b = notckd
```

Figure 3.14 : Build classifier & train the classifier output

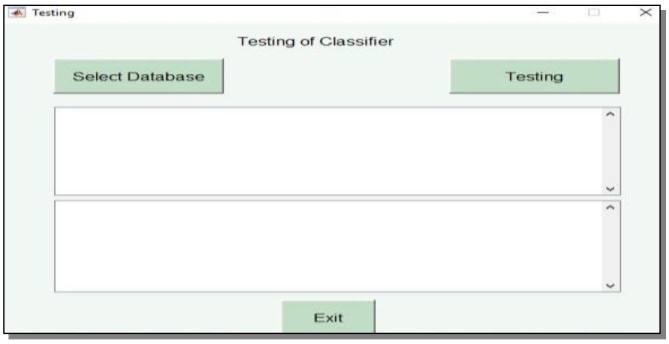
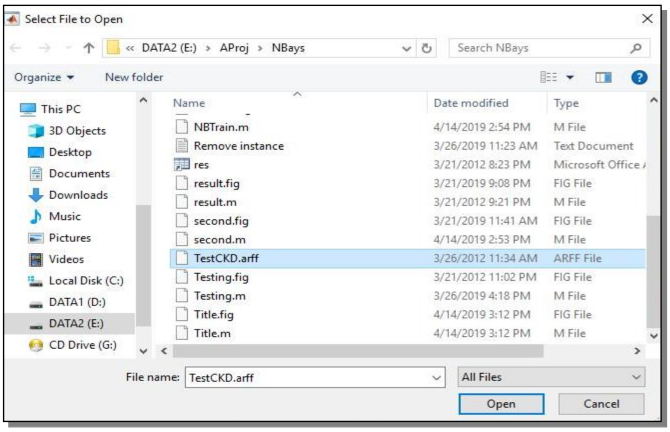


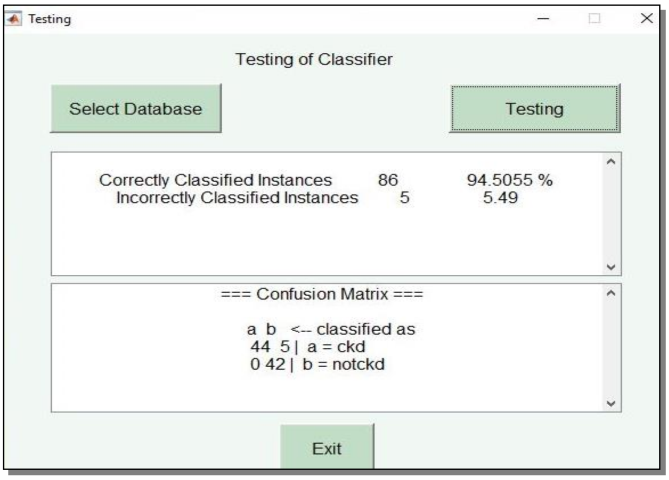
Figure 3.15 : Select Dataset for testing



Name	Date modified	Type
NBTrain.m	4/14/2019 2:54 PM	M File
Remove instance	3/26/2019 11:23 AM	Text Document
res	3/21/2012 8:23 PM	Microsoft Office /
result.fig	3/21/2019 9:08 PM	FIG File
result.m	3/21/2012 9:21 PM	M File
second.fig	3/21/2019 11:41 AM	FIG File
second.m	4/14/2019 2:53 PM	M File
TestCKD.arff	3/26/2012 11:34 AM	ARFF File
Testing.fig	3/21/2012 11:02 PM	FIG File
Testing.m	3/26/2019 4:18 PM	M File
Title.fig	4/14/2019 3:12 PM	FIG File
Title.m	4/14/2019 3:12 PM	M File

File name: TestCKD.arff

Figure 3.16 : Selecting the Test Dataset



```
Correctly Classified Instances      86        94.5055 %
  Incorrectly Classified Instances    5          5.49
```

```
=== Confusion Matrix ===

 a  b   <-- classified as
 44  5 |  a = ckd
  0 42 |  b = notckd
```

Figure 3.17 : Final output with best efficiency

Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease.

the most basic terms, which are

True Positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True Negatives (TN): We predicted no, and they do not have the disease.

IV. CONCLUSION

Data mining has been used in various fields and the importance of it is increasing day by day. In this paper survey of various data mining algorithms are made for the kidney disease. Any one of these techniques will be implemented in future[1].

The chronic kidney disease can be very well predicted using many classifiers in Data Mining. One can also predict the level of chronic kidney disease using classifiers. As per the observation of different experiments there are some classifiers which gave highest accuracy are Naïve Bayes and KNN[7].

V. REFERENCES

[1]. William M. McClellan WM. Epidemiology and risk factors for chronic kidney disease. The Medical clinics of North America. 2005;89(3):419–445. doi: 10.1016/j.mcna.2004.11.006.

[2]. Cristóbal Romero, Data mining in course management systems: Moodle case study and tutorial"http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf

[3]. Hippisley-Cox, J., and Coupland, C., 2010, "Predicting the Risk of Chronic Kidney Disease in Men and Women in England and Wales: Prospective Derivation and External Validation of the QKidney® Scores," Hippisley-Cox and Coupland BMC Family Practice, 11 -49.

[4]. Navdeep Tangri, Lesley A. Stevens, John Griffith, PhD, Hocine Tighiouart, MS Ognjenka Djurdjev, David Naimark, Adeera Levin, Andrew S. Levey, "A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure" JAMA The Journal of the American Medical Association · April 2011.

[5]. S. H. Liao, P. H. Chu, and P. Y. Hsiao, "Data mining techniques and applications - A decade review from 2000 to 2011,"

[6]. Koushal Kumar, K., and Abhishek, 2012, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal of Information Technology and Computer Science, 7, 20-25.

[7]. GeorgeDimitoglou, JamesA. Adams, andCarol M. Jim, Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability.

[8]. Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orrù, Carlo Carcassi and Giorgio La Nasa, Comparison Between an Artificial Neural Network and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome, Chapter 5, an open access article distributed under the terms of the Creative Commons Attribution License.

[9]. Ziyad, A., 2013, "Prediction of Renal End Points in Chronic Kidney Disease," Kidney International, 83(2), 189-191.

[10]. Kobayashi, T., Yoshida, T. 2014, "A Metabolomics-Based Approach for Predicting Stages of Chronic Kidney Disease," Biochemical and Biophysical Research Communications, 445, 412–416.

[11]. Lakshmi, K.R., Nagesh, Y., and VeeraKrishna, M., 2014, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering and Technology, 7(1), 242-254.

[12]. Bala, S., and Kumar, K., 2014, "A Literature Review on Kidney Disease Prediction Using Data Mining Classification Technique," International Journal of Computer Science & Mobile Computing, 3(7), 960-967.

[13]. Vijayarani, S., and Dhayanand, S., 2015, "Data Mining Classification Algorithms for Kidney

Disease Prediction," International Journal on Cybernetics and Information, 4(4), 13-25.

[14]. Ronald N. Kostoff, Uptal Patel., 2015, "Literature-Related Discovery and Innovation: Chronic Kidney Disease," Technological Forecasting and Social Change, 91, 341-351.

[15]. Ravleen Singh Dr. Tariq Hussain Sheikh, "An Overview of Data Mining Applications in Healthcare" International Journal of Advance Research in Computer Science and Management Studies ISSN: 2321-7782.

[16]. 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, pg.135 – 141.

[17]. S.Dilli Arasu, Dr. R.Thirumalaiselvi " Review of Chronic Kidney Disease based on Data Mining Techniques" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505.

[18]. Shahram Tahmasebian1, Marjan Ghazisaeedi1*, Mostafa Langarizadeh2, Mehrshad Mokhtaran1, Mitra Mahdavi-Mazdeh3, Parisa Javadian4 Applying data mining techniques to determine important parameters in chronic kidney disease and the relations of these parameters to each other J Renal Inj Prev. 2017; 6(2): 83-87.

[19]. Sahana B J, Dr Minavathi "Kidney Disease Prediction Using Data Mining Classification Techniquesand ANN" International Journal of Innovative Research in Computer and Communication Engineering ISSN(Online): 2320-9801 Vol. 5, Issue 4, April 2017.

[20]. M. Mayilvaganan, S. Malathi & R. Deepa Data Mining Techniques For The Analysis Of Kidney Disease-A Survey International Journal Of Engineering Sciences & Research Technology-DOI: 10.5281/zenodo.829799.

[21]. Tabassum S, Mamatha Bai B G, Jharna Majumdar, "Analysis and Prediction of Chronic Kidney Disease using Data Mining Techniques" International Journal of Engineering Research in Computer Science and Engineering ISSN (Online) 2394-2320.

[22]. https://www.worldkidneyday.org/faqs/chronic-kidney-disease/ 04/11/2018.

[23]. R. Agrawal and G. Psaila, "Active data mining," Current, pp. 3–8, 1995.

[24]. https://www.indiacelebrating.com/events/world-kidney-day/ 04/11/2018.

[25]. http://www.who.int/life-course/news/events/world-kidney-day-2017/en/ 04/11/2018.

[26]. B. Kjærulff, Anders L. Madsen, (2005) Probabilistic Networks — an Introduction to Bayesian Networks and Influence Diagrams, 10 May.

[27]. International comaparisons of ESRD. http://www.usrds.org/2008/pdf/V2_12_2008.pdf

[28]. Sunita B. Aher1 and Lobo L.M.R.J.2 "Comparative Study Of Classification Algorithms " International Journal of Information Technology and Knowledge Management July-December 2012, Volume 5, No. 2, pp. 239-243

[29]. https://en.wikipedia.org/wiki/Naive_Bayes_classifier / 23/4/2019

Cite this article as :

Golam Murshid, Thakor Parvez, Nagani Fezal, Lakhani Azaz, Mohammad Asif, "Data Mining Techniques to Predict Chronic Kidney Disease", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1220-1226, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952331
Journal URL : http://ijsrcseit.com/CSEIT1952331