Improving Classifier Accuracy for diagnosing Chronic Kidney Disease Using Support Vector Machines

C.Sathish Kumar, P.Thangaraju

ABSTRACT--- Preventing Chronic Kidney Disease has become one of the most intriguing task to the healthcare society. The major objective of this paper is to deal mainly with different classification algorithms namely NaiveBayes, Multi Layer Perceptron and Support Vector Machine. The work analyzes the Chronic Kidney Disease dataset taken from the machine learning repository of UCI. Pre-processing techniques such as missing value replacement, unsupervised discretization and normalization are applied to the Chronic Kidney Disease dataset to improve accuracy. Accuracy and time are the taken as the experimental outcomes of the classification models. The final conclusion states that Support Vector Machine implements much superior than all the other classification methods.

Key Words: Classification, Preprocessing, NaïveBayes, Multilayer Perceptron, Support Vector Machines, CKD.

I. INTRODUCTION

A common and important non-transmissible disease that occurs in India and around the world is Chronic Kidney Disease. CKD is related with a significant amount of morbidity, mortality and economic burden in India. A Kidney failure or related kidney diseases can occur when our kidneys lose the ability of filtering the waste from the blood. The population of Indian people that suffer from chronic kidney related ailments has doubled in the past decade and the situation has become so bad that about 8% to 10% of the adult people suffers from some of the other forms of kidney diseases. Chronic Kidney Disease (CKD) otherwise termed as Chronic Renal Failure (CRF) is a term that includes all forms of reduced function of kidneys that can include an impaired kidney, or kidneys that are at risk due to mild, moderate and severe chronic kidney failure.

It is estimated that with people between the age group of 65 and 74 worldwide, out of 5 men, one has CKD and out of 4 women, one has CKD.[1] Contagious diseases like influence, malaria, or AIDS are replaced by non-contagious diseases like kidney disease, diabetes or heart disease as the collective sources of premature mortality around the world. About 80 percent of these deaths occur in countries with middle or low-income group people, whereas 25 percent of this burden occurs to people with below 60 years of age.[1]

The current information of Global Burden of Disease project provides disturbing reports of fast increasing affliction of chronic kidney disease around the globe. Information of the Global Burden of Disease project discloses an important figure that nearly 90 percent people have lost their lives to chronic kidney disease between 1990 and 2013; it is currently the thirteenth primary cause of mortality across the globe.[2] Chronic Kidney Disease effects for nearly 10 lakh lost lives per year. From 1990, it is one of the three diseases that has regularly increased the degree of mortality. The other two diseases are diabetes and HIV/AIDS.[3]

Of late, data mining field is one of the most useful and powerful tools to extract and manipulate data and to establish patterns to provide useful data for making decisions. Data mining is applied in various domains like sales and promotion, CRM, production, medical domain, predicting expert systems, manufacturing, AI, Internet domains and mobile applications.

Clinical diagnostic and operative conclusions are taken on the basis of the instinctiveness and expertise of doctors rather than on the basis of understanding of information concealed in the data store. It often gives way to unnecessary preferences, inaccuracies and relatively high medicinal costs that can affect the quality of services delivered to suffering patients. Many tools are available to support data mining operations. One such tool is WEKA (Waikato Environment for Knowledge Analysis). WEKA is the most widely used datamining tool that supports different data mining algorithms for predictive classification.

The main goal of this paper is to predict the presence or non-presence of Chronic Kidney Disease in patients based on the CKD dataset and to improve the predictive classification accuracy.

The organization of this paper is as follows: Section 2 deals with data mining challenges.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

C.Sathish Kumar*, Research Scholar, Bharathidasan University, Tiruchirappalli & Associate Professor, PG & Research Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, Tamilnadu, India.

(Email: satgreen.in@gmail.com)

P.Thangaraju, Associate Professor, PG & Research Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, Tamilnadu, India.

(Email: trthangaraju@gmail.com)

Section 3 elaborates on the preprocessing techniques and classifier algorithms used for this research work. Section 4 discusses about literature review. Section 5 details about the data set description. Section 6 explains about the proposed methodology and section 7 concludes with the experimental results and performance evaluations.

II. CHALLENGES IN MEDICAL DATA MINING

Before the actual mining can happen, a number of problems are to be addressed in the medical domain. Below is a brief overview of the important issues faced in the field of medical data mining process:

High volume of data:

The amount of medical data is becoming huge. Because of this, a sample extraction of data has become the need of the hour. One other way of extracting data is to create an attribute subset from the data store. These two techniques use the domain knowledge to remove unnecessary attributes. This will reduce the overall size of the data store.

Update:

Latest laboratory procedures and diagnostic methods are regularly updating the information of medical data. Any new technique introduced in data mining must support these modifications.

Inconsistent data representation:

A common problem found in datasets is the erroneous entry of data. This will lead to data inconsistencies. If there is more than one way of expressing something, inconsistencies will occur. No consistent format is available to store medical data.

Missing (or) Incomplete data:

Not all the required data is available for identifying patterns in medical datasets. Omitting data values, irrelevant data, inapplicability of data in a particular medical context are some of the reasons for missing data. Learning techniques like logistic regression may need a full set of data elements for process. But even if logistic regression accepts missing data, it is not wise to overlook these data because they may contain non-dependent information. One method that is adopted to manage these incomplete data is to substitute missing values with the most known values.

Noise:

The existence of noise in datasets is an important factor.[4] Noise has become an obvious problem which may affect the different processes in data extraction. This can frequently result in errors. Techniques used in mining processes must cope with these noises.

III. DATA MINING TASKS IN HEALTHCARE

Data mining deals with the procedure of analysing huge databases to find useful patterns. It is used to extract unidentified patterns and information that are hard to be identified from vast dataset. These days, medical institutions generate and collect huge volumes of data. With the help of data mining methods, it is easy to mine the data and generate necessary patterns. This can increase work efficiency and boost the decision-making process. Doctors use information technologies to improve their decision-making capabilities. Natural human tendencies like bias towards a particular subject and fatigue-related errors can be controlled by data mining processes and thus leading to better decision-making skills.[5]

Data mining finds out models, relationships, and patterns that can help to predict and decide upon the tasks for diagnosing and treating diseases. These models are known as predictive models and are clubbed with diagnostic and medical systems for improving upon decisions, reducing bias towards a subject and minimizing the time taken for making decisions. Decision-making becomes a weak aspect whenever a huge amount of information is to be processed.[6] The data mining process analyzes the raw data and discovers its meaning. It has now become easy to judge trends and behaviour of patients or diseases. Trends and patterns in a dataset, that are not known previously can now be learnt and converted into meaningful information.[7] Categories of data mining models include a predictive model and a descriptive model. In a predictive model, data values are predicted based on the findings from various datasets. Regression, Time Series Analysis and classification are a part of predictive model. The descriptive model finds out similarities or relationships in data. A descriptive model consists of clustering and association rules.[8]

Medical field uses the following data mining methods:

1. Association:

Association is used to find the possibility of the elements occurring simultaneously in a group. Association rules define the relationships between these simultaneous-occurring elements. Specifically, it deals with dependently linked attributes. Here, we search for a particular set of attributes or events that are extremely interrelated with any other attribute or event. As an example, consider a market place, when consumers buy a particular product, they follow it by buying a related product.

2. Clustering:

Clustering generates labels of class for a data category. The Clustering principle follows the grouping of objects based on maximization of intraclass similarity and minimization of interclass similarity. Clusters are formed in such a way that highly similar objects are grouped under one cluster and dissimilar objects are placed in other clusters. Clusters so formed can be approached as a class of objects, from which rules can be deduced.

3. Regression:

The prediction of a number is done through the regression technique. A regression task commences with a data set in which the target values are identified. In the training process, a regression algorithm evaluates the value of the target as a function of the predictors for each situation in the build data. These associations between predictors and goal are précised in a model, which can then be used on a different data set in which the target standards are known.

The computation of various statistical measures and the difference among the predicted value and the expected value is calculated through the regression model.[9]

4. Classification:

Classification assigns categories to a group of data that can help in accurate predictions and analysis. The data samples are divided into target classes. Major aim of any classification technique is to predict target class for each data point. Classification can analyse disease patterns by associating a risk factor to patients. It is a supervised learning approach where the class categories are known beforehand. There are two methods of classification namely binary and multilevel classifications. Binary classification uses only two possible classes such as "high" or "low" risk patients, while in a multiclass approach more than two targets are defined, for example, "high", "medium" and "low" risk patients. Any data set can be partitioned into a training and a testing data set. Based on a given input, it predicts a certain outcome. Set of attributes are used by a training set algorithm to predict the outcome of the data set. To predict the outcome, the training set algorithm tries to find the relationship among attributes. "Goodness of any algorithm is defined as accuracy.

Preprocessing techniques:

Preprocessing the given dataset lays the groundwork for mining data. Before the discovery of useful information or knowledge, the result giving dataset must be properly prepared. But unfortunately this part is not given its due importance by most researchers due to its perceived difficulty.[10]

Replacing Missing Values:

Sometimes there are attributes in a dataset whose values are either incomplete or missing. One common method of expressing missing data is to input values that cannot be found in the data for example, missing data can be represented as "-1" or "?". Whenever a value for an attribute is empty, we normally assume that the particular case is not so important as compared to the rest of the cases in the dataset. But it may not be true as every value for an attribute may contribute useful information for any given dataset.[11]

Normalization:

Normalization scales the attribute data in such a way that all the values fall in between a very small range that can be from -1.0 to 1.0 or from 0 to 1.0.[10]

It points that the highest value for any attribute is 1 and the lowest value for the same is -1 or 0. When we cannot distribute the data in the dataset, or when the dataset distribution is not Gaussian, normalization is usually applied. It is useful when the data has varying scales.[12]

Discretization:

Discretization is the method of converting a real-valued attribute into an ordinal attribute or bins. Discretization is performed by simple binning.[13] We can either use equal frequency method or equal-width method. Equal-width method splits the range of the numeric attribute into equal number of parts or bins. The bin name is taken as the discretized version of the numeric value when any numeric value falls into a bin. The number of data points between bins may vary. The equal-frequency method tries to make the number of data points into each bins equal. The bin size is adjusted in such a way as to make the number of instances that fall into each bin approximately the same.[14]

Different classification techniques are used for various health domains. For the research work, three classification techniques are taken as bench marking algorithms to be studied for the CKD dataset.

Algorithms such as NaïveBayes, Multi Layer Perceptron and Support Vector Machines are analysed and tested with the given dataset. Brief outline of these algorithms are as follows:

NaiveBayesian classifier:

A well-known statistical and supervised method for classification is NaiveBayesian classifier. This theorem works on strong independent assumption with PB-classifier methodology. When the input dimension is high, this method can be applied successfully. The class of a particular tuple can be predicted through this probability approach. This classifier strongly supports the conditional probability. It is considered as an important algorithm for classification task. The missing value concept and the imbalance factors are appropriately dealt with this approach.[15]

Bayes theorem was proposed by Thomas Bayes (1702-1761) and hence it was aptly named after him.

The Bayesian formula is given as below:

$$Prob(H|E) = \frac{Prob(E|H)Prob(H)}{Prob(E)} \qquad (1)$$ [16]

$$Prob(H|E) = Prob(e_1|H) \times Prob(e_2|H) \times \ldots \times Prob(e_n|H) \times Prob(H) \qquad \ldots (2)$$

where,

For the given predictor (E, attributes), the posterior probability of the class (H, target) is

defined as Prob(H|E)

When class H is true, the prior probability is Prob(H)

The probability of the given predictor class Prob(E|H) is the likelihood.

The prior probability of predictor is Prob(E).[16]

The basic idea of Bayes's rule is that the outcome of a hypothesis or an event (H) can be predicted based on some evidences (E) that can be observed from the Bayes's rule.[17]

Multilayer Perceptron classifier:

The most used supervised model consisting of numerous layers for computation normally interconnected in feed-forward way mode is known as Multi layer Perceptron. A significant feed-forward Artificial Neural Network is organized by Multilayer perceptron. It attempts to imitate the learning process of human beings and model the functionalities using biological neural networks. Each neuron of a particular layer is directly connected to all other neurons of the subsequent layers in MLP. Non-linear models and classifiers can be parameterized by MLP classifier.

Retrieval Number: F9377088619/19©BEIESP
DOI: 10.35940/ijeat.F9377.088619
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

This method enhances the overall results as compared with other classical methods. MLP constitutes a network of neurons and are often termed as perceptrons. In 1958, Rosenblatt introduced the basic concept of a single perceptron. Based on the weights of the input, a single output from more than one real inputs is computed by the perceptron. This output goes is put through any nonlinear activation function.

The above process is mathematically expressed as:

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(W^T X + \mathrm{b}) \qquad \ldots (1)$$

In the above equation, x is called the input vector, w denotes the weight vector, b is called the bias and the activation function is Φ. Multilayer Perceptron is a combination of input, output and a set of hidden layers as depicted below in figure 1. To solve difficult and diverse problems Multilayer Perceptron can be trained in a supervised manner using the error backpropagation algorithm. Error correction learning rule is the base for this algorithm. Thus, it became a generalized view of an adaptive filtering algorithm.

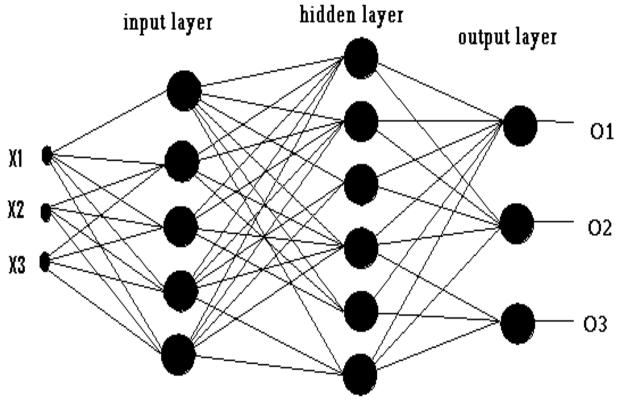


Figure 1: Multilayer Perceptron

Support Vector Machines:

SVM works on statistical learning principle. This technique classifies linear as well as nonlinear data models. [18] It applies structural risk minimum principle that formulates a goal function, and follows it by finding a partitioned hyperplane that satisfies the class requirement. Firstly, an optimal hyperplane is searched to satisfy the request of the classifier. Secondly, besides the maximum optimal hyperplane a margin of separation is created by using an algorithm thus guaranteeing the classifier accuracy. It leads to effective data classification of classes. Using nonlinear mapping to a higher dimension, the data from the two classes is separated by a hyperplane. Using support vectors and margins, the SVM identifies the hyperplane.[19] We maximize the width of the margin (*w*) to define an optimal hyperplane.

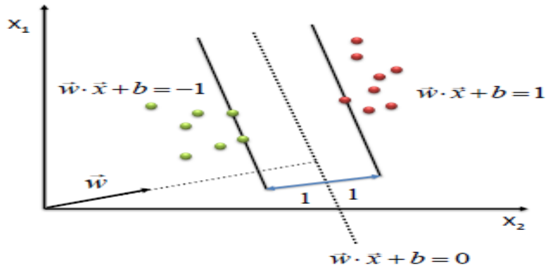

x_1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
x_2
$\vec{w}\cdot\vec{x}+b=0$

Figure 2: A hyperplane that differentiates two classes

We want to solve $max\, \frac{2}{||w||}$ such that,

$(w.x + b) \geq 1, \forall x\; of\; class\; 1$ and $(w.x + b) \leq -1, \forall x\; of\; class\; 2$

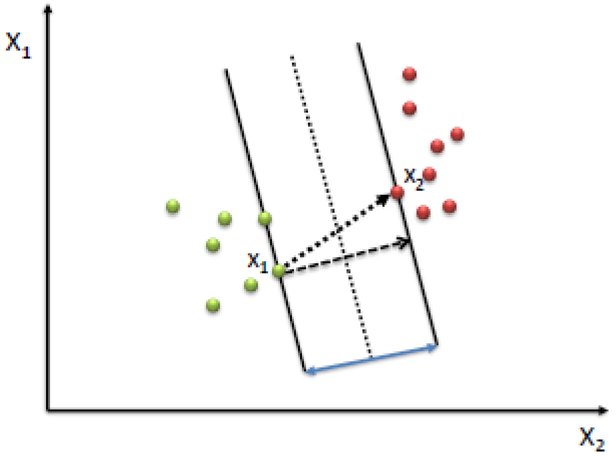


Figure 3: An optimal hyperplane

$$\frac{w}{||w||}.(x_2 - x_1) = width = \frac{2}{||w||} \qquad \ldots (1)$$

$$w.x_2 + b = 1 \qquad \ldots (2)$$

$$w.x_1 + b = -1 \qquad \ldots (3)$$

$$w.x_2 + b - w.x_1 - b = 1 - (-1) \qquad \ldots (4)$$

$$w.x_2 - w.x_1 = 2 \qquad \ldots (5)$$

$$\frac{w}{||w||}(x_2 - x_1) = \frac{2}{||w||} \qquad \ldots (6)$$

We find *w* and *b* by solving the following objective function using Quadratic Programming

$min \frac{1}{2}||w||^2$ such that

$y_i(w.x_i + b) \geq 1, \forall x_i \qquad \ldots (7)$ [20]

When compared to other risk minimization based learning algorithms, SVM provides effective generalization capabilities. Even though the SVM takes more time to train the data, they produce high accuracy rate. They can model complex nonlinear decision boundaries. The concept of overfitting is much low in SVM when compared to other models. Along with classification, SVM models are also used to predict outputs. Some of the areas where SVM can be used include time-series prediction tests, speaker recognition, handwritten digit identification and object identification.

IV. LITERATURE REVIEW

Puneet Kaur et al., in their paper, have used classification techniques for predicting student dataset with 152 student records to analyse academic performance of students and to identify slow learners among them. Student data was collected from academic sources and a model was designed. Five different classification techniques like Multilayer Perceptron, NaïveBayes, SMO, J48, REPTree are applied to classify data. Out of the five classification models, Multi Layer Perceptron performed as the best with an accuracy rate of 75% and hence it proved as the better classification method.[21]

S.Dhamodharan in his paper has used two classifier models such as NaïveBayes and FP-Tree algorithm for predicting the likelihood of patients affected with 3 main Liver diseases namely Hepatitis, Liver Cancer, and Cirrhosis using their unique symptoms. 29 different training datasets were used for experimental results and all were compared and evaluated. NaïveBayes classification produced an accuracy rate of 75.54% whereas FP Tree algorithm produced an accuracy of 72.66%.[22]

Priyadarshini Jambagi et al., proposed a diagnostic model to detect early stages of lung cancer on the basis of the analysis of the sputum color images. An extraction of features followed by a classifier model based on SVM method to classify whether the sputum cell is a cancerous or a normal cell is presented by the paper. As many as 100 color images of sputum cell is used to find the outputs.[23]

Dr. Dilip Roy Chowdhry et al., presented a paper that focuses on identifying specific diseases in a mushroom farm. They used real data from a mushroom farm. Classification models like NaïveBayes, SMO, RIDOR algorithms were applied to predict disease in mushrooms. Both SMO and Naïve Bayes algorithms produced 100% accuracy whereas RIDOR(Ripple-Down Rule Learner) algorithm provided 89.0909% accuracy.[24]

Vineeta Kunwar et al., have used classifier methods such as NaïveBayes, Artificial Neural Networks(ANN) for diagnosing patients with Chronic Kidney Disease. The experimental outputs were tested in Rapidminer tool that showed that NaïveBayes algorithm produces much better results than Artificial Neural Networks(ANN).[25]

Murat Koklu et al., have experimented on the dataset of CKD by taking four different classification algorithms. About 400 samples the dataset was used for the research. Classification models like NaïveBayes, Support Vector Machines, C4.5, Multi Layer Perceptron produced varying accuracy rates for the CKD dataset. Naïve Bayes produced an accuracy of 95.00%, SVM provided an accuracy of 97.75%, C4.5 gave an accuracy rate of 99.00% and Multi Layer Perceptron provided the best rate of accuracy as 99.75%.[26]

Lambodar Jena et al., have used chronic kidney disease dataset for data classification. They used six classifiers to predict the accuracy for the above data set. NaïveBayes classifier, Multi Layer Perceptron, SVM, J48, Conjunctive rule, Decision Table are used for classification. Parameters like Accuracy, Kappa statistics, Mean Absolute Error, Root Mean Squared Error and performance measurement for classification problems such as Receiving Operating Characteristic curve were considered and comparison is done for all the six classifiers. Time to build and test the model are also taken in to account. Multi Layer Perceptron produced the best accuracy rate.[27]

Aiswarya Iyer, et al., have proposed a paper that analyses patterns in diabetes dataset and finds solutions that can predict diabetes using classifier techniques. They applied two classification models namely Decision Tree and NaïveBayes algorithm to classify data. They used two approaches to test and train the model. One model is the 10-fold cross validation. Another model is a 70:30 percentage split of instances where 30% of the instances are provided for testing dataset while the remaining 70% are provided to train the dataset. NaïveBayes performed better with an accuracy rate of 79.5652% with the 70:30 percentage split technique.[28]

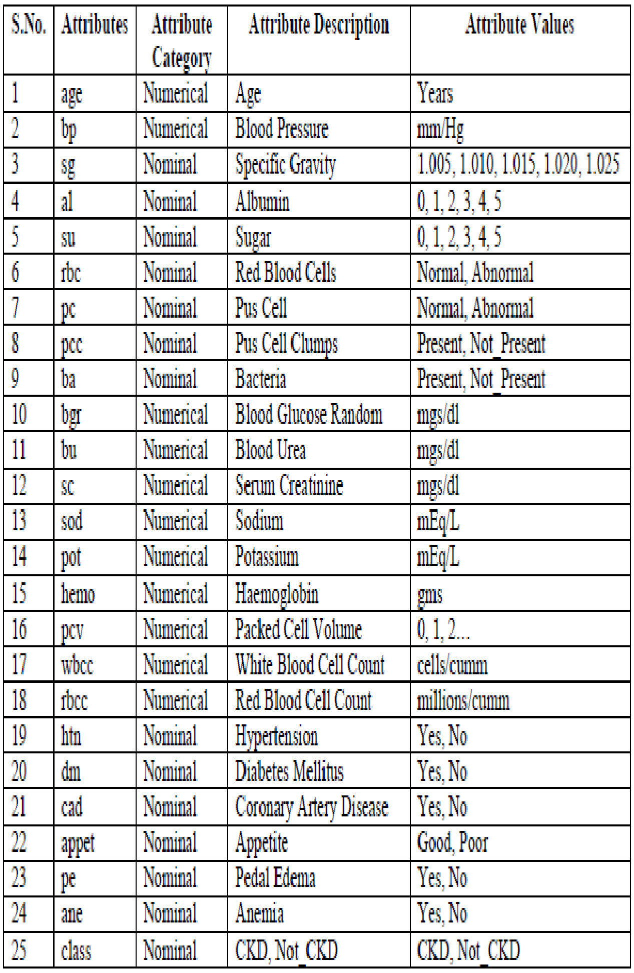
Meriem Amrane et al., have presented two different classifiers: NaïveBayes classifier and the K-nearest neighbour (KNN) algorithms for breast cancer prediction. They proposed a comparison between the two algorithms and evaluated their accuracy using cross validation technique. Results showed that when compared to NaiveBayes classifier with an accuracy rate of 96.19%, the K-nearest neighbour gives a higher accuracy of 97.51% with lower error rate.[29]

Yomna Omar, et al., have created a Lung Cancer Prognosis System (LCPS). Oncologists can use this system to evaluate accurately their patients' health condition. Different lung cancer datasets are accepted by this LCPS. Multiple data mining algorithms are used to find out the associations between the detected symptoms and likely results. The system provides oncologists with different mathematical results that can also include medical future of their patients. Three classifier models J48, NaïveBayes, K-Nearest Neighbour were applied in the classification of lung cancer dataset. J48 gives the best accuracy rate of 93.1034%, whereas Naïve Bayes provides 80.2956% accuracy rate and KNN produces 89.6552% of accuracy.[30]

V. DATASET DESCRIPTION

The Chronic Kidney Disease dataset was taken from UCI machine learning repository.[31] The dataset has 400 instances that are sampled into two different classes with 250 cases of "CKD" and 150 cases of "Not_CKD". There are 24 attributes plus one class attribute which is of binary category. Out of the 24 attributes, there are 11 numeric attributes whereas remaining 14 attributes belong to nominal category.

Chronic Kidney Disease dataset attributes as taken from UCI machine learning repository.[32]



S.No.	Attributes	Attribute Category	Attribute Description	Attribute Values
1	age	Numerical	Age	Years
2	bp	Numerical	Blood Pressure	mm/Hg
3	sg	Nominal	Specific Gravity	1.005, 1.010, 1.015, 1.020, 1.025
4	al	Nominal	Albumin	0, 1, 2, 3, 4, 5
5	su	Nominal	Sugar	0, 1, 2, 3, 4, 5
6	rbc	Nominal	Red Blood Cells	Normal, Abnormal
7	pc	Nominal	Pus Cell	Normal, Abnormal
8	pcc	Nominal	Pus Cell Clumps	Present, Not_Present
9	ba	Nominal	Bacteria	Present, Not_Present
10	bgr	Numerical	Blood Glucose Random	mgs/dl
11	bu	Numerical	Blood Urea	mgs/dl
12	sc	Numerical	Serum Creatinine	mgs/dl
13	sod	Numerical	Sodium	mEq/L
14	pot	Numerical	Potassium	mEq/L
15	hemo	Numerical	Haemoglobin	gms
16	pcv	Numerical	Packed Cell Volume	0, 1, 2...
17	wbcc	Numerical	White Blood Cell Count	cells/cumm
18	rbcc	Numerical	Red Blood Cell Count	millions/cumm
19	htn	Nominal	Hypertension	Yes, No
20	dm	Nominal	Diabetes Mellitus	Yes, No
21	cad	Nominal	Coronary Artery Disease	Yes, No
22	appet	Nominal	Appetite	Good, Poor
23	pe	Nominal	Pedal Edema	Yes, No
24	ane	Nominal	Anemia	Yes, No
25	class	Nominal	CKD, Not_CKD	CKD, Not_CKD

VI. PROPOSED METHODOLOGY

Initially the CKD dataset is taken and tested for accuracy using the three classifiers such as Naïve Bayes, Multilayer Perceptron and Support Vector Machine. Then the same dataset is reduced using three preprocessing techniques such as missing value replacement, normalization and discretization.

In the CKD dataset we find that all the 24 attributes have atleast some percentage of missing values. Especially the attribute rbc (Red Blood Cells) has 38% missing data i.e., about 152 instances of the total 400 instances are missing. Similarly wbcc (White Blood Cell count) has 27% missing data i.e., about 106 instances of the overall 400 instances are missing. Also rbcc (Red Blood Cell count) has about 131 instances of the total 400 instances which are missing. This constitutes to about 33% missing data. These missing data can significantly impact the conclusions that can be drawn from the data.[33] One common method of filling these missing data is by replacing the missing values with the arithmetic mode or mean or in relation to those attributes. The mean substitution method is used to replace all missing values. Here, the mean value of an attribute is used in place of the missing data instance for that same attribute.[34]

After replacing missing values, the entire data is normalized to have the maximum value for every attribute to be 1 and the minimum value for each attribute to be 0.

And then after normalization, discretization is applied. The entire data is discretized into multiple bins. The default bin size is taken as 10. The equal-frequency method is set as false.

After each of the above preprocessing techniques, the dataset goes through the three classifiers namely Naïve Bayes, Multilayer Perceptron and Support Vector Machine. Accuracy and time taken by each classifier is analysed after every process.

Step 1: Load the CKD data set from UCI machine learning repository

Step 2: Perform classification on the CKD data set using the three classifiers namely NB, MLP and SVM

Step 3: Find and analyse the accuracy of the above three classifiers

Step 4: Reload the CKD data set for the second time from UCI repository

Step 5: Apply three preprocessing techniques such as Missing Value Replacement, Normalization and Discretization one by one in that order to the CKD data set

Step 6: After each filter, classify the data set using NB, MLP and SVM classifiers

Step 7: Find and analyse the accuracy of the classifiers after preprocessing techniques

Step 8: Compare the accuracy of the classifiers before and after applying the preprocessing techniques

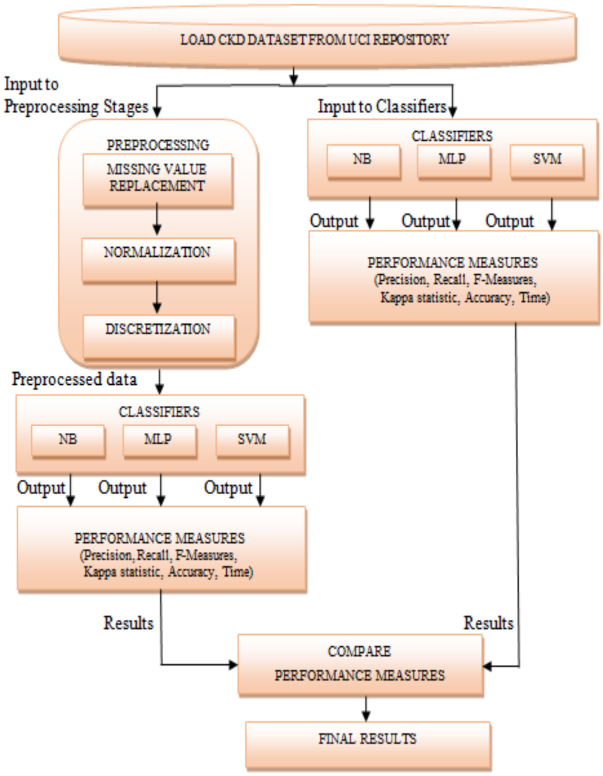

LOAD CKD DATASET FROM UCI REPOSITORY
Input to Preprocessing Stages
Input to Classifiers
PREPROCESSING
MISSING VALUE REPLACEMENT
NORMALIZATION
DISCRETIZATION
CLASSIFIERS
NB
MLP
SVM
Output
Output
Output
PERFORMANCE MEASURES (Precision, Recall, F-Measures, Kappa statistic, Accuracy, Time)
Preprocessed data
CLASSIFIERS
NB
MLP
SVM
Output
Output
Output
PERFORMANCE MEASURES (Precision, Recall, F-Measures, Kappa statistic, Accuracy, Time)
Results
Results
COMPARE PERFORMANCE MEASURES
FINAL RESULTS

Figure 4: Proposed Methodology

VII. EXPERIMENTAL RESULTS

The analysis and interpretation of CKD dataset with preprocessing and classification methods is rather a time consuming process and it requires a thorough understanding of several statistical measures. Statistical measures such as Accuracy, Precision, Sensitivity and Specificity for the NaïveBayes, Multi Layer Perceptron and Support Vector Machine algorithms are analysed. The experimental results suggest that data preprocessing techniques will actually improve the classification accuracy.

Performance Measurements:

Confusion matrix:

It is a tabular representation that gives the efficiency of an algorithm through the computation of performance metrics. Confusion matrix presents the correct and incorrect predictions of any classification model. The actual outcomes are compared with the targeted values of the dataset.

Table 1: General format of confusion matrix

Actual Outcomes	**Predicted Outcomes**	
	TP	FN
	FP	TN

Precision:

It is given as the number of True Positives divided by the total number of True Positives and False Positives. That is, it gives the result of the overall number of positive correct predictions divided by the total number of positive class values predicted. This measure is called the Positive Predictive Value (PPV)

$$Precision = \frac{TP}{TP+FP} \qquad \ldots(1)$$

Recall:

It is calculated by dividing the number of True Positives with the cumulative total of the number of True Positives and the number of False Negatives. That is, it gives the result of the overall number of positive correct predictions divided by the total number of positive class values as given in the test data. This measure is called Sensitivity or the True Positive Rate (TPR)

$$Recall\ (or)\ Sensitivity = \frac{TP}{TP+FN} \qquad \ldots(1)$$

Specificity:

It is computed by dividing TN with the total number of TN and FP. That is, it gives the result of the division between the number of correct negative predictions with the total number of negatives. This measure is also called the True Negative Rate (TNR).[35]

F-Measure:

This measure is called F_1-score. It provides the harmonic mean of recall and precision.[36]

$$F - Measure = \frac{2 \times Recall \times Precision}{Precision+Recall} \qquad \ldots(1)$$

Kappa statistic:

Cohen introduced kappa statistic. It is symbolized by the Greek letter κ[37] It measures pairwise agreement among two different raters. Each rater classifies a set of N items into C mutually exclusive categories. It is very similar to correlation coefficients. The range of kappa statistic is between -1 and +1.[38] For example, if κ = 1, there is a complete agreement between the classifier and the real world value. When no agreement can be found at all among the raters, we say that Kappa statistic κ is <= 0. Kappa statistic equation κ is:

$$\kappa = \frac{P_r(a)-P_r(e)}{1-P_r(e)} \qquad \ldots(3)$$

Here the actual observed agreement among the raters is $P_r(a)$. $P_r(e)$ is given as the expected (or) chance agreement. They are calculated as follows:

$$P_r(a) = \frac{(TP+TN)}{N} \qquad \ldots(1)$$

$$(e) = [(TP + FN) \times (TP + FP) \times (TN + FN)]N^2$$
$\ldots(2)$[39]

Here N gives the overall number of instances used. Classifiers with a better performance should have the higher κ.[37]

Accuracy:

Accuracy can be computed as the overall number of all correct predictions divided with the overall number of instances in the given dataset.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \ldots(1)$$

Table 2: Performance Statistics before and after preprocessing

Classifier Type	Precision	Recall	F-Measure	Kappa statistic
NB before preprocessing	0.956	0.95	0.951	0.8961
NB after preprocessing	**0.983**	**0.983**	**0.983**	**0.9630**
MLP before preprocessing	0.998	0.998	0.998	0.9947
MLP after preprocessing	**0.98**	**0.98**	**0.98**	**0.9574**
SVM before preprocessing	0.979	0.978	0.978	0.9526
SVM after preprocessing	**0.988**	**0.988**	**0.988**	**0.9734**

Table 3: Confusion matrix for NB classifier before preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	230	20	a = CKD
	0	150	b = NOT_CKD

Table 4: Confusion matrix for NB classifier after preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	243	7	a = CKD
	0	150	b = NOT_CKD

Table 5: Confusion matrix for MLP classifier before preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	249	1	a = CKD
	0	150	b = NOT_CKD

Table 6: Confusion matrix for MLP classifier after preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	243	7	a = CKD
	0	150	b = NOT_CKD

Table 7: Confusion matrix for SVM classifier before preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	241	9	a = CKD
	0	150	b = NOT_CKD

Table 8: Confusion matrix for SVM classifier after preprocessing

Actual Outcomes	**Predicted Outcomes**		
	a	**b**	**Classified as**
	246	4	a = CKD
	1	149	b = NOT_CKD

Table 9: Classifier Accuracy and Time before and after applying preprocessing techniques

Classifier	**Accuracy**	**Time**
NB before preprocessing	95.00%	0.03 secs
NB after replacing missing values	94.50%	0.01 sec
NB after replacing missing values and Normalization	95.00%	0.02 secs
NB after replacing missing values, Normalization and Discretization	**98.25%**	**0.01 sec**
MLP before preprocessing	99.75%	8.89 secs
MLP after replacing missing values	97.75%	8.78 secs
MLP after replacing missing values and Normalization	97.50%	8.70 secs
MLP after replacing missing values, Normalization and Discretization	**98.00%**	**93.06 secs**
SVM before preprocessing	97.75%	0.13 secs
SVM after replacing missing values	97.75%	0.02 secs
SVM after replacing missing values and Normalization	97.75%	0.01 sec
SVM after replacing missing values, Normalization and Discretization	**98.75%**	**0.13 secs**

Figure 5: Graph representation depicting Precision, Recall, F-Measure and Kappa statistic before preprocessing

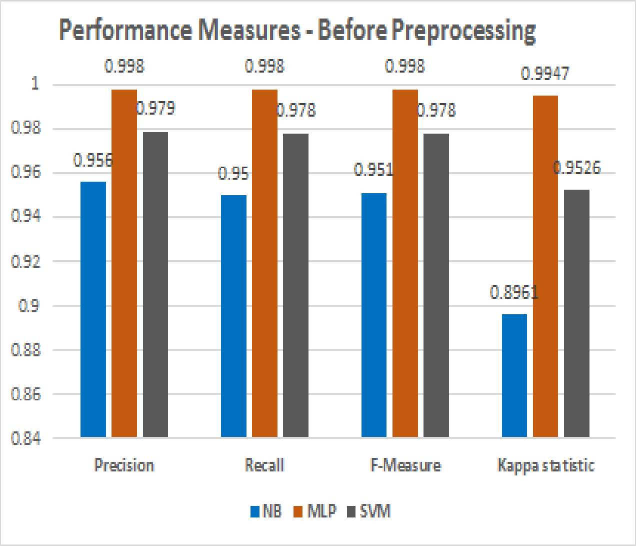


Figure 6: Graph representation depicting Precision, Recall, F-Measure and Kappa statistic after preprocessing

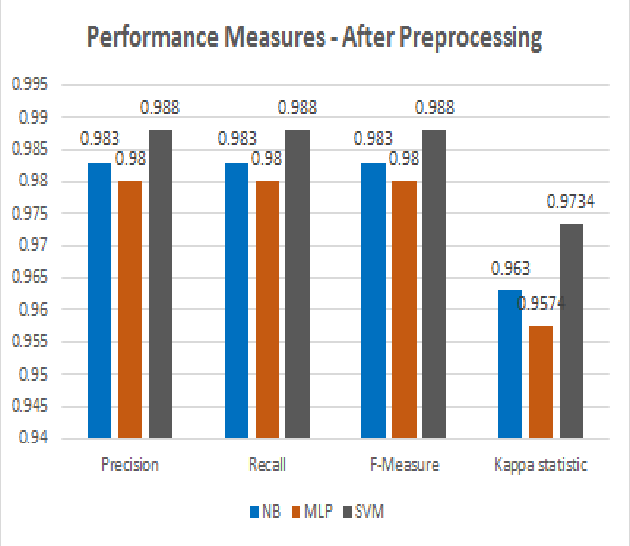


Figure 7: Graph representation depicting Accuracy before preprocessing

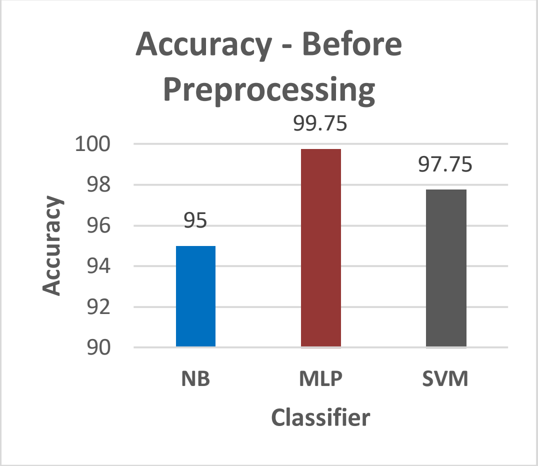
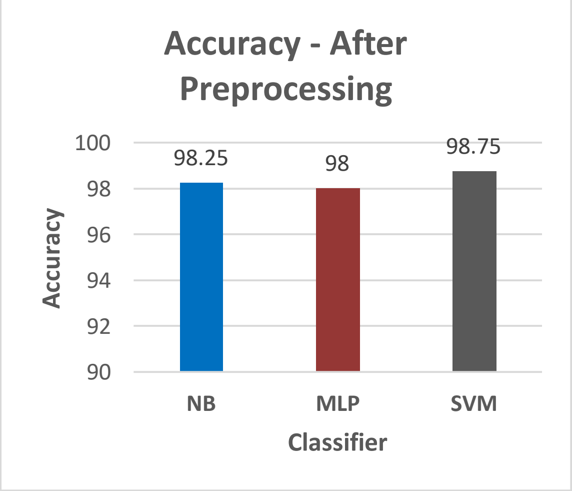


Figure 8: Graph representation depicting Accuracy after preprocessing



VIII. CONCLUSION

In this paper, the experiments were conducted and analysed by taking the chronic kidney disease data set. Three preprocessing techniques namely missing value replacement, normalization and discretization were used to filter the data set. After each filter, three classifiers namely Naïve Bayes, Multilayer Perceptron and Support Vector Machine were applied to the preprocessed data set. Results showed that after preprocessing, Support Vector Machine outperforms all the other classifiers in terms of accuracy. The time taken to build the model is less for Naïve Bayes than other classifiers. But when both accuracy and time are taken together for analysis, Support Vector Machine performs much better than other classifiers.

REFERENCES

1. "Global Facts: About Kidney Disease", National Kidney Foundation, Last modified: Mar 2015, https://www.kidney.org/kidneydisease/global-facts-about-kidney-disease#_ENREF_1
2. J. RadhaKrishnan, M.Sumit, "KI Reports and World Kidney Day", Kidney international reports, ISSN: 2468-0249, vol.2(2), pp. 125-126, Mar. 2017.
3. Vaidic CM, McDonald SP, McCredie MR, van Leeuwen MT, Stewart JH, Law M, Chapman JR, Webster AC, Kaldor JM, Grulich AE, "Cancer incidence before and after kidney transplantation", Journal of the American Medical Association, ISSN: 1538-3598,Vol.296(23), pp. 2823-2831, Dec. 2006.
4. R.Y.Wang, V.C.Storey, C.P.Firth, "A Framework for Analysis of Data Quality Research", IEEE Transactions on Knowledge and Data Mining, ISSN: 1041-4347 Volume 7, Issue 4, pp. 623-640, Aug. 1995.
5. Candelieri, A., Dolce, G., Riganello, F., Sannita, W.G., "Data Mining in Neurology", Knowledge Oriented Applications in Data Mining, IntechOpen, ISBN: 978-953-307-154-1, pp. 261-276, Jan. 2011.
6. Eapen, A.G., "Application of Data mining in Medical Applications", UWSpace,University of Waterloo Library, Ontario, Canada, 2004.
7. boirefillergroup.com, "Data Mining Methodology", 2010, http://www.boirefillergroup.com/methodology.php
8. Sirage Zeynu, Shruti Patil, "Survey on Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques and Feature Selection", International Journal of Pure and Applied Mathematics, Volume. 118, No. 8, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) Special Issue, pp. 149-156, 2018.
9. "Oracle Data Mining Concepts", 11g Release 1 (11.1), May 2008, https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#CHDBHBDI
10. A. Sivakumar, R.Gunasundari, "A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining", International Journal of Pure and Applied Mathematics, Volume. 117, No. 20, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) Special Issue, pp. 785-794, 2017.
11. Tapas Ranjan Baitharu, Subhendu Kumar Pani, "Effect of Missing Values on Data Classification", Journal of Emerging Trends in Engineering and Applied Sciences,Volume. 4(2), ISSN: 2141-7016, pp. 311-316, 2013.
12. Jason Browniee, " How to Normalize and Standardize Your Machine Learning Data in Weka", Weka Machine Learning, July 5, 2016,https://machinlearningmastery.com/normalize-standardize-machine-learning-data-weka.
13. Jason Browniee, "How to Transform Your Machine Learning Data in Weka", Weka Machine Learning, July 8, 2016, https://machinelearningmastery.com/transform- machine-learning-data-weka
14. Gerard Nico, "Statistics – (Discretizing|binning) (bin)", data_mining/discretization.txt, Last modified: 2018/03/24 https://gerardnico.com/data_mining/discretization
15. Rish.I., "An Empirical Study of the Naïve Bayes Classifier", International Joint Conferences on Artificial Intelligence 2001 Workshop on Empirical Methods in Artificial Intelligence, Vol. 3, ISSN: 1045-0823, pp. 41-46, IBM New York, 2001.
16. Sunil Ray, "6 Easy Steps to Learn Naïve Bayes Algorithm with codes in Python and R", September 11, 2017, https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained.
17. Sunita Joshi, Bhuwaneshwari Pandey, Nitin Joshi, "Comparative analysis of Naïve Bayes and J48 Classification", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 12, ISSN: 2277-128X, pp. 813-817, Dec. 2015.
18. Ashfaq Ahmed K., Sultan Aljahdali, Syed Naimatullah Hussain, "Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques", International Journal of Computer Applications, Vol. 69 – No. 11, ISSN: 0975-8887, pp. 12-16, May 2013.
19. Cristianini N., Shawe-Taylor J., "An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods", Cambridge University Press, New York, ISBN: 0-521-78019-5, 189 pp., 2000.
20. Dr. Saed Sayad, "Support Vector Machine – Classification (SVM)", http://saedsayad.com/support_vector_machine.htm
21. Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector", 3rd International Conference on Recent Trends in Computing 2015, Procedia Computer Science, Vol. 57, ISSN: 1877-0509, pp. 500-508, 2015
22. S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014, Compusoft An International Journal of Advanced Computer Technology, ISSN: 2320-0790.

23. Priyadarshini Jambagi, M.S.Shirdhonkar, " Detection of Lung Cancer Using SVM Classification", International Research Journal of Engineering and Technology, Vol. 04, Issue 06, e-ISSN: 2395-0056, p-ISSN: 2395-0072, pp. 378-381, June 2017.
24. Dr. Dilip Roy Chowdhury, Subhashish Ojha, "An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach", International Research Journal of Engineering and Technology, Vol. 04, Issue 01, e-ISSN: 2395-0056, p-ISSN: 2395- 0072, Page 529-534, Jan 2017.
25. Veenita Kunwar, Khushboo Chandel, A. Sai Sabitha, Abhay Bansal, "Chronic Kidney Disease analysis using data mining classification techniques", 6th International Conference – Cloud System and Big Data Engineering (Confluence), Publisher: IEEE Xplore Digital Library, Electronic ISBN: 978-1-4673-8203-8, Date of Conference: 14- 15 Jan 2016, Date added to IEEE Xplore: 11 July 2016
26. Imurat Koklu, Kemal Tutuncu, "Classification of Chronic Kidney Disease with Most Known Data Mining Methods", International Journal of Advances in Science Engineering and Technology, Vol. 5, Issue 1, Special Issue 2, ISSN:2321-9009, pp. 14- 18, Feb 2017.
27. Lambodar Jena, Narendra Ku. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic-Kidney-Disease", International Journal of Emerging Research in Management & Technology, Vol. 4, Issue 11, ISSN: 2278-9359, pp. 110-118, Nov 2015.
28. Aiswarya Iyer, S.Jeyalatha, Ronak Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process, Vol. 5, No. 1, Online ISSN: 2230-9608, Jan 2015.
29. Meriem Amrane, Saliha Oukid, Ikram Gagaoua, Tolga Ensari, "Breast Cancer Classification using Machine Learning", Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, Publisher: IEEE Xplore Digital Library, Electronic ISBN: 978-1-5386-5135-3, Date of Conference: 18-19 April 2018, Date added to IEEE Xplore: 21 June 2018.
30. Omar.Y., Tasleem.A., Pasquier.M., Sagahyroon.A., " Lung Cancer Prognosis System using Data Mining Techniques", In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, Vol. 5: HEALTHINF, ISBN: 978-989-758-281-3, pp. 361-368, 2018.
31. Chronic_Kidney_Disease Data Set, UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems, https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
32. Rajesh Misir, Malay Mitra, RanjitKumar Samanta, "A Reduced Set of Features for Chronic Kidney Disease Prediction", Journal of Pathology Informatics, Vol. 8:24, ISBN: 2229-5089, 2017.
33. John W. Graham, "Missing Data Analysis: Making It Work in the Real World", Annual Review of Psychology, Vol. 60, ISSN: 0066-4308, pp. 549-576, 2009.
34. Hyun Kang, "The prevention and handling of the missing data", Korean Journal of Anesthesiology, Vol. 64, No. 5, ISSN: 2005-6419(print), pp. 402-406, 2013.
35. "Saito.T., Rehmsmeier.M., "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets", PLOS ONE, Vol. 10(3), eISSN: 1932-6203, Mar 2015.
36. M.Sokolova, N.Japkowicz, S.Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation", AI 2006: Advances in Artificial Intelligence, In Proceedings of the 19th Australasian Joint Conference on Artificial Intelligence, Springer, Vol. 4304, online ISBN: 978-3-540-49788-2, pp. 1015- 1021, 2006.
37. Marston.L., "Introductory Statistics for Health and Nursing Using SPSS", SAGE Publications Ltd, ISBN: 13: 9781847874832, 2010.
38. Mary L. McHugh, "Interrater reliability: the kappa statistic", Biochemia Medica, Vol. 22(3), ISSN: 1846-7482, pp. 276-282, Oct 2012.
39. P.Thangaraju, T.Karthikeyan, "PCA-NB Algorithm to Enhance the Predictive Accuracy", International Journal of Engineering and Technology, Vol. 6, No. 1, ISSN: 0975-4024, pp. 381-387, Feb-Mar 2014.