



International Journal of Intellectual Advancements and Research in Engineering Computations

Comparison of feature selection method for chronic kidney data set using data mining classification analytical model

A. Priyadharshini¹, A.Kowsalya², A.Nandhini², A.Poornima², R.Vasugi²

¹Assistant Profesor/IT Department, Vivekanandha College of Engineering for Women (Autonomous)

²UG Students, IT Department, Vivekanandha College of Engineering for Women (Autonomous)

ABSTRACT

Medical Data Analysis is one of the applications of data mining that plays a vital role in human's lives. Data mining has become famous in healthcare management to predict, to detect or to find the hidden patterns and information in health data. In the healthcare commerce different kind of data mining approaches are used to mine the interesting pattern of diseases using the statistical medical data with the help of different machine learning techniques. The conventional disease diagnosis system uses the observation and knowledge of doctor without using the complex clinical data. The planned system assists doctor to predict sickness properly and also the prediction makes patients and medical insurance suppliers benefited. This paper implemented a feature model construction and comparative analysis for improving prediction accuracy of chronic kidney disease dataset in four phases. In first phase, Z-Score normalization algorithm is applied on the original kidney datasets collected from UCI repository. In the second phase of kidney dataset prediction, by the use of Step wise Regression Classification (SRC) Model and Built around the Random Forest Classification algorithm (BRFC) feature selection, subset (data) of Kidney dataset from whole normalized Kidney patient dataset is obtained which comprises only significant attributes. Third phase, classification algorithms are applied on the kidney data set. In the fourth phase, the accuracy will be calculated using Mean Absolute error (MAE), Root Mean square Error (RMSE), Relative Absolute Error(RAE), Root Relative Square Error(RRSE) and kappa values. MLP(Multilayer Perceptron Model) and SVM (Support Vector Model) classification algorithm is considered as the better performance algorithm after applying BRFC and SRC feature selection. Finally, the assessment is done based on accuracy values

Keywords: Medical Data Analysis, Chronic Kidney Data Set, Classification Model, Feature Selection Model, Normalization

INTRODUCTION

Data Mining is one of the most encouraging areas of research with the purpose of finding useful information from voluminous data sets. It has been used in many domains like image mining, opinion mining, web mining, text mining, graph mining etc. Its applications include anomaly detection, financial data analysis, medical data analysis, social network analysis, market analysis etc

Data Mining is particularly useful in medical field when no availability of evidence favoring a

particular treatment option is found. Large amount of complex data is being generated by healthcare industry about patients, diseases, hospitals, medical equipments, claims, treatment cost etc. that requires processing and analysis for knowledge extraction. Data mining comes up with a set of tools and techniques which when applied to this processed data, provides knowledge to healthcare professionals for making appropriate decisions and enhancing the performance of patient management tasks.

Author for correspondence:

IT Department, Vivekanandha College of Engineering for Women (Autonomous)

The meaning of CKD remains intact, but we have clarified the classification and risk stratification as indicated below. The accumulation of 'with implications for health' is intended to reflect the concept that a variety of abnormalities of kidney structure or function may exist, but not all have implications for health of individuals, and therefore need to be contextualized. Kidney damage refers to a broad range of abnormalities observed during clinical assessment, which may be insensitive and non-specific for the cause of disease but may precede reduction in kidney function. Excretory, endocrine and metabolic functions weakening together in most chronic kidney diseases. GFR (Glomerular Filtration Rate) is generally accepted as the best overall index of kidney function. We refer to a GFR of 60 ml/min/1.73 m² as decreased GFR (Table 2) and a GFR of 15 ml/min/1.73 m² as kidney failure. AKI (Acute Kidney Injury) may occur in patients with CKD and hasten the progression to kidney failure.

Complications include drug toxicity, metabolic and endocrine complications, increased risk for CVD (Cardiovascular Disease) and a variety of other recently recognized complications, including infections, frailty, and cognitive impairment. Complications may occur at any stage, often leading to death without progression to kidney failure. Complications may also arise from adverse effects of interventions to prevent or treat the disease and associated comorbidities

This research work focuses on a prediction of kidney disease. Since there are many kidney related diseases, the chronic kidney disease is very dangerous because it leads to kidney failure and also it cannot be predicted at early stages. The chronic kidney disease has stages which can be identified by regular checkup. If the disease is diagnosed than the patient's past history is analyzed. The classification model plays a vital role in the prediction of kidney diseases. The aim of this research work is to develop an efficient predictive healthcare decision support system using data mining techniques. A common kidney or kidney dataset is trained in this system using C4.5, KNN, MLP, SVM and Naïve bayes classification algorithms and tested with the sample data which predict the patient's outcome of Chronic Kidney Diseases.

Data mining has been with success utilized in data discovery for prognostic functions to form a lot of active and correct call. Different data mining techniques i.e. Decision Tree, Bayesian Network, K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Multi layer perceptron etc. are used to predict kidney disease in early stage which also helps to avoid the patient's complications. The main objective of this research work is to predict kidney disease using Step wise Regression Model (SRM) and Built around the Random Forest Classification algorithm (BRFC), the result is obtained by comparing the algorithms and analysis the performance of the algorithm. Different data mining techniques are used to pull data. The experimental comparison of C4.5, KNN, MLP, SVM and NBC are done based on the performance measures of classification accuracy and execution time.

RELATED WORKS

Anu Chaudhary et al., [1] developed a prediction system exploitation A-priori and k-means formula for cardiovascular disease and renal failure prediction. In her survey A-prior and k-mean formula algorithms are went to predict renal failure patient with forty two attributes. They analyzed the information exploitation machine learning tools like distribution and attribute statistics, followed by A-prior and k-means algorithms. They evaluated the information exploitation Receiver operational Characteristic (ROC) plot and standardization plots.

Ashfaq Ahmed et al., [2] have given a piece exploitation machine learning techniques, particularly Support Vector Machine [SVM] and Random Forest [RF]. These were went to study, classify and compare cancer, liver and cardiovascular disease knowledge sets with variable kernels and kernel parameters. Results of Random Forest and Support Vector Machines were compared for various knowledge sets like carcinoma unwellness dataset, disease dataset and cardiovascular disease dataset. It's over that variable results were determined with SVM classification technique with completely different kernel functions.

Giovanni Caocci et al., [3] so as to predict future urinary organ Transplantation Outcome, they

taken discrimination between a man-made Neural Network and supplying Regression. Comparison has been done supported the Sensitivity and specificity of supplying Regression and a man-made Neural Network within the prediction of urinary organ rejection in 10 coaching and corroborative datasets of urinary organ transplant recipients. From the experimental results that each the formula approaches were complementary and their combined algorithms won't to improve the clinical decision-making method and prognosis of urinary organ transplantation.

Lakshmi.K.R et al., [4] analyzed Artificial Neural Networks, call tree and Logical Regression supervised machine learning algorithms. These algorithms are used for urinary organ chemical analysis. For classification method they used an information mining tool named Tanagra. The tenfold cross validation is employed so as to gauge the classified knowledge proceeded by the comparison of these knowledge. From the experimental result they absorbed that ANN performed higher than the choice tree and Logical Regression algorithms.

Neha Sharma et al., [5] detected and expected urinary organ diseases as a prelude to correct treatment to patients. The system was used for detection in patients with kidney disease and also the results of their IF-THEN rules expected the presence of a disease. Their technique used 2 fuzzy systems and a neural network referred to as a neural blur system, supported the results of the input file set obtained. Their system was a mix of fuzzy systems that created results exploitation correct mathematical calculations, rather than probabilistic based mostly classifications. Usually results supported arithmetic tends to possess higher accuracies. Their work was ready to acquire helpful knowledge in conjunction with optimizations in results.

Swathi Baby P et al., [6] incontestable that data processing strategies may be effectively employed in medical applications. Their study collected knowledge from patients affected with excretory organ diseases. The results showed knowledge mining's pertinence during a sort of medical applications. K-means (KM) rule will verify range of clusters in massive knowledge sets. Their study analyzed tree AD, J48, star K, theorem wise, random forest and tree - based ADT naive theorem

on J48 renal disorder knowledge Se and noted that the techniques offer applied mathematics analysis on the utilization of algorithms to predict excretory organ diseases in patients.

Veenita Kunwar et al., [7] in their study had foreseen chronic renal disorder (CKD) mistreatment naive theorem classification and artificial neural network (ANN). Their results showed that naive theorem created correct results than artificial neural networks. it had been conjointly ascertained that classification algorithms were wide used for investigation and identification of CKDs.

Vijayarani and Dhayanand., [8] classification method is employed to classify four varieties of excretory organ diseases. Comparisons of Support Vector Machine (SVM) and Naïve mathematician classification algorithms are done supported the performance factors, classification, accuracy and execution time. As results, the SVM achieves enhanced classification performance. Therefore it's thought-about because the best classifier when put next with Naïve mathematician classifier rule. However, Naïve mathematician classifier classifies the information with minimum execution time. during this study, we tend to apply data processing techniques, recently hierarchic among the highest ten as best classifiers, to predict chronic renal disorder on the idea of the data attributes within the info employed in order to reason patients World Health Organization are littered with the chronic renal disorder (ckd) and patients World Health Organization don't seem to be littered with it (notckd).

Vijayarani and Dhayanand ., [9] have conferred a piece to predict renal disorder by classifying four varieties of excretory organ diseases: Acute Nephritic Syndrome, Chronic renal disorder, Acute kidney failure and Chronic nephritis, mistreatment Support Vector Machine (SVM) and Artificial Neural Network (ANN), then examination the performance of these 2 algorithms on the idea of accuracy and execution time. The results show that the performance of the ANN is healthier than the SVM rule.

Sahil Sharma, [10] applied varied machine learning algorithms to a tangle within the domain of diagnosis and analyzed their potency in predicting the results. The matter selected for the study is that the designation of the chronic

nephropathy. The dataset used for the study consists of four hundred instances and twenty four attributes. The authors evaluated twelve classifications on techniques by applying them to the chronic nephropathy knowledge. So as to calculate potency, results of the prediction by candidate ways were compared with the particular medical results of the topic.

Dr. S. Vijayarani mentioned that, within the health care business, the information mining is preponderantly used for sickness prediction. Huge data processing techniques square measure existing for predicting diseases specifically classification, clustering, association rules, summarizations, regression and etc. the most objective of this analysis work is to predict excretory organ diseases exploitation Naïve Thomas Bayes and Support Vector Machine. This analysis work chiefly targeted on finding the most effective classification rule supported the classification accuracy and execution time performance factors. From the experimental results it's ascertained that the performance of the SVM is healthier than the Naïve Thomas Bayes classifier rule.

METHODOLOGY

In the projected system, a classical approach is projected for locating the diseases of urinary organ cancer victimization data processing classification techniques of Random Forest and Naïve mathematician. The techniques offer profit to the doctors, physicians, medical students and patients to form call relating to the diagnosing of the urinary organ cancer diseases.

The projected KNN primarily based classifier determines neighborhoods directly from coaching observations and it works with numeric feature vector of the urinary organ cancer dataset. The foremost advantage of this approach is that the correct operative coming up with supported image diagnostic information of the urinary organ cancer patients. The projected approach is employed to discover the urinary organ cancer patients affected and also the experimental application shows the results of the potency of the projected approach.

In addition to it for analyzing aid information, major steps information mining approaches like preprocess data, replace missing values, feature choice, machine learning and build call square measure applied on train dataset. Finally the

random forest methodology has been dead on the coaching dataset of urinary organ cancer sickness for the classification method.

- Decision tree predicts a category victimization predefined classification tree with contains each numerical and categorical feature vector.
- To guarantee the validity of result's allotted by distribution varied values of K.
- The application is often utilized by anybody particularly for medical practitioners via web for diagnosing purpose.
- Select the class-outliers, that is, coaching information that square measure classified incorrectly by Random forest (for a given N time kth)

System Architecture

- Step1: Read the chronic Kidney Dataset from UCI Machine learning Repository. The dataset have 400 records.
- Step 2: Normalize the kidney patient dataset using Z-Score Normalization.
- Step 3: Feature extraction will be done by using Step wise Regression Model (SRM) and Built around the Random Forest Classification algorithm (BRFC) feature selection algorithm.
- Step4: The feature will be selected and put in to data frame.
- Step5: Classification algorithms are applied on the selected feature.
- Step6: KNN Classification to create centroid point of data a new group contains the most important data points and others will be considered as outliers
- Step 7: C4.5 classification, multiple trees are induced in the forest, the number of trees is pre-decided by the parameter Ntree.
- Step 8: SVM Classification to create a new group contains the most important data points and others will be considered as outliers.
- Step 9: NBC Classification prediction values for rpcc and rcc and compare train dataset and accuracy calculate.
- Step 10: To apply kidney dataset using MLP classification model and accuracy calculated.
- Step11: The results are obtained, MLP and SVM gives better accuracy when compare to other algorithms.
- Step12: Accuracy will be analyzed.
- Step13: Finally the evaluation metrics will be calculated.

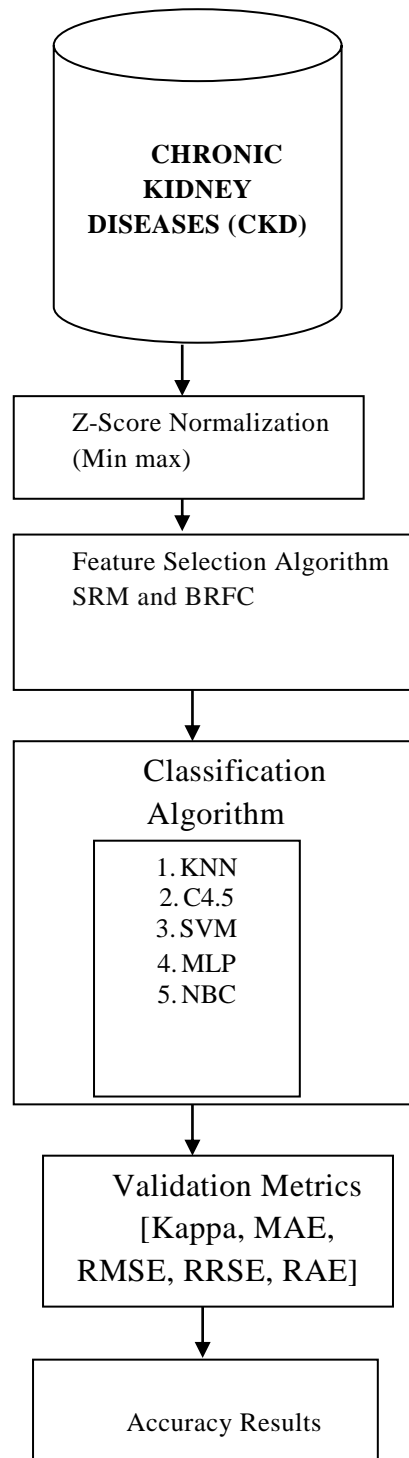


Fig 3.1 System architecture

Normalization

Normalization is scaling technique or a pre process stage. Where, we are able to discover new dimension from associate degree existing one series. It is often useful for the prediction or

statement operates heaps. Therefore maintain the big distinction of prediction and statement the standardization technique is needed to form them nearer.

Z-Score Normalization

Z-score is that the variety of normal deviations from the mean an information purpose (kidney). However additional technically it's calculated of what percentage normal deviations below or on top of the population means that a rough score. A z-score is additionally referred to as a customary score and it are often placed on a standard distribution curve. Z-scores vary from -3 normal deviations (which would fall to the left of the conventional distribution curve) up to +3 normal deviations (which would fall to the way right of the conventional distribution curve). so as to use a z-score, you wish to spot the mean μ and conjointly the population variance σ .

$$z = (x - \mu) / \sigma$$

FEATURE SELECTION

Feature extraction is that the model of choosing a set of the terms gift within the coaching set and victimization solely this set as options in text classification. Feature extractions provide 2 main functions. First, it makes coaching and applying a classifier additional powerful by decreasing the scale of the adequate vocabulary. Feature extraction method is of explicit significance for classifiers that, unlike NB, square measure costly to coach. Second, feature extraction typically will increase classification accuracy by eliminating noise options. A noise feature is one that, once joined to the document illustration, will increase the classification error on new knowledge. Facilitating knowledge visual image is dashing up the execution of mining algorithms and reducing descending times

Stepwise Regression

Stepwise regression n may be a combination of the forward and backward choice techniques. Stepwise regression may be a modification of the forward choice in order that when every step within which a variable was added , all candidate variables within the model square measure checked to examine if their significance has been reduced below the required tolerance level. If a no important variable is found, it's aloof from the model. Stepwise regression needs 2 significance levels: one for adding variables and one for

removing variables. The cutoff likelihood for adding variables ought to be but the cutoff likelihood for removing variables in order that the procedure doesn't get into associate degree infinite loop.

Backward Elimination

- This can be the only of all excretory organ variable choice procedures and may be simply enforced while not special software package. In things wherever there's a fancy hierarchy, backward elimination are often run manually whereas taking account of what variables square measure eligible for removal.
- Begin with all the chronic excretory organ predictors within the model
- Take away the predictor with highest p-value bigger than Θ
- Refit the model and goto two stop when all p-values are less than \square_{crit} .
- The \square_{crit} is typically referred to as the "p-to-remove" and doesn't have to be five-hitter. If prediction performance is that the goal, then a 15-20% cut-off may match best, though ways designed additional directly for best prediction ought to be most popular.

Forward choice

- This simply reverses the backward technique.
- Start with no excretory organ variables within the model.
- For all predictors not within the model, check their p-value if they're intercalary to the model.
- Choose the one with lowest p-value but $\text{crit.}\alpha$
- Continue till no new predictors are often intercalary.

ADVANTAGES

- The procedure yields one final model, though there are a unit usually many equally sensible models. Stepwise regression doesn't take into consideration a researcher's information regarding the predictors.
- It's going to be necessary to force the procedure to incorporate vital predictors. One mustn't over-interpret the order within which predictor's are a unit entered into the model.
- One mustn't jump to the conclusion that everyone the vital predictor variables for predicting y are known, or that everyone the

unimportant predictor variables are eliminated. It is, of course, attainable that we have a tendency to could have committed a kind I or kind II error on the means.

BUILT AROUND THE RANDOM FOREST CLASSIFICATION ALGORITHM (BRFC)

The r designed around Random Forest rule is employed within the current work each as a classifier associated as an engine for the feature choice rule; therefore we tend to offer below a brief outline of its most significant qualities. It's designed as associate ensemble of weak classifiers that mix their results throughout the ultimate classification of every object. Individual classifiers square measure designed as classification trees. Every tree is made exploitation totally different bootstrap sample of the coaching set, roughly 1/ three of objects isn't used for building a tree. At every step of the tree construction a special set of attributes is willy-nilly way} designated associated a split is performed exploitation an attribute which results in a best distribution of information between nodes of the tree

BRFC mechanisms goal is to work out whether or not or not it's attainable to pick a set of trees from a forest that's ready to surpass this forest. The analysis work aim isn't to search out the optimum set of individual classifiers among an oversized ensemble of trees, however rather to check the extent to that it's attainable to boost accuracy of a BRFC by specializing in some explicit subsets of trees. The "final" goal of this work is therefore to spot some explicit properties that square measure shared by these sub-forests, and also the tree choice approach we tend to propose during this analysis work could be a opening move toward this direction.

Algorithm

Let D is a coaching dataset in associate degree M -dimensional area X , and let Y be the category feature with total range of c distinct categories. the strategy for building a random forest follows the method as well as 3 steps:

- Step 1: coaching knowledge sampling: use the cloth methodology to come up with K subsets of coaching knowledge by willy-nilly sampling D with replacement;

- Step 2: Feature mathematical space sampling and tree classifier building: for every coaching dataset D_i ($1 \leq i \leq K$), use a call tree formula to grow a tree. At every node, willy-nilly sample a mathematical space X_i of F options ($F \ll M$), figure all splits in mathematical space X_i , and choose the most effective split because the cacophonic feature to come up with a baby node. Repeat this method till the stopping criteria is met and a tree $h_i(D_i, X_i)$ designed by coaching knowledge D_i below mathematical space X_i is so obtained.
- Step 3: call aggregation: ensemble the K trees to create a random forest associate degreed use the bulk vote of those trees to create an ensemble classification call.

To enhance accuracy, the randomness injected needs to minimize the correlation ρ whereas maintaining strength. The forests studied here accommodate exploitation willy-nilly chosen inputs or mixtures of inputs at every node to grow every tree. The ensuing forests provide accuracy that compare favorably with Ada-boost. This category of procedures has fascinating characteristics:

- Its accuracy is pretty much as good as Ada-boost and typically higher.
- It's comparatively sturdy to outliers and noise.
- It's quicker than cloth or boosting.
- It offers helpful internal estimates of error, strength, correlation and variable importance.
- It's straightforward and simply parallelized.

CLASSIFICATION ALGORITHM

Algorithm

C4.5 is AN algorithmic program accustomed manufacture a choice tree that is increase of previous ID3 calculation. It en-large the ID3 algorithmic program is managing each continuous and distinct property, missing values and pruning trees once construction. The choice trees created by C4.5 are often used for grouping and sometimes cited as a applied math classifier. C4.5 creates call trees from a group of coaching urinary organ information same approach as Id3 algorithmic program. Because it could be a supervised learning algorithmic program it needs a group of coaching examples which may be seen as a pair: input object and a desired output worth (class). The algorithmic program analyzes the coaching set and frame a

classifier that has to have the dimensions to accurately prepare each coaching and take a look at cases

Naive Thomas Bayes Algorithm

The Naive Bayesian classifier relies on Bayes' theorem with independence assumptions between predictors. Naive Thomas Bayes classifiers are a family of straightforward probabilistic classifiers supported applying theorem. Thomas Bayes theorem provides some way of conveying the posterior likelihood, $P(c/x)$, from $P(c)$, $P(x)$, and $P(x/c)$. It assumes that the result of the worth of a predictor (x) on a given class (c) is freelance of the values of alternative predictors. This assumption is named category conditional independence. The Naïve Bayesian classification predicts that the tuple ' x ' belongs to the category ' c ' victimization the formula.

$$P(c/x) = (x/c) / (P(x))$$

- $P(c/x)$ is that the posterior likelihood of sophistication (target) given predictor (attribute).
- $P(c)$ is that the previous likelihood of sophistication.
- $P(x/c)$ is that the chance that is that the likelihood of predictor given category.
- $P(x)$ is that the previous likelihood of predictor.

KNN Classification

K-Nearest Neighbor (Knn) –Techniques
KNN could be a supervised learning algorithmic program that classifies new information supported minimum distance from the new information to the K nearest neighbor. The projected work has used geometrician Distance to outline the closeness. Pseudo-code for the KNN classifier is declared below:

- Step 1: Input: $D = \{x_1, \dots, x_n\}$ new instance to be classified
- Step 2: for every labeled instance (x_i, c_i) Calculated (x_i, x)
- Step 3: Ordered (x_i, x) from lowest to highest, $(i=1, \dots, N)$
- Step 4: Select the K nearest instances to x : $D_x K$
- Step 5: Assign to x the foremost frequent category in $D_x K$

MLP (Multilayer Perceptron)

A multilayer perceptron (MLP) could be a feed forward artificial neural network model that maps urinary organ datasets of input file onto a collection of applicable outputs. Associative MLP classification could be a multiple layers of nodes in a much directed graph, with every layer totally connected to following one. A side from the input nodes, every node could be a nerve cell (or process element) with a nonlinear activation perform. MLP classification urinary organ dataset utilizes a supervised learning technique known as back propagation for coaching urinary organ the network. MLP could be a modification of the quality linear perceptron and might distinguish knowledge that isn't linearly dissociable urinary organ dataset method.

SVM Classification

Support Vector Machine SVM have attracted an excellent deal of attention within the last decade and actively tested to varied domains applications. SVMs are unit largely used for learning classification, regression or ranking operate. SVM are unit supported applied math learning theory and structural risk diminution principal and have the intent of crucial the placement of call boundaries conjointly called hyper plane that manufacture the optimum separation of categories. SVM is that the most strong and precise classification technique, there are a unit several issues. The info analysis in SVM relies on bulging quadratic programming, and it's computationally expensive, as finding quadratic programming ways need giant matrix operations likewise as time intense numerical computations.

Analysis Metric

Mean Absolute Error

Statistical exactness metrics value the accuracy of a system by examination the numerical recommendation scores against the particular user ratings for the user-item pairs within the take a look at dataset. Mean Absolute Error (MAE) between ratings and predictions could be a wide used metric

Root Mean sq. Error

The Root Mean sq. Error (RMSE) (also known as the foundation mean sq. deviation, RMSD) could

be a oftentimes used live of the distinction between values expected by a model and therefore the values truly determined from the setting that's being modelled. These individual variations also are known as residuals, and therefore the RMSE serves to combination them into one live of prognostic power. The RMSE of a model prediction with relevancy the calculable variable Xmodel is outlined because the root of the mean square error:

Root Relative Squared Error

Correlation – usually measured as a parametric statistic – indicates the strength and

direction of a linear relationship between 2 variables (for example model output and determined values). Variety of various coefficients square measure used for various thing

Kappa Metrics

It returns the constant value. It measures the agreement between classification and truth values. It of one represents good agreement, whereas a price of zero represents no agreement

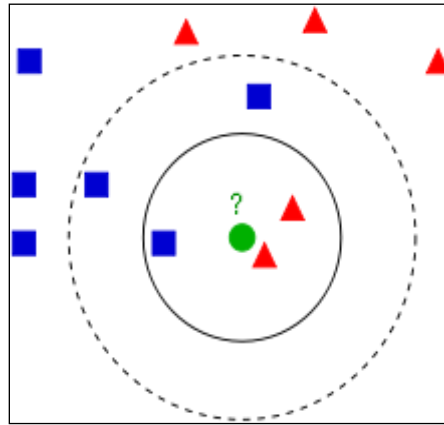


Fig 3.2 KNN Classification

EXPERIMENTAL RESULTS

The dataset employed in this paper has been obtained from UCI supply. The knowledge set contains data of four hundred samples from the southern a part of Asian country with their ages travel between a pair of - ninety years. There are a unit in total twenty - four options, most of that area unit clinical in nature and also the rest area unit physiological. Table 4.1 summarizes varied

parameters. As a locality of knowledge pre - process, missing worth and outlier's area unit imputed with norm of that feature for continuous knowledge and attribute model value for categorical knowledge. Nominal knowledge is a unit born-again to numerical values. For instance, Nominal values 'Normal' area unit labeled "1" and 'Abnormal' area unit labeled "0".

Table 4.1 Dataset Attribute Values

ATTRIBUTE	REPRESENTATION	INFORMATION ATTRIBUTE	DESCRIPTION
Specific Gravity	Sg	Numerical	1.005, 1.010, 1.015, 1.020,1.025
Albumin	Al	Numerical	0.1.2.3.4.5
Sugar	Su	Numerical	0.1.2.3.4.5
Red Blood Cells	Rbc	Nominal	Normal, abnormal
Pus Cell	Pc	Nominal	Normal, abnormal
Bacteria	ba	Nominal	Present, not present
Hypertension	Htn	Nominal	Yes, no
Diabetes Mellitus	Dm	Nominal	Yes, no

Coronary Artery Disease	Cad	Nominal	Yes, no
Appetite	appet	Nominal	Good, poor
Pedal Edema	pe	Nominal	Yes, no
Anemia	ane	Nominal	Yes, no
Pus Cell clumps	pcc	Nominal	Present, not present
Age	age	Numerical	Years
Blood Pressure	bp	Numerical	Mm/Hg
Blood Urea	bu	Numerical	Mgs/dl
Serum Creatinine	sc	Numerical	Mgs/dl
Sodium	sod	Numerical	mEq/L
Potassium	pot	Numerical	mEq/L
Hemoglobin	Hemo	Numerical	Gms
Packed Cell Volume	pcv	Numerical	
White Blood Cell Count	wbcc	Numerical	Cells/cumm
Red Blood Cell Count	rbcc	Numerical	Millions/cmm
Blood Glucose Random	bgr	Numerical	Mgs/dl
Class	class	Nominal	Ckd /notckd

The step wise regression model feature extraction attributes describe the following table 4.2,

Table 4.2 Feature Selection Dataset-Stepwise Regression Model

Attributes Type			
Gender Categorical			
1	hemo	9	Pcv
2	sg	10	Bgr
3	htn	11	Cad
4	rbcc	12	ane
5	rbc	13	class
6	al		
7	dm		

The built around the Random Forest Classification Algorithm feature extraction attributes describe the following table 4.3,

Table 4.3 Feature Selection Dataset-BARFC

Attributes Type			
Gender Categorical			
1	Sg	9	rbc
2	Al	10	bgr
3	Hemo	11	class
4	Pcv		
5	Sc		
6	Rbcc		
7	Dm		
8	htm		

The table 4.4 describes the Kappa metrics value for SRC and BARFC algorithm model. The table contain classification algorithm and selection attribute details for SRC and BAFRC selective attribute model.

Table 4.4 Kappa Value- SRC and BAFRC

S.NO	Classification Algorithm	SRC Selection Model (12 Attributes)	BAFRC Selection Model (10 Attributes)
		(hemo, sg, htn, rbcc, rbc ,al ,dm ,bu, pcv ,bgr ,cad, ane)	Sg, Al, hemo, Pcv,sc, rbcc,dm, htm, rbc, bgr)
1	C4.5	0.954	0.966
2	SVM	0.963	0.974
3	KNN	0.942	0.958
4	NB	0.877	0.902
5	MLP	0.956	0.963

The figure Fig 5.1 describes the Kappa metrics value for SRC and BARFC algorithm model. The figure contain classification algorithm and selection

attribute details for SRC and BAFRC selective attribute model

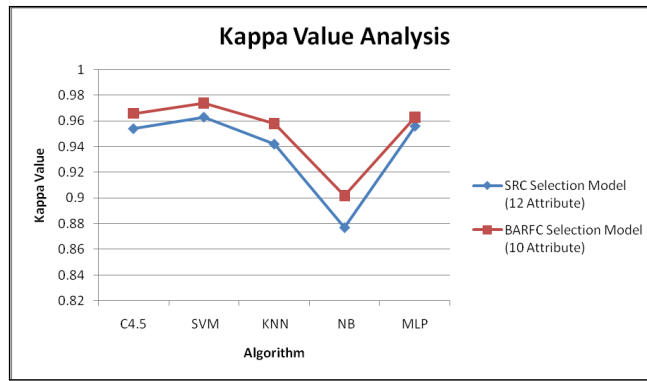


Fig 5.1 Kappa Value- SRC and BAFRC

The table 4.5 describes the accuracy value for SRC and BARFC algorithm model. The table contain classification algorithm and selection

attribute details for SRC and BAFRC selective attribute model.

Table 4.5 Accuracy Value- SRC and BAFRC

S.NO	Classification Algorithm	SRC Selection Model (12 Attribute)	BAFRC Selection Model (10 Attribute)
		(hemo, sg, htn ,rbcc, rbc ,al ,dm ,bu ,pcv ,bgr ,cad, ane)	Sg, Al, hemo, Pcv,sc, rbcc,dm, htm, rbc, bgr)
1	C4.5	94.66	94.78
2	SVM	95.45	95.85
3	KNN	89.67	89.75
4	NB	90.16	91.22
5	MLP	93.25	95.95

The figure Fig 5.2 describes the Kappa metrics value for SRC and BARFC algorithm model. The figure contain classification algorithm and selection

attribute details for SRC and BAFRC selective attribute model.

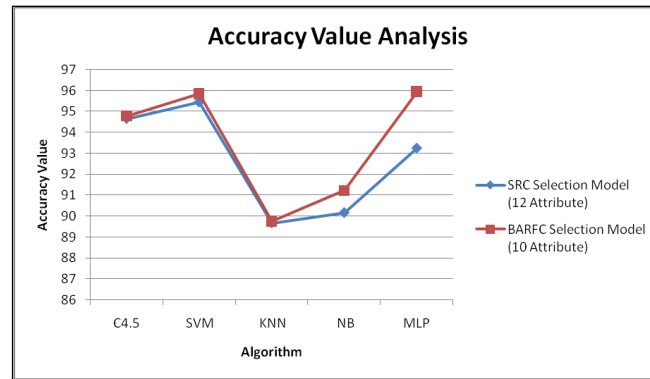


Fig 5.2 Accuracy Value- SRC and BARFC

CONCLUSION AND FUTURE ENHANCEMENT

In this paper, a novel system is proposed for predicting the diseases like kidney cancer using data mining classification technique such as Random forest, Naive Bayes search and classification. The system gives benefit to the doctors, physicians, medical students and patients to make decision regarding the diagnosis of the cancer diseases particularly kidney cancer. Wide use of imaging techniques enables early detection of renal tumors. This provides an opportunity for removing only the cancer lesion with a necessary margin facilitating preservation of as much healthy kidney parenchyma volume as possible. Finally, the thesis uses the K-Nearest Neighbor (KNN) Algorithms as supervised classification model. And the proposed system calculates and compares the

accuracy of Random forest, Naive Bayes and KNN and the experimental result demonstrates that the MLP and SVM provides better accuracy for chronic kidney dataset

The proposed methodology is used to predict the kidney cancer region into separable compartments i.e. the kidney, tumor etc. However, the method requires further improvement mostly regarding feature selection and segmentation of the kidney into multiple components: renal cortex, renal column, renal medulla and renal pelvis. In addition to that, it is intended to increase the database on which the system will be tested. And also the proposed method in this project can be employed for detecting the heart diseases in future with the heart and liver dataset and classification of the diseases.

REFERENCES

- [1]. Anu Chaudhary, Puneet Garg, "Detecting and Diagnosing a Disease by Patient Monitoring System", International Journal of Mechanical Engineering And Information Technology, 2, 6, 2015, 493-499.
- [2]. Ashfaq Ahmed, K., Aljahdali, S., Hussain, S.N.: "Comparative prediction performance with support vector machine and random forest classification techniques", International Journal Computer Applications. 69(11), 2016, 12–16 .
- [3]. 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, <http://dx.doi.org/10.5772/53104>, 2017.
- [4]. Lakshmi. K.R, Nagesh. Y and VeeraKrishna. M "Performance Comparison Of Three Data Mining Techniques For Predicting Kidney Dialysis Survivability", International Journal of Advances in Engineering & Technology, Mar., 7(1), 2016, 242-254.
- [5]. Neha Sharma, Er. Rohit Kumar Verma, "Prediction of Kidney Disease by using Data Mining Techniques", Prediction of Kidney Disease by using Data Mining Techniques, 2016.
- [6]. Swathi Baby P and Panduranga Vital T, "Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT), 2015.

- [7]. Veenita Kunwar, Khushboo Chandel, A. Sai Sabitha, and Abhay Bansal, “Chronic Kidney Disease Analysis Using Data Mining Classification Techniques”, IEEE, 2016.
- [8]. Vijayarani, S., Dhayanand, S.: “Data mining classification algorithms for kidney disease prediction”, International Journal of Cybern. Inf. (IJCI) 4(4), 2017, 13–25.
- [9]. Vijayarani, S., Dhayanand, S.: “Kidney disease prediction using SVM and ANN algorithms”, International Journal Comput. Business Res. 6(2), 2017.
- [10]. Sharma, S., Sharma, V., & Sharma, A. (2017). Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis. arXiv preprint arXiv:1606.09581,2017.