Data Mining Classification Technique towards the Prediction of Kidney Disease

YERROLLA SREEKANTH[1], S. Nagaraju[2]
[1]Student, Master of Technology, KITS, Warangal
[2]Associate Professor, KITS, Warangal

Abstract: ***The gigantic measures of data produced by healthcare transactions are excessively mind boggling and voluminous, making it impossible to be handled and examined by conventional strategies. Data mining gives the strategy and innovation to change these hills of data into helpful data for decision making. The Healthcare business is for the most part "data rich", which isn't achievable to deal with physically. These a lot of data are essential in the field of data mining to separate helpful data and create connections among the characteristics. Kidney disease is a perplexing task which requires much understanding and knowledge. Kidney disease is a noiseless killer in created nations and one of the primary supporters of disease trouble in creating nations. In the human services industry the data mining is basically utilized for anticipating the diseases from the datasets. The Data mining classification systems, specifically Decision trees, ANN, Naive Bayes are examined on Kidney disease data set.***

Index Terms: Kidney Disease, Data Mining, Decision tree, ANN, Naive Bayes, K-NN, Rough Set, SVM, Genetic Algorithms (GAs) / Evolutionary Programming (EP), Logistic Regression, Clustering

I. INTRODUCTION

Data mining is the non-inconsequential extraction of certain already unknown and possibly valuable data about data [4]. Data Mining is a standout amongst the most essential and inspiring zone of research with the target of finding important data from enormous data sets. In introduce period, Data Mining is getting to be prevalent in healthcare field on the grounds that there is a need of proficient diagnostic technique for recognizing unknown and important data in wellbeing data [1]. Restorative data mining is utilized as a part of the knowledge securing and examinations the data acquired from investigate reports, medicinal reports, stream graphs, confirm tables, and change these hills of data into valuable data for decision making[2]. This paper showed the utility of Classification systems for anticipating Kidney disease with various data mining devices.

Organization of the paper: Section II Kidney diseases variables, side effects and kind of kidney disease. Area III portrays writing audits. Area IV portrays data mining Classification strategies. Exploratory outcomes are exhibited in Section V lastly, Section VI finishes up the paper and brings up some potential future work.

II. KIDNEYDISEASE

The kidneys' functions are to filter the blood. All the blood in our bodies goes through the kidneys a few times each day. The kidneys evacuate squanders, control the body's liquid adjust, and direct the adjust of electrolytes. As the kidneys filter blood, they make pee, which gathers in the kidneys' pelvis - funnel-molded structures that deplete down tubes called ureters to the bladder. Every kidney contains around a million units called nephrons, every one of which is a minuscule filter for blood. It's conceivable to lose as much as 90% of kidney work without encountering any side effects or issues. Kidney disease is a noiseless killer [5].

There are number of factors which increase the risk of Kidney disease:

- Diabetes
- Hypertension
- Smoking
- Obesity
- Heartdisease
- Family history of Kidneydisease
- Alcoholintake
- Drug abuse/drugoverdose
- Age
- Race/Ethnicity
- Malesex

Symptoms of kidney disease:

- Changes in your urinaryfunction
- Difficulty or pain duringvoiding
- Blood in theurine
- Swelling & Pain in the back orsides
- Extreme fatigue and generalizedweakness
- Dizziness & Inability toconcentrate
- Feeling cold all thetime
- Skin rashes anditching
- Ammonia breath and metallictaste

- Nausea andvomiting
- Shortness ofbreath

Types of Kidney diseases:

❖ Pyelonephritis (infection of kidney pelvis): Bacteria may infect the kidney, usually causing back pain and fever. A spread of bacteria from an untreated bladder infection is the most common cause ofpyelonephritis.
❖ Glomerulonephritis: An overactive immune system may attack the kidney, causing inflammation and some damage. Blood and protein in the urine are common problems that occur with glomerulonephritis. It can also result in kidneyfailure.

❖ Kidney stones (nephrolithiasis): Minerals in urine form crystals (stones), which may grow large enough to block urine flow. It's considered one of the most painful conditions. Most kidney stones pass on their own but some are too large and need to betreated.

❖ Nephrotic syndrome: Damage to the kidneys causes them to spill large amounts of protein into the urine. Leg swelling (edema) may be asymptom.
❖ Polycystic kidney disease: A genetic condition resulting in large cysts in both kidneys that impair their function.

❖ Acute renal failure (kidney failure): A sudden worsening in kidney function. Dehydration, a blockage in the urinary tract, or kidney damage can cause acute renal failure, which may bereversible.

❖ Chronic renal failure: A permanent partial loss of kidney function. Diabetes and high blood pressure are the most commoncauses.
❖ End stage renal disease (ESRD): Complete loss of kidney function, usually due to progressive chronic kidney disease. People with ESRD require regular dialysis forsurvival.

❖ Diabetic nephropathy: High blood sugar from diabetes progressively damages the kidneys, eventually causing chronic kidney disease. Protein in the urine (nephrotic syndrome) may alsoresult.
❖ Hypertensive nephropathy: Kidney damage caused by high blood pressure. Chronic renal failure may eventuallyresult.

❖ Kidney cancer: Renal cell carcinoma is the most common cancer affecting the kidney. Smoking is the most common cause of kidneycancer.

❖ Interstitial nephritis: Inflammation of the connective tissue inside the kidney, often causing acute renal failure. Allergic reactions and drug side effects are the usualcauses.

❖ Minimal change disease: A form of nephrotic syndrome in which kidney cells look almost normal under the microscope. The disease can cause significant leg swelling (edema). Steroids are used to treat minimal changedisease.

❖ Nephrogenic diabetes insipidus: The kidneys lose the ability to concentrate the urine, usually due to a drug reaction. Although it's rarely dangerous, diabetes insipidus causes constant thirst and frequenturination.

❖ Renal cyst: A benign hollowed-out space in the kidney. Isolated kidney cysts occur in many normal people and almost never impair kidneyfunction.

III. LITERATUREREVIEW

This area comprises of the audits of different specialized and survey articles on data mining methods connected to anticipate Kidney Disease.

- DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi et al [6]. depicted in their exploration to comprehend machine learning strategies to anticipate kidney stones. They anticipated great precision with C4.5, Classification tree and Random backwoods (93%) trailed by Support Vector Machines (SVM) (91.98%). Logistic and NN has likewise demonstrated great precision comes about with zero relative total blunder and 100% effectively characterized outcomes. ROC and Calibration bends utilizing Naive Bayes has likewise been built for anticipating exactness of the data. Machine learning approaches give better outcomes in the treatment of kidney stones.

- J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van nook Berghe, K.U.Leuven et al [7]. Investigated data mining systems for foreseeing intense kidney damage after elective cardiovascular surgery with Gaussian process and machine learning strategies (classification task and regression task).

- K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna et al [8]. exhibited execution examination of Artificial Neural Networks, Decision Tree and Logical Regression are utilized for Kidney dialysis survivability. The data mining systems were assessed in light of the precision measures, for example, classification exactness, affectability and specificity. They accomplished outcomes utilizing 10 overlap cross-approvals and perplexity grid for every method. They discovered ANN demonstrates better outcomes. Subsequently ANN demonstrates the solid outcomes with Kidney dialysis of patient records.

- Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri et al [9]l. portrayed in their examination by utilizing administered systems to anticipate the early risk of AVF disappointment in patients.

- They utilized classification ways to deal with foresee likelihood of entanglement in new hemodialysis patients whom have been alluded

by nephrologists to AVF surgery.

- Abeer Y. Al-Hyari et al [10].proposed in their exploration by utilizing Artificial Neural Network (NN), Decision Tree (DT) and Naïve Bayes (NB) to anticipate interminable kidney disease. The proposed NN calculation and the other data mining algorithms showed high potential in fruitful kidney disease.

- Xudong Song, Zhanzhi Qiu, Jianwei Mu et al [11].introduced data mining decision tree classification technique, and proposed another variable exactness rough set decision tree classification calculation in view of weighted utmost number express locale.

- N. SRIRAAM, V. NATASHA and H. KAUR et al [12].presented data mining approach for parametric assessment to enhance the treatment of kidney dialysis persistent. Their trial result demonstrates that classification exactness utilizing Association mining between the reaches 50– 97.7% is acquired in light of the dialysis parameter blend. Such a decision-based approach causes the clinician to choose the level of dialysis required for singular patient.

- Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh et al [13]. Their exploration depicts a productive Diagnosis of Kidney Images Using Association Rules. Their approach is separated into four noteworthy steps: pre-preparing, highlight extraction and determination, affiliation run age, and age of analysis proposals from classifier.

- Divya Jain et al [14].presented impact of diabetes on kidney utilizing C4.5 calculation with Tanagra device. The execution of classifier is assessed as far as review, accuracy and blunder rate.

- Koushal Kumar and Abhishek et al [15].their examine portrays correlation of each of the three neural networks, for example, (MLP, LVQ, RBF) based on its precision, time taken to construct model, and preparing data set size.

IV. DATAMININGTECHNIQUESUSEDFORPREDICTIONS

Classification is an imperative data mining task, and the motivation behind classification is to propose a classification capacity or classification display (called classifier).The classification model can outline data in the database to a particular class. Classification development strategies include: Decision Tree, Naive Bayes, ANN, K-NN, Support Vector Machine, Rough set, Logistic Regression, Genetic Algorithms (GAs)/Evolutionary Programming (EP), Clustering and so forth [3].

Decision Tree: The decision tree is a structure that incorporates root hub, branch and leaf hub. Each interior hub signifies a test on characteristic, each branch indicates the result of test and each leaf hub holds the class name. The highest hub in the tree is the root hub. The decision tree approach is all the more capable for classification issues. There are two steps in this strategies assembling a tree and applying the tree to the dataset. There are numerous prevalent decision tree algorithms CART, ID3, C4.5, CHAID, and J48.

Artificial Neural Network (ANN): is an accumulation of neuron – like preparing units with weight connections between the units. It maps a set of information data onto a set of fitting yield data. It comprises of 3 layers: input layer, shrouded layer and yield layer. There is connection between each layer and weights are relegated to every connection. The essential capacity of neurons of information layer is to partition enter xi into neurons in concealed layer. Neuron of concealed layer includes input flag xi with weights wji of particular connections from input layer. The yield Yj is capacity of $Yj = f(\Sigma wji\ xi)$ Where f is a straightforward edge capacity, for example, sigmoid or hyperbolic digression work.

Naive Bayes: Naive Bayes classifier depends on Bayes theorem. This classifier calculation utilizes restrictive independence, implies it expect that a quality incentive on a given class is independent of the estimations of other properties. The Bayes theorem is as per the following: Let X={x1, x2... xn} be a set of n characteristics. In Bayesian, X is considered as proof and H is some hypothesis implies, the data of X has a place with particular class C. We need to decide P (H|X), the likelihood that the hypothesis H holds given proof i.e. data test X. As per Bayes theorem the P (H|X) is communicated as $P(H|X) = P(X|H)\ P(H)/P(X)$.

K-Nearest Neighbor: The k-nearest neighbor's calculation (K-NN) is a technique for arranging objects in view of nearest preparing data in the component space. K-NN is a sort of occasion based learning. The k-nearest neighbor calculation is among the least difficult of all machine learning algorithms. Be that as it may, the exactness of the k-NN calculation can be seriously corrupted by the nearness of uproarious or unessential highlights, or if the component scales are not predictable with their significance.

Logistic Regression: The term regression can be characterized as the estimating and breaking down the connection between at least one independent variable and dependent variable. Regression can be characterized by two classes; they are direct regression and logistic regression. Logistic regression is a summed up by straight regression. It is basically utilized for evaluating parallel or multi-class dependent factors and the reaction variable is discrete, it cannot be demonstrated specifically by straight regression i.e. discrete variable changed into consistent esteem. Logistic regression essentially is utilized to order the low dimensional data having non-straight limits. It likewise gives the distinction in the level of dependent variable and gives the rank of individual variable as per its significance. Thus, the principle witticism of Logistic regression is to decide the consequence of every factor accurately.

Rough Sets: A Rough Set is controlled by a lower and upper bound of a set. Each individual from the lower bound is a sure individual from the set. Each non-individual from the upper bound is a sure non-individual from the set. The upper bound of a rough set is the joining between the lower bound and the supposed limit locale. An individual from the limit area is perhaps (yet not positively) an individual from the set. Therefore, rough sets might be seen as with a three-esteemed enrollment work (yes, no, maybe). Rough sets are a mathematical concept managing Uncertainty in data. They are generally joined with other strategies, for example, manage enlistment or clustering

techniques.

Support Vector Machine (SVM): Support vector machine (SVM) is a calculation that endeavors to locate a straight separator (hyper-plane) between the data purposes of two classes in multidimensional space. SVMs are appropriate to managing connections among highlights and repetitive highlights.

Genetic Algorithms (GAs)/Evolutionary Programming (EP): Genetic algorithms and evolutionary programming are utilized as a part of data mining to define hypotheses about dependencies between factors, as affiliation principles or some other interior formalism.

Support Vector Machine (SVM): Support vector machine (SVM) is a calculation that endeavors to locate a straight separator (hyper-plane) between the data purposes of two classes in multidimensional space. SVMs are appropriate to managing cooperations among highlights and repetitive highlights.

Clustering: Clustering is the way toward gathering comparable components. This method might be utilized as a pre-handling step before nourishing the data to the arranging model. The credit esteems should be standardized before clustering to maintain a strategic distance from high esteem characteristics ruling the low esteem properties. Further, classification is performed in light of clustering.

V. EXPERIMENTAL RESULTS ANDANALYSIS

This area exhibits the trial results and examination improved the situation this investigation. The data mining procedures utilized for kidney disease forecast has been clarified in area IV. For the tests, different diverse classification methods have been connected to anticipate kidney disease. Table 1 demonstrates the outcomes for classification strategies to foresee Kidney disease with various mining instruments for this work.

Table1. Results of classification techniques used for kidney disease:

Author	Publication Year	Type of Kidney Disease	Tool	Techniques	Accuracy
Kaladhar et al.[6]	2012	Kidney stone	WEKA	Naive Bayes	0.99%
				Logistic	1.00%
				J48	0.97%
				Random Forest	0.98%
			ORANGE	Naive Bayes	0.79%
				K-NN	0.7377%
				Classification tree	0.9352%
				C4.5	0.9352%
				SVM	0.9198%
				Random Forest	0.9352%
K.R.Lakshmi et al.[8]	2014	Kidney dialysis	TANAGRA	ANN	93.852%
				Decision Tree(C5)	78.4455%
				Logical Regression	74.7438%
J.Van Eyck et al[7]	2012	AKI	MATLAB	Gaussian process(aROC)	0.758%
				Gussian process(RMSER)	0.408%
Morteza Khavanin Zadeh et al.[9]	2012	Early AVF Failure	WEKA	W-Simple Cart	85.11%
				WJ48	80.85%
Abeer Y. Al- Hyari et al[10]	2012	Chronic Kidney disease	WEKA	Decision tree	--
Xudong Song et al[11]	2012	Renal failure Hemodialysis	WEKA	Decision tree	60-80%
N. SRIRAAM et al[12]	2005	Kidney Dialysis	-	Association Rule	97.7%
Divya Jain et al [14]	2014	Nephrotic syndrome(total protein)	TANAGRA	C4.5	11% (error rate)
Jicksy Susan Jose et al[13]	2012	Kidney Image	MATLAB	Association Rule	92%
				Navie Bayes	
Koushal Kumar et al[15]	2012	Kidney Stone	WEKA	ANN	MLP 0.9613%
					LVQ 0.8459%
					RBF 0.8732

VI. CONCLUSIONS

The general goal is to contemplate the different data mining strategies accessible to anticipate the Kidney disease and to contrast them with locate the best technique for forecast.

We dissected with a specific end goal to enhance the rough set model; the variable accuracy rough set model was proposed by the presentation of the blunder parameter. It permits a few mistakes in the division procedure, which idealized the estimation space concept, diminished the tree's limbs considerably, and enhanced speculation abilities. In any case, the variable accuracy rough set decision tree development process still has a conspicuous inadequacy during the time spent figuring the unequivocal locale: the more the quantity of traits is the more prominent the estimation of express district. With a specific end goal to take care of this issue, they proposed another weight restrict number unequivocal area estimation technique.

We broke down that the most regularly utilized DM system, for example, Decision Trees, ANN and Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs) coming about too performing on restorative databases. Additionally demonstrates that DTs, ANNs and Naive Bayes are the well-performing algorithms utilized for Kidney disease. Be that as it may, it is extremely hard to name a solitary DM method as the best for the Kidney diseases. Depending on solid circumstances, at some point a few procedures perform superior to others, yet there are situations when a blend of the best properties of a portion of the previously mentioned DM strategies comes about more viable.

We additionally broke down that there is no single classifier which create best outcome for each dataset. The execution of a classifier is assessed utilizing testing data set. In any case, there are likewise issue with testing data set. Some time it is unpredictable and some time it turns out to be anything but difficult to group the testing data set. To maintain a strategic distance from these issues they utilized cross approval strategy with the goal that each record of data set is utilized for both preparing and testing.

We additionally investigated that by utilizing diverse data mining classification systems and devices to anticipate kidney disease as well as the precision of kidney pictures and impact of other disease on kidney.

References

[1]H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.
[2]K.Sudhakar and Dr. M. Manimekalai," Study of Heart Disease Prediction utilizing Data mining", IJARCSSE, Volume 4, Issue 1, January 2014.
[3]J. Han and M. Kamber, "Data mining: concepts and procedures", second Ed. The Morgan Kaufmann Series, 2006.
[4]Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.
[5]http://www.webmd.com/urinary-incontinence-oab.
[6]DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi,"Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-investigation", Open Access Scientific Reports, Volume 1
• Issue 12 • 2012.
[7]J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van cave Berghe, K.U.Leuven," Data mining methods for anticipating intense kidney damage after elective heart surgery", Springer, 2012.
[8]K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna,"Performance examination of three data mining strategies for anticipating kidney disease survivability", International Journal of Advances in Engineering and Technology, Mar. 2014.
[9]Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri," Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients", International diary of doctor's facility investigate, Volume 2, Issue 1,2013, pp 49-54.
[10]Abeer Y. Al-Hyari," CHRONIC KIDNEY DISEASE PREDICTION SYSTEM USING CLASSIFYING DATA MINING TECHNIQUES", library of college of Jordan, 2012.
[11]Xudong Song, Zhanzhi Qiu, Jianwei Mu," Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field", International Journal of Advancements in Computing Technology(IJACT) ,Volume4, Number3, February 2012.
[12]N. SRIRAAM, V. NATASHA and H. KAUR," DATA MINING APPROACHES FOR KIDNEY DIALYSIS TREATMENT" , diary of Mechanics in Medicine and Biology, Volume 06, Issue 02, June 2006.
[13]Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh," An Efficient Diagnosis of Kidney Images utilizing Association Rules", International Journal of Computer Technology and Electronics Engineering (IJCTEE),Volume 2, Issue 2,april 2012.
[14]Divya Jain, Sumanlata Gautam," Predicting the Effect of Diabetes on Kidney utilizing Classification in Tanagra", International Journal of Computer Science and Mobile Computing, Volume 3, Issue 4, April 2014.
[15]Koushal Kumar and Abhishek,"Artificial Neural Networks for Diagnosis of Kidney Stones Disease", I.J. Data Technology and Computer Science, 2012, 7, pp 20-25.