

Analysis of Chronic Kidney Disease Prediction Using Decision Tree and K-Nearest Neighbor Classification

Regin Bose K¹, Bhuvaneshwar N², DineshKumar D³,
Belwin J Brearley⁴

¹Professor, Department of Computer Science and Engineering, Chennai
Institute of Technology, Chennai, Tamil Nadu, India
reginbosek@citchennai.net

²UG Student, Department of Computer Science and Engineering,
Chennai Institute of Technology, Chennai, Tamil Nadu, India
bhuvan08022002@gmail.com

³UG Student, Department of Computer Science and Engineering,
Chennai Institute of Technology, Chennai, Tamil Nadu, India
dineshdb3103@gmail.com

⁴Assistant Professor (SG),
Department of Electrical and Electronics Engineering,
B.S.Abdur Rahman Crescent Institute of Science and Technology,
Chennai, Tamil Nadu, India
belwinbrearley@gmail.com

Abstract

Chronic Kidney Disease (CKD) is a major threat in medical analysis and is one of the major contributors of death as a non-communicable disease, affecting 10 to 15 percent of the worldwide population. Accurate detection of CKD in its initial stages is thought to be critical for minimizing the effects of health complications of the patients such as hypertension, iron deficiency, bone disorders occurring due to imbalance of minerals, malnourishment, pH fluctuations and abnormalities, and neurological complications through timely intervention with appropriate treatments. This study offers the steps for predicting status of CKD using medical records which comprise the data preprocessing, managing missing values, and feature extraction. Several studies on the detection of CKD at an early stage have been conducted using machine learning techniques. The two classification models such as Decision Tree (DT) and K-Nearest Neighbor (KNN) is used in this study. The performance of each classification methods is compared with each other and identified that KNN is the best classifier. Also using the relevant data extracted from the clinical dataset after feature

extraction, the patient is said to have CKD or not CKD based on KNN approach with accuracy of 97 percent.

Keywords: Chronic kidney disease, prediction, machine learning, K-Nearest Neighbor, Decision Tree, accuracy, recall, f1-score.

Introduction

The kidneys play the vital role in essential function of the body. If the kidneys of a person are damaged and if he has CKD, then its impossible to keep him healthy by filtering his blood. Typically, the symptoms of CKD do not show up in its nascent stages. However, as the condition degrades, the blood of the patient is contaminated with bodily wastes, causing symptoms that are identical to that of a flu. High blood pressure, anemia, brittle bones, poor dietary habits, and neurological damage are all possibilities of the kidney disease. Kidney disease elevates the chances of developing cardiovascular diseases because kidneys are required for many bodily processes. Even though these problems may appear slowly with gradual increase in intensity and with zero symptoms, they can eventually lead to renal failure, which can occur suddenly. When the kidneys fail, dialysis or a kidney transplant is required. The primary causes of the kidney disease are diabetics and high blood pressure. Between 2015 and 2017, 76 percent of kidney failure cases had one of these two illnesses as the major diagnosis: 29 percent of new KFRT (Kidney Failure with replacement therapy) patients had hypertension as the primary diagnosis, while 47% of new KFRT patients had diabetes as the primary diagnosis, which is the major cause of KFRT. Kidney disease affects one in three adults in the United States. Certain demographic groupings are vulnerable. The risk factors of kidney disease are diabetes, being obese, high blood pressure, having heart disease, being over age of 60, family history of renal failure and having previously suffered from any kidney injury [1].

In the study of [2] found that 387.5 million people in underprivileged and bourgeois countries had CKD, compared to 110 million in affluent countries (men of about 48.3 million and women of about 61.7 million). Bangladesh, a heavily populated developing nation in Southeast Asia in which CKD rates continue to hike. A global survey of six countries shows that the overall prevalence of CKD is around 14 percent [3]. Patients with End-Stage Renal Disease (ESRD), which necessitates complex medical procedures like dialysis and kidney transplantation, are more likely to develop CKD [4], and that financial burden results in ongoing medical and psychological issues [5,6].

The American Recovery and Reinvestment Act (ARRA) [7], which was implemented in 2009, had passed the indication that paper medical records should be replaced by electronic ones in order to make it unchallenging for Registered Medical Practitioners to access patient

records and this required health professional across the US to revamp their internal record systems by establishing the centralized database. It was a difficult step to implement ARRA. Medical professionals have kept physical records of patient information for decades. Now the entire process had to be reversed, requiring expenditures in software and career upskilling.

This study aims to compare several approaches of intelligent ML-based to predict renal disease. The accuracy rate for majority of studies was about 90 percent, which was regarded as excellent. The uniqueness of this study is that a variety of methods are employed to increase the accuracy from earlier papers to 97 percent. With F1-scores of 97 percent approximately, the decision tree and K-Nearest Neighbor (KNN) algorithms exhibited the best performance.

Literature Survey

In 2015 a project is developed to diagnose and predict system based on predictive mining which was proposed by P. Swathi Baby [8]. Here, the kidney disease dataset is analyzed by using Weka and orange software. In this research, machine learning techniques such as AD (Active Directory) Trees, J48 (Java 48), K-STAR (Korean Superconducting Tokamak Advanced Research), Naive Bayes, and Random Forest are utilized to assess the performance of each algorithm in terms of statistical analysis and renal disease prediction. Their research concluded that K-Star and Random Forest are the most accurate models for the input dataset and development time (i.e.) less than 0.6 second, and ROC (Receiver Operating Characteristics) values are 1.

In order to forecast kidney dialysis survivorship, K.R. Lakshmi offered performance evaluation of three data mining algorithms in 2014 [9]. The three different data mining algorithms are Artificial Neural Network (ANN), Decision Tree (DT) and Logistic Regression (LR). The data gathered from several dialysis centers has been used to put the principles in this study to the test. After the performance evaluation this research had concluded that ANN is better for kidney dialysis to improve accuracy and performance.

Research on predicting kidney dialysis patient's survival using data mining approaches was presented by Shital Shah [10]. Data mining is utilized in this study to extract information about how these variables interact with patient survival. Two data mining methods which are responsible for information extraction in the form of decision rules are RS theory (Stimulus-Response theory) and DT algorithm. The "most invariant" patient's individual visits are mined for data as they create "signatures" for their respective categories. It is concluded that overall classification accuracy of these data techniques is improved by employing individual visit dataset rather than aggregate dataset. When compared to rule sets

that are aggregate based, individual visit-based rule sets had higher accuracy of forecasting.

Mr. S Dayanand proposed a project in 2015 to use Support Vector Machines (SVM) and Artificial Neural Network (ANN) to estimate the kidney disorders [11]. The accuracy and the execution times are compared for both algorithms in order to determine their functionality. According to his experimental findings, the ANN performs better than SVM.

With some restrictions, the current chronic renal disease prediction system works well. The need for a new CKD prediction system remains. Given the paucity of research in this area, a computerized clinical decision support system for chronic renal disease is a priority for early research.

Proposed System

This study which focuses on the prediction of chronic kidney disease in humans employs two classification models, namely DT and KNN. Each classifier used the dataset to predict the kidney disease, and its performance was assessed objectively using the accuracy, degree of precision, and F-measure.

3.1 System Architecture

Figure 1. Architecture of Proposed System

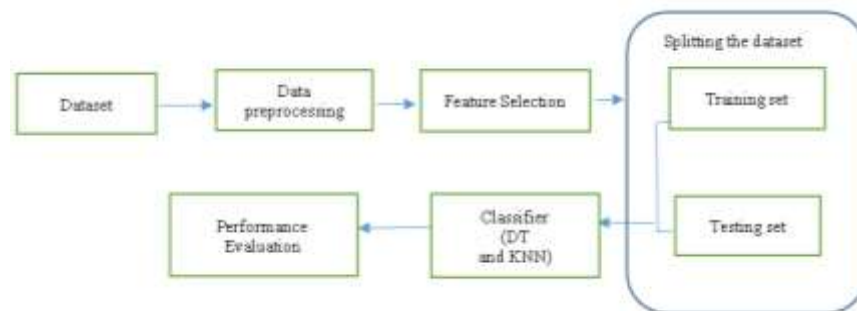


Figure 1. portrays the steps of the proposed systems to be followed. Each step and its functionality are detailed as follows:

3.1.1 Dataset

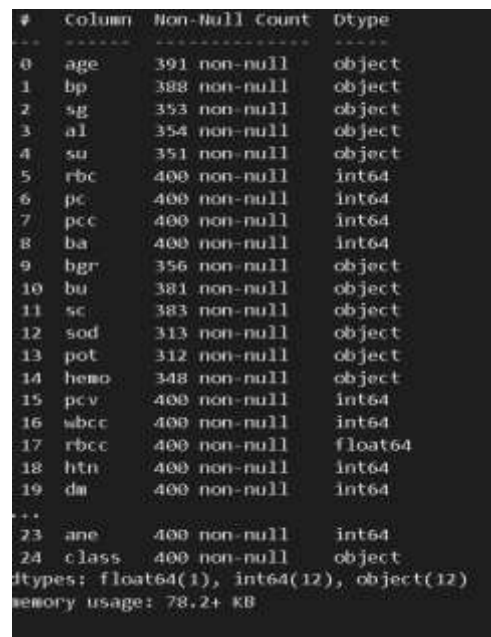
The CKD dataset [12] was used for the investigation. This dataset has 14 columns and 400 rows. There are two possible values for the output column class: "1" and "0." The patient who has the CKD is indicated by the value 1 and the patient who does not have the CKD is indicated by value 0. The dataset is compiled from various hospitals, clinics, and medical research facilities. A mock Kidney Function Test (KFT) dataset has been created from this information to investigate kidney disease. In this side-by-side analysis, a dataset with four hundred occurrences and twenty-five attributes is employed. This KFT dataset includes some of the following

attributes: age, blood pressure, albumin, sugar, pus cells, blood urea, serum creatinine, sodium, potassium, hemoglobin, red and white blood cell count, hypertension, and diabetes mellitus.

3.1.2 Data Preprocessing

Data preprocessing is used to eliminate undesired noise and outliers before model development that could model divergence from the intended training set. The effectiveness of the model is addressed at this step. Data must first be gathered, cleansed, and made available for model creation. Next, the dataset is checked for any null/void values (Figure2).

Figure 2. Absence of Missing Values



#	Column	Non-Null Count	Dtype
0	age	391 non-null	object
1	bp	388 non-null	object
2	sg	353 non-null	object
3	al	354 non-null	object
4	su	351 non-null	object
5	rbc	400 non-null	int64
6	pc	400 non-null	int64
7	pcc	400 non-null	int64
8	ba	400 non-null	int64
9	bgr	356 non-null	object
10	bu	381 non-null	object
11	sc	383 non-null	object
12	sod	313 non-null	object
13	pot	312 non-null	object
14	hemo	348 non-null	object
15	pcv	400 non-null	int64
16	wbcc	400 non-null	int64
17	rbcc	400 non-null	float64
18	htn	400 non-null	int64
19	dm	400 non-null	int64
20
23	ane	400 non-null	int64
24	class	400 non-null	object

dtypes: float64(1), int64(12), object(12)
memory usage: 78.2+ KB

There is a 70/30 split between the training and testing datasets, which improves this research accuracy and effectiveness. The split model is then trained using a variety of classification methods. The DT classification and KNN was two of the classification techniques employed in this study.

3.1.3 Feature Selection

According to Genari[13] "Features are relevant if their values vary systematically with category membership". Correlation-based filter selector is used for this feature selection (Figure 3). Building the classification model for certain job after finding the representative set of characteristics is the key challenge in ML. This paper implements correlation to address the issue of feature selection for ML. The main premise is that effective feature sets consist of features that incorporate a strong correlation to the class but no correlation to one another. A feature assessment formula that is based on ideas from test theory offers

an operational description of this premise. This assessment formula, a heuristic search technique and suitable correlation measure are all combined in the CFS (Correlation-based Feature Selection) algorithm. In tests using synthetic and real-world datasets, CFS was assessed. Here the two machine learning algorithms DT and KNN are employed. Using artificial datasets, experiments shown that CFS rapidly finds and screens irrelevant, redundant, and noisy features, provided that their relevance does not substantially depend on other features. [14].

Figure 3. Filter Feature Selector

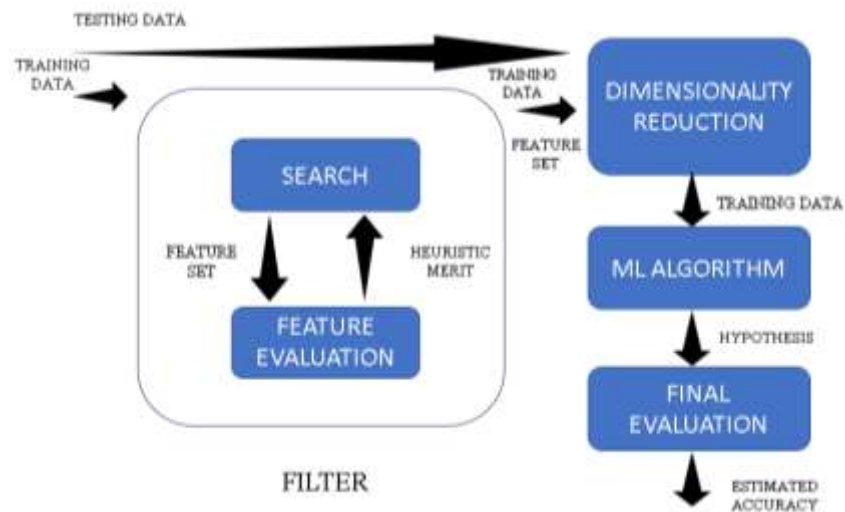


Figure 4. Wrapper Feature Selector

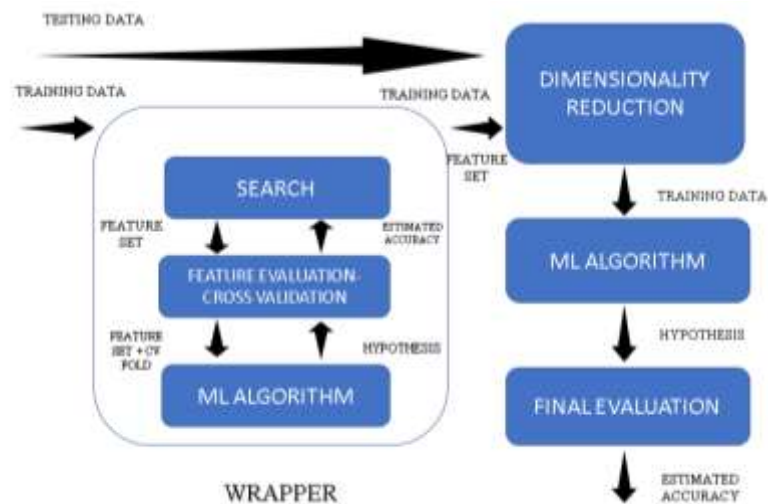


Table. 1 Feature Correlation Values

	age	bp	sg	al	su	rbc	pc	pcr	ba	bgr	...	hemo	pcv	wbcc	rbcc	
132	-0.074372	1.495105	-1.013883	0.755031	1.457545	-0.364890	-0.484322	0.360477	0.263664	0.901042	...	-1.283127	-0.471178	1.585596	-0.197080	-1.287
309	-0.016117	-1.218544	0.407027	-0.788474	-0.438396	-0.364890	-0.484322	0.360477	0.263664	-0.268475	...	1.742018	0.473244	0.428440	1.156887	0.770
341	0.882944	-0.477004	1.403458	-0.788474	-0.438396	-0.364890	-0.484322	0.360477	0.263664	-0.252502	...	0.405681	-0.522270	0.250063	0.848212	0.770
196	-0.132627	1.741618	-1.315838	1.526783	-0.438396	2.740554	2.064742	0.360477	0.263664	-0.268475	...	-1.459043	-0.471178	0.782998	0.141437	-1.287
246	-0.190983	2.401158	-0.409405	1.526783	-0.438396	2.740554	-0.484322	-2.774104	0.263664	-0.568480	...	-1.283127	-0.353125	-0.262772	-0.281709	-1.287
...
216	0.741189	-0.477004	-1.315838	-0.788474	-0.438396	-0.364890	-0.484322	0.360477	0.263664	-0.551868	...	0.194562	0.355191	-1.377631	-1.338575	0.770
259	-0.948199	0.262537	0.407027	-0.788474	-0.438396	-0.364890	-0.484322	0.360477	0.263664	-0.580785	...	1.355604	0.768376	-0.418852	0.860785	0.770
49	0.508179	-0.477004	-1.315838	0.755031	-0.438396	-0.364890	2.064742	-2.774104	0.263664	-0.071889	...	-0.890113	-0.178046	3.438558	0.141437	-1.287
238	1.207240	1.741618	-0.711549	0.240528	0.194918	-0.364890	-0.484322	0.360477	0.263664	0.687538	...	-1.001663	-0.235073	-1.377631	-1.338575	-1.287
343	-0.831689	-1.218544	1.403458	-0.788474	-0.438396	-0.364890	-0.484322	0.360477	0.263664	-0.488978	...	1.390788	1.063508	-0.151298	1.072358	0.770
120 rows * 24 columns																

For subsequent prediction, all positively correlated values are taken for estimation. Every molecule of albumin consists of exactly five different value sets. The albumin level is determined by using a protein test on the urine sample. A high protein content in the urine indicates that illness, fever, or strenuous activity have harmed the kidneys' filtration systems. To establish the diagnosis, numerous tests should be run over a period of several weeks. Blood creatinine, serum creatinine, and creatinine are all used interchangeably. The result of the molecule creatine being broken down in muscles is creatinine. Creatinine is eliminated from the body through the kidneys. The level of creatinine in blood is measured by this test. The metabolism that generates the required energy for muscle contractions includes creatine as a component. Creatine and creatinine production occurs at similar rates. Elevated blood creatinine levels can be the effect of a high-protein diet, congestive heart failure, dehydration, to name a few possible causes (Table.1). Range of male creatinine levels is 0.7 and 1.3 mg/dL, while the range for women is 0.6 and 1.1 mg/dL. Moreover, as the blood exerts increased pressure against the walls of blood arteries, hypertension, also known as high blood pressure, develops. If hypertension is not adequately managed or treated, it can result in heart attacks, strokes, and chronic renal disease. Conversely, hypertension can manifest from CKD [15].

3.4 Algorithms

The CKD can be predicted by using the following ML algorithms:

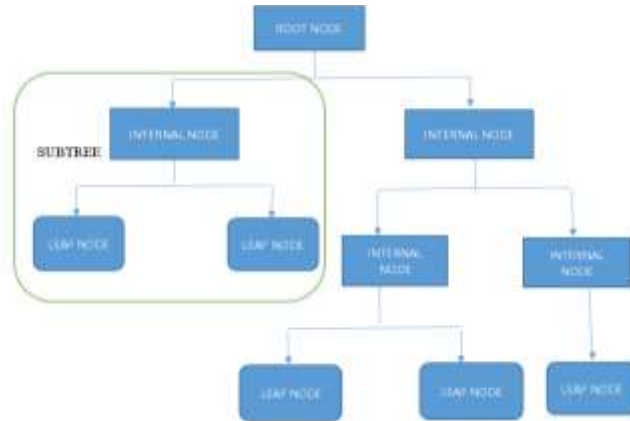
- i) Decision Tree Algorithm
- ii) K-Nearest Neighbor

3.4.1 Decision Tree Algorithm

Both discrete and continuous attributes may be forecasted using the Decision Tree model, which combines classification and regression. This method forecasts discrete attributes depending on the connections between input columns in a dataset. By using the values of those columns, which are referred to as states, it forecasts column states that can be

identified as predictable. The approach finds the input columns connected to the predicted column.

Figure 5. DT Block Diagram

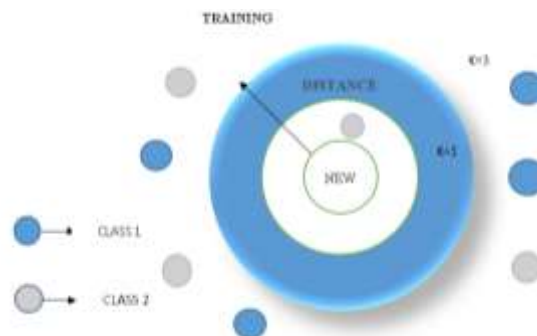


Since it mimics the steps, a person takes when making a real-life decision, the decision tree is simple to understand. Dealing with challenges involving decision-making may make use of it quite a bit. It is a good idea to think about all options for resolving a problem. In decision tree classification the data cleaning is not important.

3.4.2 K-Nearest Neighbor (KNN)

KNN is the supervised learning methodology, which is a highly simple machine learning algorithm. It is categorized accordingly depending on how much a new instance resembles previous categories, it is. By using the KNN approach, all your data can be stored, and new data may be categorized according to its resemblance to the existing one. KNN approach can categorize new data into predefined categories quickly. The KNN approach can be applied to regression even though it is frequently employed for classification problems.

Figure 6. Working procedure of KNN



The K-nearest neighbor is a popular classification model. To categorize data, K-nearest neighbor can be used which is the non-parametric slow

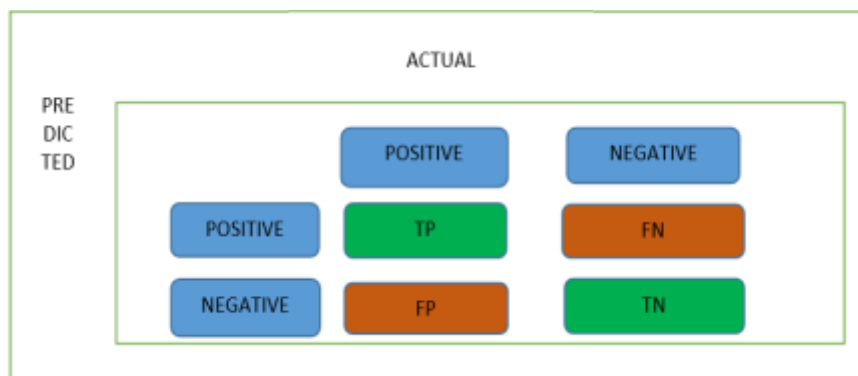
learning approach. This classifier sorts objects based on their distance and proximity to one another. It gives importance to both the proximity of the object and the delivery of important information.

3.5 Confusion Matrix

Confusion matrix is used to evaluate the effectiveness of machine learning categorization models. This matrix was used to assess each model. The confusion matrix shows how frequently our models make accurate and unreliable guesses. False positives and negatives were associated with poorly predicted values, whereas accurate predictions were associated with true positives and negatives.

- i) True Positive (TP): This refers to a result where the ML model properly predicted that the outcome belongs to the positive class.
- ii) True Negative (TN): The model is deemed to have correctly identified the negative class in this case.
- iii) False Positive (FP): This is the result that occurs when the model fails to accurately identify the positive class.
- iv) False Negative (FN): The model is considered to have incorrectly classified the negative class in this case.

Figure 7. Block diagram of confusion matrix



Result Analysis

4.1 Decision tree (DT) classifier

The decision tree algorithm is discussed earlier in section 3.4.1. In this research the accuracy of the DT algorithm is 96 percent. The precision of both CKD and non-CKD patients using DT algorithm is 96 percent respectively, whereas the F1-score values of CKD and non-CKD patients using DT algorithm is 95 and 97 percent respectively.

4.2 K-Nearest Neighbor (KNN)

The KNN algorithm is discussed earlier in section 3.4.2. In this research the accuracy of the KNN algorithm is 97 percent. The precision of CKD and

non-CKD patients are 98 and 97 percent respectively, whereas the F1-score values of CKD and non-CKD patients are 98 and 97 percent respectively.

Comparative Analysis:

The two algorithms DT and KNN are subjected to comparison, based on precision, recall, accuracy, and F1-score values respectively.

5.1 Estimation Parameters

5.1.1 *Precision*: The precision is the proportion of relevant records found to all relevant and irrelevant records found. Often, it is shown as a percentage.

Precision= (relevant record retrieval count) / ((Number of relevant records retrieved) + (incorrect record retrieval count))

5.1.2. *Recall*: The recall is calculated as the proportion of relevant records that were successfully retrieved to all relevant entries in the database. Often, it is shown as a percentage.

Recall= (Number of relevant records retrieved) / ((Number of relevant records retrieved) + (Number of relevant records ignored))

5.1.3. *Accuracy*: It predicts the class label correctly and the predictor's accuracy refers to how well a given predictor can guess the value of the predicted attribute of new data.

Accuracy= (TN + TP)/ (FN + FP + TN + TP)

5.1.4. *F1-score*: It is named as the harmonic mean of precision and recall.

F1-Score= (2* precision * recall)/ (precision + recall)

Table 2: Comparison between DT and KNN based on the precision, accuracy, recall and F1-score

TECHNIQUES USED	PRECISION	ACCURACY	RECALL	F1-SCORE
DECISION TREE	0.9574468085106383	0.9583333333333334	0.9375	0.9473684210526315
K-NEAREST NEIGHBOR	0.9787234042553191	0.975	0.9583333333333334	0.968421052631579

From the table 2, the precision and accuracy of the KNN classification Algorithm is greater as compared to Decision Tree classification technique for CKD dataset. From the relevant test data of the dataset (after feature extraction) using KNN algorithm, the patient's records are accessed to distinguish between the patient affected by chronic kidney disease from that of whom does not affect by the chronic kidney disease.

Conclusion and Future Enhancement

According to this research, the KNN algorithm is used to predict the chronic kidney disease more accurately. The precision and accuracy of the KNN classifier model is 98% and 97% respectively. As compared to the previous research, the accuracy percent of the KNN model used in this investigation is substantially higher, indicating the models that are used in this study are highly reliable as compared to the models implemented in the previous research.

Future research must be built on this work by developing the web application for this model incorporating the machine learning algorithms and the biggest dataset used in this study. The model for real time analysis of the kidney disease prediction should be implemented in future to foretell the disease during the time of the consultation using ML algorithms which will be helpful for patients for early diagnosis.

Bibliography

1. <https://www.kidney.org/news/newsroom/fsindex>
2. K. T. Mills, T. Xu, W. Zhang, "A systematic analysis of worldwide population-based data on the global burden of chronic kidney disease in 2010," *Kidney International*, vol. 88, no. 5, pp. 950–957 (2015).
3. B. Ene-Iordache, N. Perico, B. Bikbov, "chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study," *The Lancet Global Health*, vol. 4, no. 5, pp. e307–e319 (2016).
4. M. J. Lysaght, "Maintenance dialysis population dynamics: current trends and long-term implications," *Journal American Society Nephrology*, vol. 13, suppl 1, pp. S37–S40 (2002).
5. M. Bakhshayeshkaram, J. Roozbeh, S. T. Heydari, "A population-based study on the prevalence and risk factors of chronic kidney disease in adult population of shiraz, southern Iran," *Galen Medical Journal*, vol. 8, no. 935, p. 935 (2019).
6. K. U. Eckardt, J. Coresh, O. Devuyst, "Evolving importance of kidney disease: from subspecialty to global health burden," *The Lancet*, vol. 382, no. 9887, pp. 158–169 (2013).
7. <https://www.truenorthitg.com/pros-and-cons-paper-medical-records>
8. P. Swathi Baby, T. Panduranga Vital, "Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms" *International Journal of Engineering Research & Technology (IJERT)* ISSN: 2278-018, Vol. 4 Issue 07, pp 206-210, (2015).
9. K.R.Lakshmi¹, Y.Nagesh² and M.VeeraKrishna³, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", *International Journal of Advances in Engineering & Technology*, Vol. 7, Issue 1, pp. 242-254 (2014).
10. Andrew Kusiak, Bradley Dixon^b, Shital Shaha, "Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, *Computers in Biology and Medicine* 35, pp 311–327 (2005).
11. Dr. S. Vijayarani, Mr.S.Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms" *IJCBB*, ISSN (online): 2229-6166, Volume Issue 6 (2015).

12. Kaggle-> "Chronic Kidney Disease Dataset," <https://www.kaggle.com/abhia1999/chronic-kidney-disease>.
13. Mark A. Hall," Correlation-based Feature Selection for Machine Learning", <https://www.cs.waikato.ac.nz/~mhall/thesis.pdf> (1999)
14. <https://www.semanticscholar.org/paper/Correlation-based-Feature-Selection-for-Machine-Hall/6bc43977fb11cceed0b9aa55b23c6dd29dd9a132>
15. Gazi Mohammed Ifraz, Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis, and Mohammad Monirujjaman Khan," Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods", Volume 2021 | ArticleID 6141470 | <https://doi.org/10.1155/2021/6141470> | (2021)
16. Taiwo Oladipupo Ayodele, "Types of Machine Learning Algorithms", New Advances in Machine Learning, Yagang Zhang (Ed.), InTech. (2010)
17. A. Asuncion and D. J. Newman. UCI Machine Learning Repository. Available: <http://www.ics.uci.edu/~mllearn/MLRepository.html> (2007)