Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Predicting Chronic Kidney Disease Using Machine Learning Approach

B. Meena, S. Narasimha Rao, T. Purna Chandra Sai, S.P. Sriranganath Department of Computer Science Engineering Raghu Engineering College-Visakhapatnam-531162
meenacserec@gmail.com[1], 19981a05f4@raghuenggcollege.in[2], 19981a05g2@raghuenggcollege.in[3], 20985a0520@raghuenggcollege.in[4]

Abstract:

A major issue, chronic kidney disease has been growing at a constant rate. An individual may only survive for a few days without kidneys which leads to dialysis and kidney transplantation. The kidneys are harmed in chronic kidney disease (CKD) and can't cleanse blood as they normally do. Heart conditions, anemia, bone conditions, excessive potassium and calcium levels, and anemia are among the extremely frequent consequences that arise with kidney failure. The worst-case scenario results in total renal failure, necessitating a kidney transplant in order to survive. The quality of life can be increased more by CKD early identification by using machine learning methods like Random Forest, Naive Bayes, Decision tree, SVM, KNN. Random forest gives more accuracy than remaining models, so we used random forest classifier to predict CKD.

Keywords: chronic kidney disease, renal failure, machine learning, random forest.

I. Introduction

Artificial intelligence known as machine learning allows software applications in predicting the outcomes more accurately without the use of explicit coding. Machine learning algorithms make output value predictions using historical data.

It is a system that offers the user advice and tips for maintaining their health, as well as instructions on how to find out the diseases. In today's world, the health sector is essential for treating people's illnesses. CKD does not cause symptoms early in the disease, so testing is the only way to detect it [1]. The user can understand the ailment simply by entering the symptoms and helpful information, and it is also helpful for the user if he or she does not want to travel to the hospital or other clinics. It is also a result of poor medical infrastructure and a low doctor-to-population ratio. We can predict diseases using Python and Machine Learning and with the help of the dataset that is taken from the hospitals.

Everyone in today's world is occupied with their jobs. When they see any symptoms of a sickness, people start to worry about their health. Globally, chronic kidney disease (CKD), which can lead to cardiovascular disease, renal failure and early death, is a serious public health concern. Early kidney impairment diagnosis may aid in correction, which is not always possible. CKD can't be predicted because it doesn't show any symptoms which might leads to complete health damage however machine learning algorithms helps us in prediction, analyzing data and earlier treatment of the diseases. There is a high probability that the disease will be detected when it is in its final stage, which can lead to kidney failure in some cases [8].

II. DATASET DESCRIPTION

The dataset description is given below.

- A dataset is a collection of structured or unstructured data that is organized and presented in a specific format for analysis or processing. It can be in the form of a spreadsheet, a database, a CSV file, a text file, or any other type of file that contains information.
- Datasets serve multiple purposes, ranging from training machine learning models and carrying out statistical analysis, to offering valuable insights into the underlying patterns and trends present within the data. They can range in size from small to large, depending on the amount of data collected and the purpose for which it is being used.
- In general, a good dataset should be well-defined and organized, with clear documentation on the format, structure, and meaning of the data. It should also be representative of the population or phenomenon being studied and be largeenough to provide reliable and meaningful results.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Attributes	Acronym	Type
Age	age	Numerical (years)
Blood pressure	bp	Numerical (mm/Hg)
Specific gravity	sg	Nominal(1.005,1.010,1.015,1.02,1.025)
Albumin	al	Nominal(0,1,2,3,4,5)
Sugar	su	Nominal(0,1,2,3,4,5)
Red blood cells	rbc	Nominal (normal,abnormal)
Pus cell	pc	Nominal (normal,abnormal)
Pus cell clumps	pcc	Nominal (present,notpresent)
Bacteria	ba	Nominal (present,notpresent)
Blood glucose random	bgr	Numerical (mgs/dl)
Blood urea	bu	Numerical (mgs/dl)
Serum creatinine	sc	Numerical (mgs/dl)
Sodium	sod	Numerical (mEq/L)
P otassium	pot	Numerical (mEq/L)
Haemoglobin	hemo	Numerical (gms)
Packed cell volume	pcv	Numerical
White blood cell count	wc	Numerical (cells/cumm)
Red blood cell count	rc	Numerical (millions/cmm)
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)
Anaemia	ane	Nominal(yes,no)

Figure 1: Dataset

III. LITERATURE REVIEW

The existing technique relies on measuring blood creatinine levels to analyses urine samples. For this, medical techniques including ultrasonography techniques and screening are employed. Patients who have a history of heart disease, hypertension, or kidney disease in any of their family are checked during the screening process. This method involves measuring the albumin-to-creatinine ratio (ACR) in a first-morning urine sample as well as estimating GFR from the serum creatinine level. Urine test and other methods like MRI, CT etc. can also be used to diagnose but, these diseases require DNA samples from the patient and every test should be done thoroughly by qualified doctors. These tests take tedious process and time from the start and the costisn't economic too. ML algorithms are used to forecast renal disease since their invention.

A. KNN

- Assuming that the new case and the prior case are comparable, the K-Nearest Neighbor technique assigns the new instance to the category that is closest to the pre-existing categories in order to reduce the similarity between the two categories.
- By simply preserving the dataset during the training phase, the KNN algorithm categorizes fresh data into a category that is relatively close to the training data which is helpful to train the data in an easy way.
- KNN is a machine learning technique that can assist in predicting and diagnosing chronic kidney disease (CKD). KNN is a non-parametric and lazy learning method that doesn't rely on any assumptions regarding the underlying data distribution and only trains when given a new data point. In KNNClassification, the output is a class membership [4].
- In the context of CKD prediction, KNN can be used to classify patients into different stages of the disease based ontheir clinical and laboratory data. For example, if the K nearest neighbors to a new patient have CKD stage 3, then the algorithm would predict that the new patient also has CKD stage 3.
- KNN has been used in several studies on CKD prediction, including:
- "Chronic Kidney Disease Prediction Using Machine Learning Techniques: A Review," by Y. H. Low, C. M. Tan, C.
- H. Wong, and J. J. Wong, which discusses the use of KNN along with other machine learning techniques in CKDprediction.

B. Naive Bayes

- To forecast a class of datasets, Bayes is one of the quick and simple Classification methods in ML. Both binary andmulti-class classifications can be done using it.

- Naive Bayes is a probabilistic machine learning method that can aid in predicting and diagnosing chronic kidney disease (CKD). It employs Bayes' theorem to calculate the likelihood of a patient having CKD based on their clinical and laboratory data, and despite its simplicity, it is an effective algorithm. Naive Bayes algorithm uses the conceptof Maximum Likelihood for prediction [9].

- In the context of CKD prediction, Naive Bayes can be used to classify patients into different stages of the disease based on their data. The algorithm works by calculating the conditional probability of each feature given each class and then multiplying them together to get the probability of a patient belonging to a certain class. For example, if a patient has high levels of creatinine, low levels of hemoglobin, and is over 60 years old, Naive Bayes would calculate the probability of the patient having CKD stage 3 based on the conditional probability of each feature givenCKD stage 3.
- Naive Bayes has been used in several studies on CKD prediction, including:

- "A Machine Learning Approach for Predicting Chronic Kidney Disease," by S. Sultana and M. S. Allam, which usesNaive Bayes along with other machine learning algorithms to predict CKD.

C. Support vector Machine

- SVM is a supervised machine learning technique that is used to address classification or regression issues. SVM provides extremely precise results and is mostly used in machine learning and deep learning approaches to train and test the data. SVM vary from conventional classification methods because they choose a distance measure that minimizes the separation between the nearest sample points for all categories. Decision boundary produced by the SVM is referred to as the largest margin. Decision boundaries in the form of hyperplanes are utilized to separate data points, with support vectors serving as the determining factors for the position and orientation of these hyperplanes due to their proximity to them.[2].
- Support Vector Machines (SVMs) are a widely used machine learning technique that can aid in predicting and diagnosing chronic kidney disease (CKD). They are a robust and versatile algorithm capable of handling both classification and regression tasks. For the organization of linear and nonlinear data a new technique is introduced and need as SVM [1].
- In the context of CKD prediction, SVMs can be used to classify patients into different stages of the disease based on their clinical and laboratory data. The algorithm works by finding the hyperplane that maximally separates the data points into different classes. For example, if we have data points representing patients with CKD stage 3 and patientswithout CKD stage 3 SVMs would find the hyperplane that best separates these two classes.
- SVMs have been used in several studies on CKD prediction, including:
- "Prediction of Chronic Kidney Disease Using Support Vector Machine and Random Forest Models," by Z. Xia, X. Wang, and J. Zheng, which uses SVMs along with Random Forest models to predict CKD progression.

D. Decision Tree

- Decision Tree is a versatile algorithm capable of handling both classification and regression tasks. It uses a tree-like structure where internal nodes represent dataset features, branches correspond to decision rules, and leaf nodes represent the output.
- In the context of CKD prediction, Decision Trees can be used to classify patients into different stages of the disease based on their clinical and laboratory data. The algorithm works by recursively partitioning the data based on the values of different features until a stopping criterion is met. For example, if we have data points representing patientswith different stages of CKD, Decision Trees would partition the data based on different features such as creatinine level, age, and gender, until it can accurately classify patients into different stages of CKD.
- Decision Trees have been used in several studies on CKD prediction, including:
- "Prediction of Chronic Kidney Disease Using Decision Tree Algorithm," by A. Al-Emran, which uses DecisionTrees to predict CKD progression.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

IV. METHODOLOGY

Aim of the proposed model is to predict whether the patient will suffer chronic kidney disease in the future if they continue their lifestyle. In proposed system, first we have taken the data set and then removing anomalies, handling the missing values and selecting the most important attributes from the dataset and then training our model on it. we used random forest because it gives more accuracy when compared to the existing models. Random forest gives the prediction depending on the majority of the votes cast by the trees.

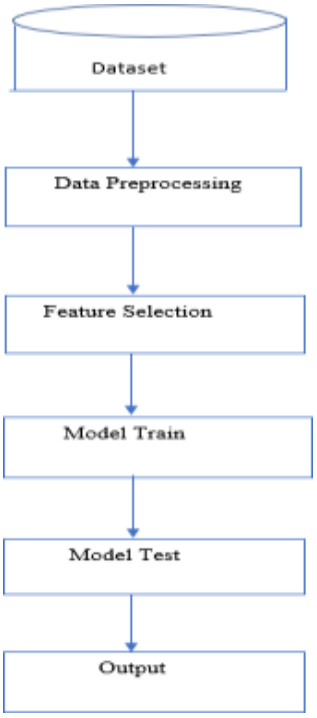


Figure 2: Model flowchart.

A. **Dataset:** Collect and clean the data from various sources. The data should include demographics, medical history,and laboratory test results.

B. **Data Pre-processing:** Pre-process the data to ensure its quality and to prepare it for use in machine learning algorithms. This approach involves several actions, such as managing missing values, scaling the dataset, converting it to binary data, and standardizing it. This includes handling missing data, normalizing, and scaling the data,and encoding categorical variables.

C. **Feature Selection:** The process of Feature Selection involves the computational identification of the most significant features that impact the prediction variable or output. Identify the most important features that will be used in the predictionmodel.

D. **Model Selection**: Choosing an appropriate machine learning technique for a particular task entail evaluating various options, including logistic regression, decision trees, and support vector machines.

E. **Model Training:** Train the selected model on the pre-processed data, using the identified features as inputs and theCKD status as the output.

F. **Model Testing:** Assessing the model's effectiveness on a test dataset involves utilizing metrics such as accuracy, precision, recall, and F1 score.

Random Forest is a classifier which averages input data to increase the projected accuracy of the dataset. Random forest selects the decision trees randomly to increases the accuracy of the input dataset. Instead of using the predictions from just one tree, the random forest forecasts the outcome using the votes of the majority of projections.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

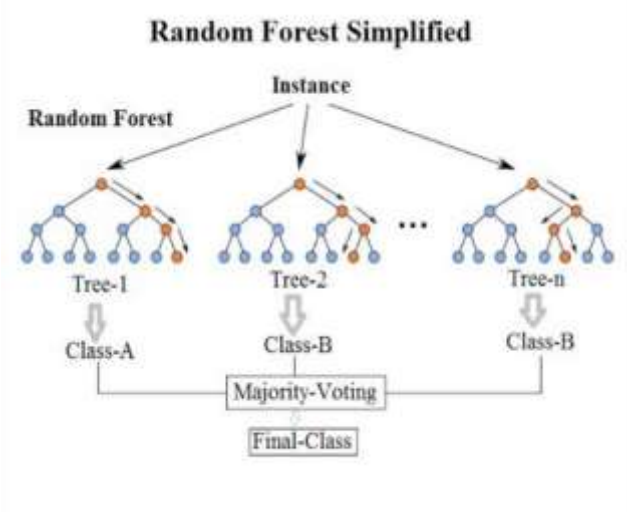


Figure 3: Random Forest Mechanism

- Random Forest is a commonly used machine learning method that can aid in predicting and diagnosing chronic kidney disease (CKD). By employing an ensemble-based approach that integrates multiple decision trees, the Random Forest algorithm improves the model's predictive accuracy and resilience.
- When predicting CKD, Random Forest can be utilized to categorize patients into different stages of the disease using their clinical and laboratory data. This method involves creating several decision trees on various subsets of the data and merging their predictions to generate the final result. By using decision trees on subsets of the data, the algorithm minimizes overfitting and enhances the model's ability to generalize.

In our proposed system we are used special attributes are:

1. **Pus cell:** In the context of chronic kidney disease (CKD), the presence of pus cells in the urine can indicate inflammation and infection of the kidneys or urinary tract. Pus cells are typically not present in healthy urine, and their presence can be a sign of an underlying medical condition. In CKD patients, the presence of pus cells in the urine can be an important diagnostic marker and can help guide treatment decisions. Urine tests can be used to detect the presence of pus cells, and the number of pus cells can be quantified to provide a measure of the severity ofthe inflammation or infection.
2. **pus cell clumps:** Pus cell clumps, also known as WBC casts, are another important indicator of kidney disease in thecontext of chronic kidney disease (CKD). WBC casts are formed when pus cells aggregate and become trapped in the kidney tubules, and their presence in the urine can indicate inflammation and damage to the kidneys.

In CKD patients, the presence of WBC casts in the urine can be a more specific indicator of kidney inflammationthan the presence of individual pus cells. However, the presence of WBC casts can also indicate other underlyingmedical conditions, such as glomerulonephritis or interstitial nephritis.

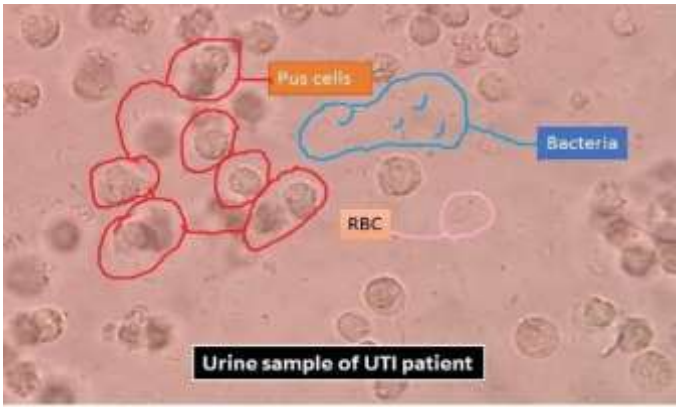


Figure 4: pus cell in urine sample.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

V. EXPERIMENTAL RESULTS

Classifier	Accuracy	precision	Recall	F1 Measure	TPR	TNR	FNR	FPR	Mean	Variance
KNN	**98**	**1.00**	**0.9**	**0.98**	**1.0**	**0.05**	**0.63**	**0**	**0.9**	**0.02**
Naïve Bayes	**97**	**1.00**	**0.9**	**0.97**	**1.0**	**0.07**	**0.64**	**0**	**0.2**	**0.02**
SVM	**96**	**0.96**	**0.6**	**0.75**	**0.9**	**0.05**	**0.58**	**0.05**	**0.9**	**0.01**
Decision Tree	**97**	**1.00**	**0.9**	**0.97**	**1.0**	**0.07**	**0.64**	**0**	**0.9**	**0.02**
Random Forest	**100**	**1.00**	**1.0**	**1.00**	**1.00**	**0.00**	**0.61**	**0**	**0.9**	**0.08**

The above experimental values are calculated by using confusion matrix.

In confusion matrix, 2 things must be noted. They are actual values and predicted values. Actual values are those that are presentin our dataset. Predicted values refers to the values that are predicted by our model.

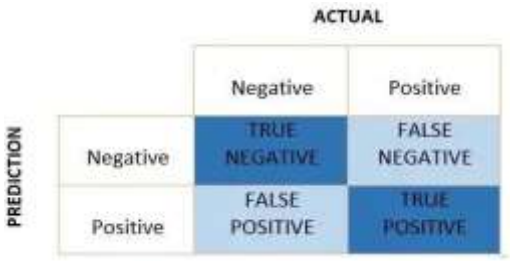


Figure 5: Confusion Matrix

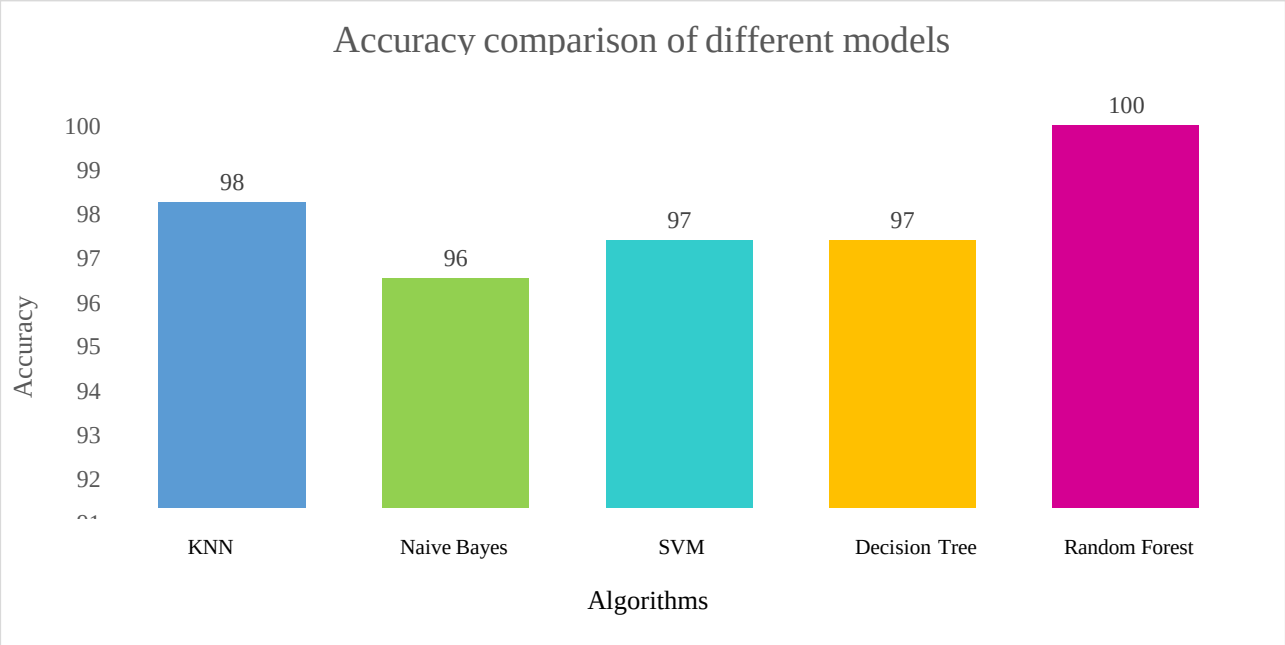
There are 4 components in confusion matrix.

- True Negative (TN): Both the actual value and the expected value are negative.
- True Positive (TP): The expected and actual values are both true.
- False Negative (FN): actual value is yes but the expected value is no.
- False positive (FP): actual value is no, but the expected value is yes.

By using the above components, we can derive Precision, Recall, F1 score and Accuracy.

- **Accuracy:** the proportion of correct predictions made by the model, calculated as **(TP + TN) / (TP + TN + FP + FN).**
- **Precision:** the proportion of true positive predictions among the instances predicted as positive, calculated as **TP / (TP + FP).**
- **Recall** (also known as sensitivity): the proportion of true positive predictions among the actual positive instances,calculated as **TP / (TP + FN).**
- **F1 score:** the harmonic means of precision and recall, calculated as **2 * (precision * recall) / (precision + recall).**

The below figure shows graphical representation of accuracy comparison of different models:



VI. CONCLUSION

The prevention and slowing of the progression of chronic renal disease to kidney failure depends heavily on early prediction, which is important for both patients and doctors. The 16 top attributes are selected for prediction out of the 25 available attributes. We developed a model to forecast the development of CKD using Random Forest. First step was to
apply the five machine learning algorithms to original datasets that included all the features and then train our models on the dataset. Random Forest gives the highest accuracy (100%) among the models, so it gives the prediction more accurately.

VII. REFERENCES

[1] https://www.ijert.org/comparative-study-of-chronic-kidney-disease-prediction-using-knn-and-svm

[2] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00657-5#:~:text=Yashfi%20%5B14%5D%20proposed%20to%20predict,the%20highest%20accuracy%20of%2097.12%25.

[3] https://www.ijert.org/research/chronic-kidney-disease-prediction-using-machine-learning-IJERTV9IS070092.pdf

[4] https://www.ijert.org/research/comparative-study-of-chronic-kidney-disease-prediction-using-knn-and-svm-IJERTV4IS120622.pdf

[5] https://ieeexplore.ieee.org/document/9333572

[6] https://www.sciencedirect.com/science/article/pii/S2153353923000032

[7] https://ijarcce.com/wp-content/uploads/2015/02/IJARCCE3L.pdf

[8] https://www.ijert.org/chronic-kidney-disease-prediction-using-machine-learning

[9] https://www.ijamtes.org/gallery/41.%20may%20ijmte%20-%20490.pdf

[10] https://www.mdpi.com/2504-2289/6/3/98