International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Chronic Kidney Disease Prediction using Random Forest and Suggesting Suitable Diet Plan

Dr.D.Rathna Kishore[1], M.Keerthi Reddy[2], V.Pavani[3], P.Madhumitha[4]
Associate Professor[1], B.Tech Students[2,3,4]
[*] Department of Information Technology
[**] ANDHRA LOYOLA INSTITUTE OF ENGINEERING & TECHNOLOGY

Abstract- Chronic kidney disease (CKD) is a global health problem with high morbidity, mortality, and other illnesses. Kidney disease affects 1 in 10 people worldwide. As the number of CKD patients increases, effective predictors for early detection of CKD are needed. Therefore, a better diagnosis of chronic kidney disease is needed to prevent continued progression. Machine learning models help clinicians achieve this goal with their fast and accurate prediction capabilities. In this project, we are using Random Forest Algorithm to determine if an individual has been diagnosed with CKD or Not CKD. And Stage of kidney disease and based on the person's potassium level a diet plan is suggested. The CKD dataset was taken from the Kaggle machine learning repository. This system helps predict the early detection of chronic kidney disease.

Keywords- Chronic Kidney Disease, Prediction, Diet Plan, Machine Learning, Random Forest, Classification, Potassium level.

I. INTRODUCTION

The Chronic kidney disease, also known as chronic renal disease or CKD, is a condition characterized by a gradual loss of kidney function over time. Chronic kidney disease refers to conditions that harm your kidneys and impair their ability to keep you healthy by filtering waste from your blood. If your kidney disease worsens, wastes can accumulate to dangerously high levels in your blood, making you sick. Complications such as high blood pressure, anemia (low blood count), weak bones, poor nutritional health, and nerve damage are possible. Kidney disease increases your chances of developing heart and blood vessel disease. These issues may develop gradually over time. Early detection and treatment can often prevent the progression of chronic kidney disease. 37 million American adults have CKD, and millions more are at risk. Early detection can help prevent kidney disease from progressing to kidney failure. Heart disease is the leading cause of death in all CKD patients. Glomerular Filtration Rate (GFR) is the most accurate test for determining a person's level of kidney function and the stage of chronic kidney disease. It is calculated using the patient's blood creatinine, age, race, gender, and other information. CKD has affected the entire world, but it has been especially devastating in low and middle-income countries. Chronic kidney disease affects approximately 13.4% of the global population, with the number of deaths increasing dramatically each year. According to studies and research conducted by various health organizations over the last two decades, CKD is said to have caused a large number of deaths and other severe complications. The number of people who suffer from end-stage renal disease is also becoming high and this is considered the last stage of CKD. So, we are developing a system that can predict kidney damage at an early stage. Our focus is on predicting Chronic Kidney Disease using algorithms. A large amount of data is being generated by the medical industry that medical history is being unused. So, it is required to process those unused data. It comes up with the set of techniques that when applied to this processed data, generates reports for making the appropriate decisions and making people aware that kidney failure can be predicted at early stage. The objective of this paper is to predict the CKD at the early stage and provides diet plan as well as provide immediate reports to the patients. Instead of waiting for the report to get ready for more than days the system will help to provide the immediate report on that day itself. Hence this will save the time for patients so that at an early stage they get to know about any kidney disease from which they are suffering.

II. LITERATURE REVIEW

In this paper, CKD is being predicted using Boosting classifiers like AdaBoost and rule induction techniques like Ant-Miners. The boosting algorithm is an ensemble machine learning algorithm that converts weak classifiers to strong models in order to improve accuracy. Researches showed that many data mining techniques had been applied for CKD classification. Among those algorithms, AdaBoost classifier and J48 rule induction method performed well.[1].

In this study, the effects of using clinical features to classify patients with chronic kidney disease by using support vector machine algorithm is investigates. The performance measures of SVM classifier have been evaluated in order to find the best scores for sensitivity, specificity and accuracy metrics [2].

In this paper, presents the importance of detection of CKD. Chronic renal failure is defined as either kidney damage or glomerular filtration rate less than 60ml/min for three months or

more. This is invariably a progressive process that results in end stage renal disease. The importance of detection of CKD and detailed information of CKD is provided with causes and treatment. It is increasingly recognised that the burden of CKD not limited to its implications on demand[5].

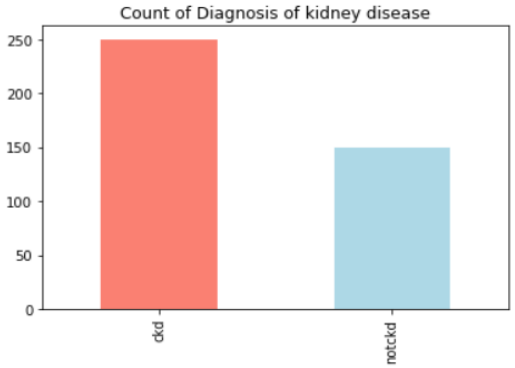
In this study they implemented a prototype in which they applied a machine-learning algorithm for a quick judgment of chronic kidney disorder by applying characteristic choice and aggregate training. In order to enhance the kind of chronic kidney disorder, Correlation-based characteristic Choice was applied for feature collection and Adaptive Boosting was employed for aggregate training. Classifiers such as the KNN algorithm, Bayes, and SVM algorithms were adopted [7].

III. PROPOSED METHOD

A. Dataset Description-

The dataset is taken from the Kaggle machine learning repository. Pre-processing is considered either by replacing or separating from the dataset to eliminate the missing values. RandomForest, a classification algorithm, is used to build a model. The dataset contains 400 illustrations, each of which has 25 features: Blood Pressure, Potassium, Anemia, Albumin, Sugar, Pus Cell, Appetite, Pus Cell clumps, Bacteria, Blood Glucose Random, Hypertension, Specific Gravity, Sodium, Hemoglobin, Age, Packed Cell Volume, Red Blood Cells, White Blood Cell Count, Red Blood Cell Count, Blood Urea, Diabetes Mellitus, Serum Creatinine, Coronary Artery Disease, Pedal Edema and Class [11].The class target variable contains values “CKD” or “NOT CKD”. While “CKD” Chronic Kidney diseases specify positive tests and “NOT CKD” specify a negative test. At hand 250 cases in class “CKD” and 150 cases in class “NOT CKD”.

Distribution of CKD patient- We made a model to predict CKD however the dataset was slightly imbalanced having around 400 classes labeled as 0 means negative means no CKD and 268 labeled as 1 means positive means CKD.



B. Data Preprocessing

Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction.

1)Missing Values removal

When the data obtained is real-world data, missing values will be present. This causes the prediction accuracy to shift even further. Using the mean or average of the observed property or value is an efficient technique to handle missing values. As a result, we have more precise data and better prediction outcomes.

2)Data Transformation

In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1(yes/1 or No/0). Now the resultant csv file comprises of all the integer and decimal values for different CKD related attributes.

3)Splitting of data

So after data cleaning, data is normalized for training and testing the model. When data is split, we train the algorithm on the training data set while keeping the test data set aside. This training process will generate a training model based on logic, algorithms, and feature values in training data. The goal of normalization is to bring all attributes onto the same scale.

C. Apply Machine Learning

When data has been ready, we apply Machine Learning Technique. We have used classification technique Random Forest algorithm to predict Chronic Kidney Disease (CKD).

1)Random Forest

A random forest is a machine learning technique for classifying and predicting outcomes. It takes advantage of ensemble learning, which is a technique for merging several classifiers to solve complex problems. Many decision trees make up a random forest algorithm. The random forest method is trained via bagging or bootstrap aggregation. Bagging is a meta-algorithm that groups machine learning algorithms together to improve accuracy. Based on decision tree predictions, the (random forest) algorithm determines the outcome. By averaging or averaging the output of various trees, its predictions. The precision of the output improves as the number of trees increases. Using a random forest technique, the drawbacks of a decision tree algorithm are avoided. It reduces dataset overfitting and improves precision. It creates forecasts without the need for a huge number of package configurations (like sci-kit-learn).

D. Classifier's Performance Measures

The confusion matrix commonly is used to evaluate the classifier, which measures the quality of the classification process. In addition, there are also various standard evaluation measures for correct and incorrect classification results of the classifier. The most common measure to evaluate the performance is accuracy. It is defined as the proportion of the total number of instances that were correctly classified. Another evaluation metric is sensitivity, is the mean proportion of actual true positives that are correctly identified. On the other hand, specificity is the mean proportion of true negatives which are identified correctly.

E. User Interface:

For the user interface we have used Tkinter which is a Python library for creating desktop application graphical user interfaces (GUIs). It's not difficult to create desktop programmed with Tkinter. Our primary GUI toolkit will be Tk, which is Python's default GUI framework. Tkinter (short for Tk interface), a Python interface, will be used to access Tk.Tkinter is a Python module that makes creating graphical user interfaces very simple. Tkinter is the only graphical user interface framework included with the Python Standard Library. The fundamental benefit of Tkinter is that it is cross-platform, meaning that the same code may run on Windows, Mac OS X, and Linux. Tkinter is a small module that has a lot of power.

F. Diet recommendation module

As dietary management plays an important role in slowing the progression of chronic kidney disease. Patients with diabetes and high blood pressure should follow a strict diet to avoid kidney failure. So, in this module, based on zone detected (using potassium level) a suitable diet plan is suggested. We have three diet plans with low, medium, high potassium diet plans.

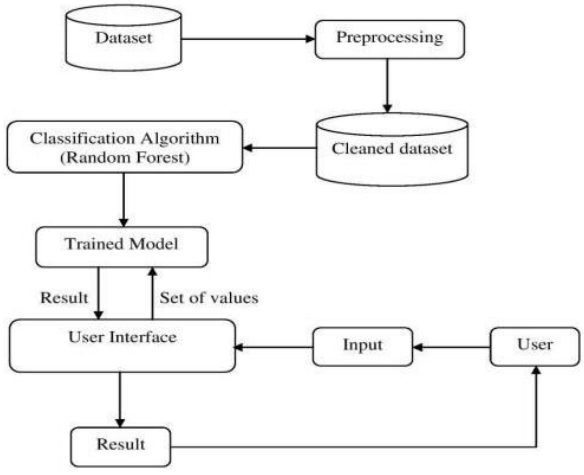


Figure 1. Operational flow for the proposed method.

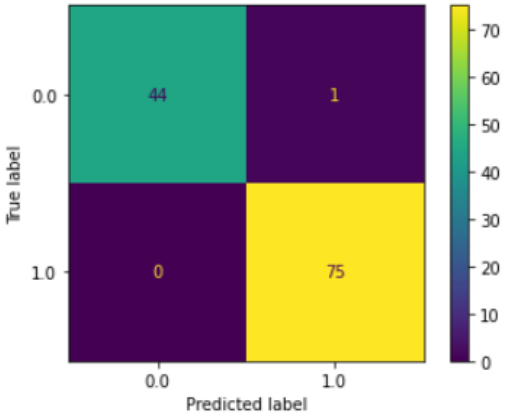
Step1: Import the necessary libraries.
Step2: Load the Chronic kidney disease dataset .
Step3: Preparing data for training i.e data preprocessing.
Fill in missing values with the mean (for numerical values) or mode (for categorical values).
Step4: Next, we will divide the data into train and test split. The following code will split the dataset into 70% training data and 30% of testing data .
Step5: Building the Random Forest Classifier.
To create our random forest classifier and then train it on the train set. We will obtain a trained model.
Step6: We evaluate the performance of the model with the testing data.
Step7: The system will predict the person is with CKD or Not CKD by entering the parameters in GUI Interface.
Step8: Along with the prediction , we will get the stage of the kidney disease based on the GFR(glomerular filtration rate).
Step9: Based on the person's potassium level a diet plan is suggested.

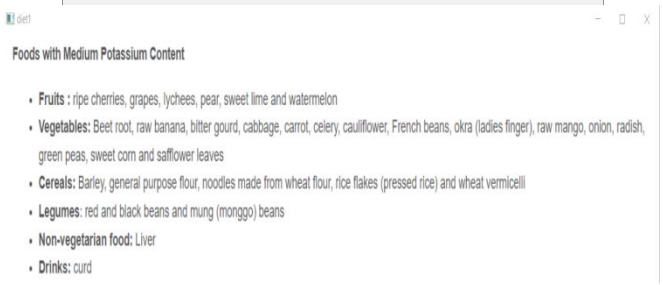
IV. Experimental Results

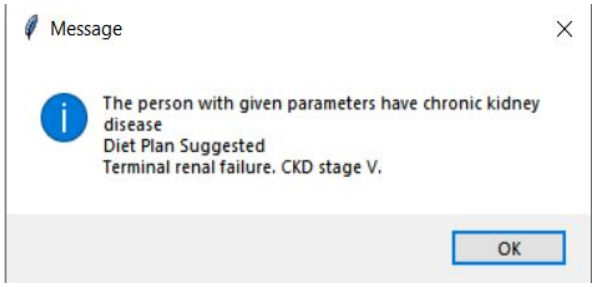
The classification method is implemented in a Jupyter notebook, which produces the results given below. The TP rate refers to True Positives, which are cases that are correctly classified in a dataset, whereas the FP rate refers to False Positives, which are occurrences that are incorrectly classified in a class. Precision denotes the proportion of relevant examples among the retrieved instances (also known as positive predictive value), whereas recall denotes the proportion of resolved appropriate instances over the entire number of appropriate instances (also known as sensitivity).

	precision	recall	f1-score	support
0.0	1.00	0.98	0.99	45
1.0	0.99	1.00	0.99	75
accuracy			0.99	120
macro avg	0.99	0.99	0.99	120
weighted avg	0.99	0.99	0.99	120

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other.The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made. Confusion matrix obtained for the model is







CONCLUSION

Chronic Kidney Infection is one of the health problems which needs better diagnosis. Prognostication of this disease in the early stages may stop the progression of this disease. Therefore, our system aims to predict this at an early stage. Prediction is done using the machine learning technique, random forest algorithm. The models obtained from CKD patients are trained and authenticated with the mentioned input parameters. And 99% classification accuracy has been achieved. The main aim of the machine learning module is to label the patient's CKD status and categories them into various stages. The Diet recommendation module is purely based on blood potassium levels. Overall this system detects and suggests a diet which will be useful to the doctors as well as patients. In addition, to help minimize the incidence of CKD, it has been attempted to predict if a person with this syndrome's chances of chronic risk factors such as hypertension, family history of kidney failure and diabetes using the appropriate dataset.

REFERENCES

[1]. Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6.

[2]. Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-4, doi: 10.1109/ICAICT.2018.8747140.

[3] Y. Amirgaliyev, S. Shamiluulu and A. Serek, "Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-4, doi: 10.1109/ICAICT.2018.8747140.

[4] G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms inhadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatre, 2017, pp. 973-979.

[5] Chronic kidney disease: a research and public health priority; Norberto Perico and Giuseppe Remuzzi1 ,Advance Access pub- lication 3 July 2012

[6] Lemons, K., 2020. A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer. International Journal of Undergraduate Research and Creative Activities, 12(1), pp.1–5. DOI: http://doi.org/10.7710/2168-0620.0287..

[7]. M. S. Wibawa, I. M. D. Maysanjaya, I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnosis," in the 2017 5thInternational Conference on Cyber and ITService Management (CITSM).

[8] Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques ; Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee ,The 2016 Management and In- novation Technology International Conference (MITiCON-2016)

, May 2016

[9] M.A. Ameta and M.K. Jain, "Data Mining Techniques for the Prediction of Kidney Diseases and Treatment: A Review."

[10]www.vikaspedia.in/health/diseases/kidney-related/diet-in-kidney-diseases/diet-in-chronic-kidney-disease

[11] www.kaggle.com/datasets/mansoordaku/ckdisease

[12] www.kidney.org/atoz/content/about-chronic-kidney-disease

AUTHORS

Dr. D Rathna Kishore [1] M.Tech, Ph.D., Associate Professor, Department of IT.

M.Keerthi Reddy[2] B.Tech, Andhra Loyola Institute of Engineering & Technology

V.Pavani[3] B.Tech, Andhra Loyola Institute of Engineering & Technology.

P.Madhumitha[4] B.Tech, Andhra Loyola Institute of Engineering & Technology.