CHRONIC KIDNEY DISEASE ANALYSIS USING DATA MINING CLASSIFICATION TECHNIQUES AS A MACHINE LEARNING PERSPECTIVE

Manasa Ayyalasomayajula, Ayushi Yadav, P.Rakesh, Kalla Khyathi Lekha

ABSTRACT

Chronic Kidney Disease (CKD) means the kidneys are damaged and are not filtering the blood the way they should. The primary role of kidneys is to filter extra water and waste from the blood to produce urine. If a person suffers from CKD, it means the wastes are collected inside the body which is chronic and can gradually cause damage to the kidneys in the long run. There are many causes for CKD like diabetes, high blood pressure, heart disease. If a particular kidney is not working, then one may notice one or more symptoms like abdominal pain, back pain, diarrhoea, fever, nosebleeds, rash, vomiting. Data Mining is one among the foremost encouraging areas of analysis with the aim of finding helpful information from voluminous knowledge of datasets. Data Mining is especially helpful in medical field when no handiness of proof favouring a treatment is found. Hence the Chronic Kidney Disease (CKD) dataset taken from UCI repository is analysed using different data mining classification. It is observed that LR, SVM techniques have given an accuracy of 93.8% and 94.0% respectively. The techniques RF, KNN, GNB and Voting Classifier have given an accuracy of 100% respectively. For this project, the CKD dataset has been taken from the UCI repository. Classification techniques like Logistic Regression, Support Vector Machine (SVM), K- nearest neighbours (KNN), Random Forest, GNB and Voting Classifier are used in the project.

INTRODUCTION

Chronic kidney Disease (CKD) means your kidneys are damaged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body. This disease is chronic because of the damage gradually over a long period. It is a flattering common disease worldwide. Due to CKD may have some health troubles. There are many causes for CKD like diabetes, high blood pressure, heart disease. Along with these critical diseases, CKD also depends on age and gender. If your kidney is not working, then you may notice one or more symptoms like abdominal pain, back pain, diarrhoea, fever, nosebleeds, rash, vomiting. There are two main diseases of CKD: (i) diabetes and (ii) high blood pressure.

So that controlling of these two diseases is the prevention of CKD. Usually, CKD does not give any sign till kidney is damaged badly. CKD is being increased rapidly as per the studies hospitalization cases increase 6.23 per cent per year but the global mortality rate remains fixed. There are few diagnostic tests to check the condition of CKD: (i) estimated glomerular filtration rate(eGFR) (ii) urine test (iii) blood pressure.

A. EGFR

eGFR value shows that how your kidney cleaning the blood. If your eGFR value is greater than 90, that means the kidney is normal. If eGFR value is less than 60, that means you have CKD.

B. URINE TEST

The doctor also asks for urine test for kidney functionality because kidneys make urine. If the urine contains blood and protein, that means your kidney is not working properly.

C. BLOOD PRESSURE

Doctor measures blood pressure as Blood pressure range shows how your heart is pumping blood. If eGFR value reaches less than 15, that means the patient has end-stage kidney disease. At this point, there are only available treatments: (i) dialysis and (ii) kidney transplant. Patient's life after dialysis depends on such factors as age, gender, frequency and duration of dialysis, physical movement of the body and mental health. If dialysis is not possible, the doctor has only one solution, i.e., kidney transplantation. However, it is extremely expensive. Therefore, it is critical noteworthiness in early recognition, monitoring and handling of the disease. It is essential to predict the striding of CKD with appropriate accuracy due to its dynamic and secretive nature in the early stages and patient abnormality. Medical treatment of CKD is prescribed by the stage. Anything other than this, it is very imperative to characterize the organization of the infection because it gives a few indications. It underpins the assurance of fundamental intercessions and medications.

Qin *et al.* proposed data assertion and sample diagnosis achievable in CKD diagnosis. KNN is used for data assertion. Six classifiers algorithms used for accuracy of diagnosis: logistic regression, random forest, support vector machine, K-nearest neighbour, naive Bayes classifier and feed-forward neural network. In these classifiers random forest gives better accuracy, i.e., 99.75%.

Vasquez-Morales *et al.* developed a neural network model for risk prediction of Chronic Kidney Disease development on 40000 instances dataset and their model accuracy was 95%.

Chen *et al.* applied three models on the dataset that is provided by UCI. They used KNN, SVM and soft independent modelling of class analogy (SIMCA) for finding the risk calculation of patient using these classifiers. In which the SVM and KNN model attained, the best accuracy of 99.7% and SVM model has the greatest capability to endure noise disturbance.

PROBLEM IDENTIFICATION & OBJECTIVES

Chronic Kidney Disease (CKD) means the kidneys are damaged and are not filtering the blood the way they should. The primary role of kidneys is to filter extra water and waste from the blood to produce urine. If a person suffers from CKD, it means the wastes are collected inside the body which is chronic and can gradually cause damage to the kidneys in the long run. There are many causes for CKD like diabetes, high blood pressure, heart disease. If a particular kidney is not working, then one may notice one or more symptoms like abdominal pain, back pain, diarrhoea, fever, nosebleeds, rash, vomiting.

For this project, the CKD dataset has been taken from the UCI repository. Classification techniques like Artificial Neural Network (ANN), C5.0, logistic regression, linear support vector machine (LSVM), K-nearest neighbours (KNN) are used in the project. The field of this project is Data Mining but we incorporated Machine Learning algorithms for an even better training and testing of the data to produce much efficient results. In the project at first the chronic kidney disease dataset is collected and uploaded. Later we read the data from the dataset. Then we perform Data Visualization i.e., we first analyse the number of disease and non-disease data. Data Visualization is a representation of the information in the form of pie charts, bar graphs, diagrams etc. Then we pre-process the data i.e., we check for any null values or any duplicate values that exist. Data pre-processing is a technique which is used to transform the data into a useful and an efficient format. After that we split the data for Training and Testing and build different classification models like KNN, RFC, SVM, GNB, Voting Classifier and draw a comparison graph to check which model gives the highest accuracy.

LITERATURE REVIEW

TITLE: REVIEW OF CHRONIC KIDNEY DISEASE BASED ON DATA MINING TECHNIQUES

AUTHORS: S. Dilli Arasu, Dr. R. Thirumalaiselvi

ABSTRACT: The Chronic Kidney disease is the most important health issues concerning the people as a whole. Chronic diseases lead to morbidity and increase of death rates in India and other low and middle income countries. The chronic diseases account to about 60% of all deaths worldwide. 80% of chronic disease deaths worldwide also occur in low and middle income countries. In India, probably the number of deaths due to chronic disease found to be 5.21 million in 2008 and seems to be raised to 7.63 million in 2020 approximately. Data mining is the process of extraction of hidden information from the large dataset. Major data mining techniques such as clustering, classification, association analysis, regression, time series and sequence analysis were used to predict kidney diseases. In this paper, the various data mining techniques are surveyed to predict kidney diseases and major problems are briefly explained.

TITLE: CHRONIC KIDNEY DISEASE (CKD) PREDICTION USING SUPERVISED DATA MINING TECHNIQUES

AUTHORS: S. Rajarajeswari, T. Tamilarasi

ABSTRACT: Diseases are causing high rates of mortality in the modern world, chronic kidney disease (CKD) is one of the major causes of mortality, and it has a long-term disability. The predisposing factors for CKD include diabetes mellitus, hypertension, cardiovascular diseases, smoking, obesity, family history of kidney disease and congenital kidney problems. CKD is associated with many complications such as, proteinuria, anaemia of CKD, CKD mineral and bone disorder, dyslipidaemia and electrolytes imbalance. Renal replacement therapy (dialysis and kidney transplantation) is the treatment of choice for CKD. Data mining is an accurate technique helps to predict the disease using various methods includes logistic regression, naive bayes classification, k-nearest neighbours, and support vector machine. Apart from these previous techniques, it was necessary to use a classification method for data segmentation according to their diagnosis and regression method for finding risk factors. In this present study, data are classified using proposed Identification of Pattern Mining, Decision Tree methods and regression techniques are used to obtain the best levels and this can be taken as metrics that the proposed methods can help in diagnosing a patient with CKD.

TITLE: PREDICTION OF CHRONIC KIDNEY DISEASE - A MACHINE LEARNING PERSPECTIVE

AUTHORS: Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michał Jasiński, Łukasz Jasiński, Radomir Gono, Elżbieta Jasińska, And Vadim Bolshev

ABSTRACT: Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority oversampling technique with full features. Techniques like KNN, RFC, SVM, GNB, Voting Classifier are used, and a comparison graph is drawn to estimate which technique gives the highest accuracy. It was estimated that linear support vector machine gave a result of 98.46% and when one deep neural network was applied the accuracy was the highest which is 99.6%.

TITLE: PREDICTION OF CHRONIC KIDNEY DISEASE USING DATA MINING CLASSIFICATION TECHNIQUES AND ANN

AUTHORS: Sahana B J

ABSTRACT: Information mining is a procedure of separating helpful data from gigantic measure of dataset. Information mining has been a present pattern for getting analytic outcomes. In therapeutic application, tremendous measure of unmined information is gathered by the social insurance industry with a specific end goal to find concealed data for successful conclusion and basic leadership. There are numerous information mining systems like grouping, clustering and so on. The goal of our paper is to anticipate Chronic Kidney Disease (CKD) utilizing arrangement methods like Naive Bayes and to foresee the phases of endless kidney illness utilizing the Artificial Neural Network (ANN) like C4.5.

TITLE: CHRONIC KIDNEY DISEASE PREDICTION USING CLASSIFICATION TECHNIQUES

AUTHORS: Pawan Agarawal, Sanjay Kumar

ABSTRACT: Healthcare industry faces the necessity to manage the growth and process the data into a new actionable insight. In this paper the usage of big data analytics and data mining in healthcare, overcomes the challenges of analysing the hidden information and extracting the useful information from the massive amount of data such as patients EHR's, which provides an intuition of predicting the chronic kidney disease at the early stage. The aim of the paper is to predict the chronic kidney disease using classification algorithm on structured dataset. There are many classification algorithms namely Logistic Regression, Random Forest, Naïve Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbours are used in the prediction model to evaluate the performance based on the accuracy, precision and f1-measure on the given dataset. This paper studies the usage of classification algorithm in the prediction model to help the physicians to make right decision to extend the life span.

METHODOLOGY

OVERVIEW OF TECHNOLOGIES

The technologies or the classification techniques used in this project are Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbours, Gaussian Naïve Bayes' and Voting Classifier.

a. LOGISTIC REGRESSION

Logistic regression is also a type of supervised learning algorithm. It is a statistical model. The probability of target value is predicted from logistic regression. It is divided the target attribute into two-classes: success or not success. For success, it returns 1 whereas it returns 0 for not succeeding. Logistic regression is represented by the equation:

$$\boldsymbol{P = 1/ (1 + e\hat{}\ (- (b0 + b1x + b2x\hat{}2))}$$

where P is the predicted value, b0, b1, b2 are biases and x is an attribute. It is used in various field of machine learning application in social sciences and medical arena, for example, for spam detection, diabetes detection, cancer detection, etc. Logistic regression is the advanced version of linear regression. Through this technique, we only concern about the probability of the outcome variable.

b. RANDOM FOREST

The random tree is a type of supervised classifiers. It produces lots of distinct learners. The stochastic process is used to form the tree. It is a type of ensemble learning technique for classification. It works the same as decision tree, but a random subset of attributes uses for each split. This algorithm uses for both classification problems and regression problems. A group of random trees is known as a forest. The random trees classifier takes the input feature set and classifies input for every tree in the forest. The output of the random tree selects from the majority of votes. In the tree, every leaf node holds a linear model. The bagging training algorithm is used to train the model.

c. SUPPORT VECTOR MACHINE

Linear support vector machine is the modern particularly fast machine learning algorithm for solving multiclass classification problem for the large dataset based on a simple iterative approach. It is created the SVM model in linear CPU time of the dataset. LSVM can be used for the high dimensional dataset is the sparse and dense format. It is used for solving the large dataset machine learning problems in less expensive computing resource. Support Vector Machine is a supervised classifier algorithm. It is used kernel trick for solving the classification problem. Based on these transformations ideal edge is found between the possible outputs. SVM is used for the nonlinear kernel, such as RBF. For the linear kernel, LSVM is an appropriate choice. LSVM classifier is sufficient for all linear problems.

d. K-NEAREST NEIGHBOURS

KNN is a simple type of supervised algorithm. It can be used for both classification and regression problems. It is largely used for classification problems. KNN does not use a particular training stage and use all the data for training so that it is a lazy learning algorithm and also it does not consider anything about the underlying data. KNN stores the whole dataset because it has no model so that there is no learning required. When the new data enter for predicting the outcomes, it compares K neighbours so that selection of K's value is very important. The distance is calculated between two already label data. The distance helps to find the nearest neighbour of the new data. A Euclidian method is used for finding the distance.

e. GAUSSIAN NAIVE BAYES'

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be:

$$\mathbf{P\ (xi\ |\ y) = 1\ /\ \sqrt{2\pi\sigma y\hat{}2}\ exp\ (-\ (xi\ -\ \mu y)\ \hat{}2\ /\ 2\sigma y\hat{}2)}$$

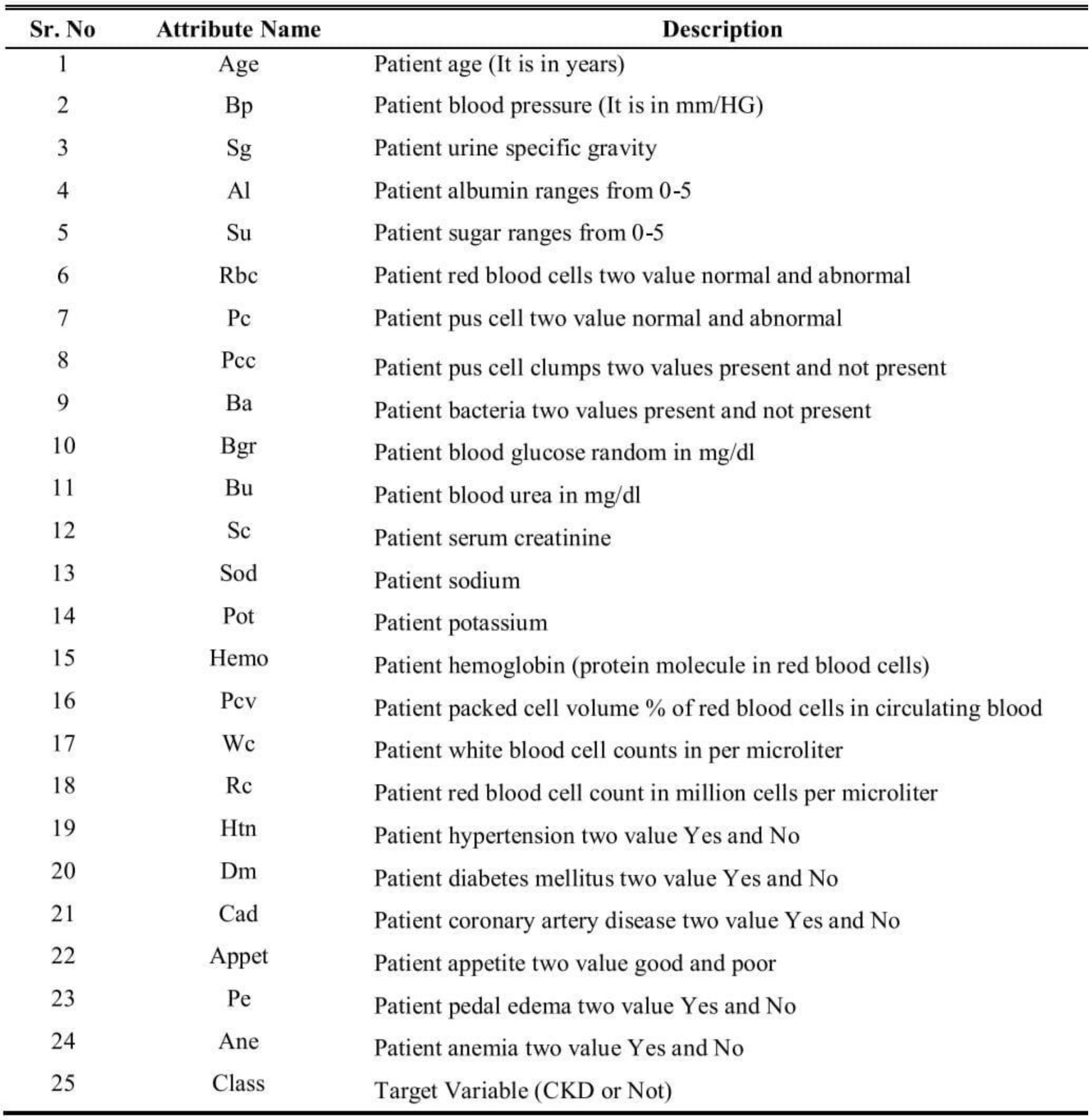
Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

f. VOTING CLASSIFIER

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

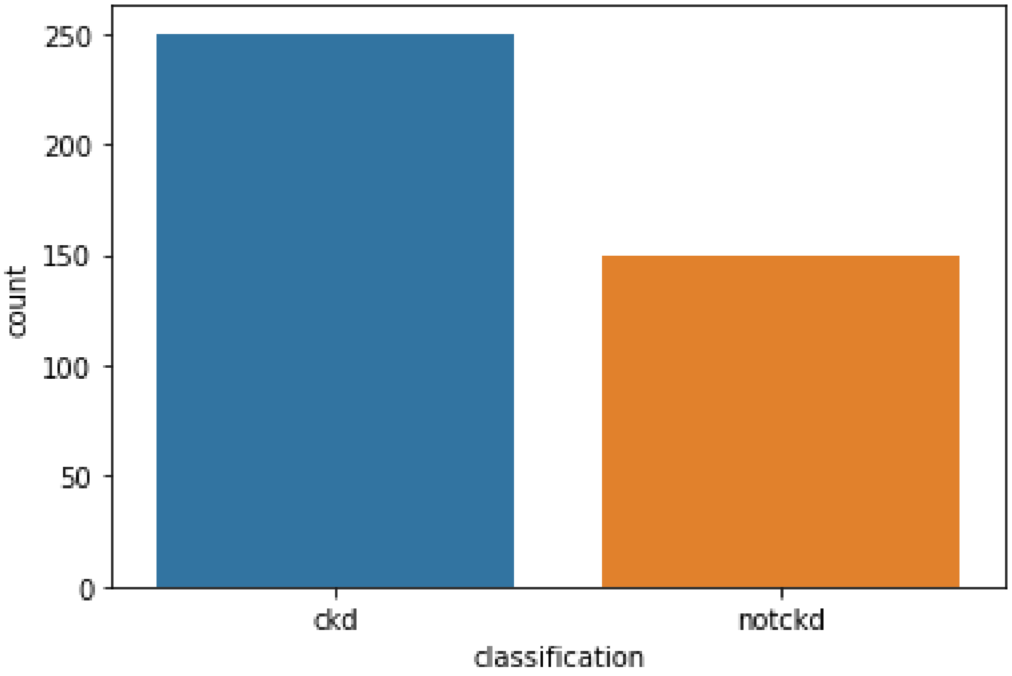
Description of Attributes in the Dataset.



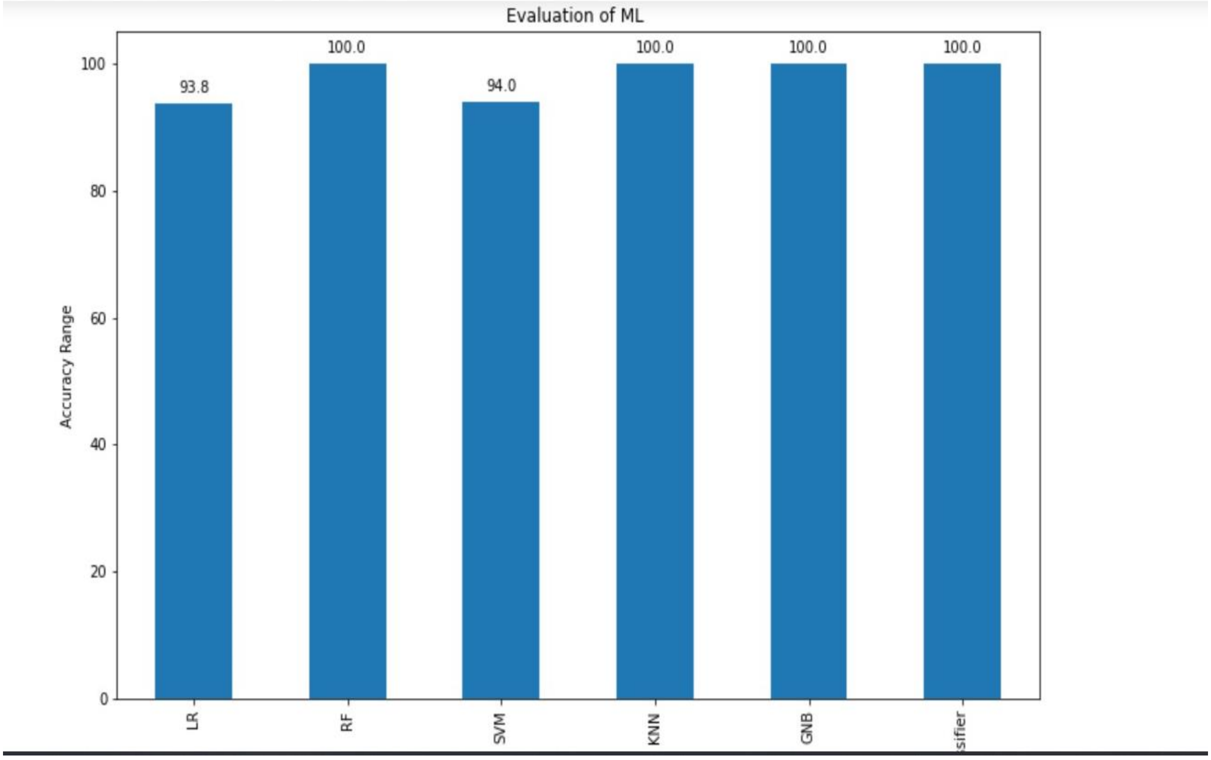
Sr. No	Attribute Name	Description
1	Age	Patient age (It is in years)
2	Bp	Patient blood pressure (It is in mm/HG)
3	Sg	Patient urine specific gravity
4	Al	Patient albumin ranges from 0-5
5	Su	Patient sugar ranges from 0-5
6	Rbc	Patient red blood cells two value normal and abnormal
7	Pc	Patient pus cell two value normal and abnormal
8	Pcc	Patient pus cell clumps two values present and not present
9	Ba	Patient bacteria two values present and not present
10	Bgr	Patient blood glucose random in mg/dl
11	Bu	Patient blood urea in mg/dl
12	Sc	Patient serum creatinine
13	Sod	Patient sodium
14	Pot	Patient potassium
15	Hemo	Patient hemoglobin (protein molecule in red blood cells)
16	Pcv	Patient packed cell volume % of red blood cells in circulating blood
17	Wc	Patient white blood cell counts in per microliter
18	Rc	Patient red blood cell count in million cells per microliter
19	Htn	Patient hypertension two value Yes and No
20	Dm	Patient diabetes mellitus two value Yes and No
21	Cad	Patient coronary artery disease two value Yes and No
22	Appet	Patient appetite two value good and poor
23	Pe	Patient pedal edema two value Yes and No
24	Ane	Patient anemia two value Yes and No
25	Class	Target Variable (CKD or Not)

Description of Used Mathematical Symbol.

Symbol	Name
p	Significant Value
α	Significant Level
P	Predicted Value
$b_{0,}$ $b_{1,}$ b_2	Bias
x	Attribute
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
accuracy	Classification Accuracy
Error	Classification Error
Precision	Classification precision rate
Recall	Classification recall rate
F-Measure	Classification F1 Score



The classification of people who are diagnosed with CKD and those who are not diagnosed with CKD.



It is observed that LR, SVM techniques have given an accuracy of 93.8% and 94.0% respectively. The techniques RF, KNN, GNB and Voting Classifier have given an accuracy of 100% respectively.

ACKNOWLEDGEMENT

We wish to convey gratitude to our principal Dr. Y. Madhavee Latha, for providing us the environment and mean to enrich our skills and motivating us in our endeavor and helping us to realize our full potential.

We express our sincere gratitude to Dr.Sireesha.R, Head of the Department of Computer Science and Engineering for inspiring us to take up a project on this subject and successfully guiding us towards its completion.

We express our sincere gratitude to Prof.A.Radha Rani, Director of Computer Science and Engineering for inspiring us to take up a project on this subject and successfully guiding us towards its completion.

We would like to thank our internal guide Mr. K.Sandeep Varma and all the faculty members for their valuable guidance and encouragement towards the completion of our project work.

With Regards and Gratitude:

Manasa Ayyalasomayajula (Reg No: 121810302023)

Ayushi Yadav (Reg No: 121810302055)

P. Rakesh (Reg No: 121810302060)

Kalla Khyathi Lekha (Reg No: 121810302062)

CONCLUSION

We learnt that using Data Mining Classification techniques we can predict which technique gives the highest accuracy in effective diagnosis of the Chronic Kidney Disease (CKD). Data mining is defined as the process of extracting the huge hidden data from a large dataset, categorizing valid and unique patterns in data. Huge amount of unmined data is collected by the healthcare department in order to discover the hidden facts for effective diagnosis and also decision making.

This project is a medical sector application which helps the medical practitioners in predicting the CKD based on the CKD parameters. It is an automation for CKD prediction and it efficiently and economically speedily identifies the disease, its types and complications from the clinical database.

The Chronic Kidney Disease (CKD) dataset taken from UCI repository is analysed using different data mining classification. It is observed that LR, SVM techniques have given an accuracy of 93.8% and 94.0% respectively. The techniques RF, KNN, GNB and Voting Classifier have given an accuracy of 100% respectively.

All the machine learning models but Logistic and KNN classifiers give satisfactory result and have the negligible difference between precision and recall values.

In comparison with them precision for Logistic and KNN classifiers is low whereas recall is high. It indicates that these two classifiers give many False positive results due to unbalanced dataset. Logistic and KNN algorithms have not enough capacity to distinguish between positive class and negative class as the related AUC score is very low.

REFERENCES

1. S. Dilli Arasu, Dr. R. Thirumalaiselvi, “Review of Chronic Kidney Disease based on Data Mining Techniques”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp.

2. S. Rajarajeswari, T. Tamilarasi, “Chronic Kidney Disease (CKD) Prediction Using Supervised Data Mining Techniques”, Int. J. Advanced Networking and Applications Volume: 12 Issue: 06 Pages: 4776-4780 (2021) ISSN: 0975-0290.

3. Pankaj Chittora, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michał Jasiński, Łukasz Jasiński, Radomir Gono, Elżbieta Jasińska, And Vadim Bolshev, “Prediction of Chronic Kidney Disease - A Machine Learning Perspective”, Received January 10, 2021, accepted January 15, 2021, date of publication January 22, 2021, date of current version February 1, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3053763.

4. Sahana B J, “Prediction of Chronic Kidney Disease Using Data Mining Classification Techniques And ANN”, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, NCETEIT - 2017 Conference Proceedings.

5. Pawan Agarawal, Sanjay Kumar, “Chronic Kidney Disease Prediction Using Classification Techniques”, ISSN- 2394-5125 VOL 7, ISSUE 17, 2020.