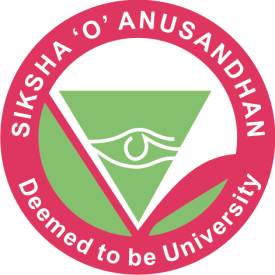
****

**SIKSHA ‘O’ANUSANDHAN**

## DEEMED TO BE UNIVERSITY

**Admission Batch: 2020 Session: 2020-2021**

**A DMAA Project Report On**

**Analysis of chronic Kidney disease by using orange tool**

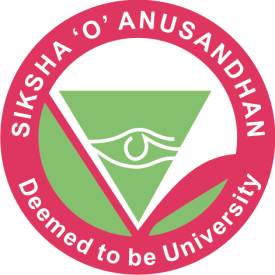
***Submitted by:-***

**Ambika Sahoo**

**MCA 1St year**

**ITER, SOA UNIVERSITY**

**ReGD No-206120070**



**DEPARTMENT OF MCA (2020-22)**

**ITER, SOA UNIVERSITY**

**Jagamohan Nagar, Jagamara, Bhubaneswar, Odisha – 751030**

**Abstract:-**

This dataset are often wont to predict the chronic renal disorder and it are often collected from the hospital nearly 2 months of period. Chronic renal disorder (CKD) is a crucial and customary no communicable condition globally. In national and international guidelines, CKD is defined and staged consistent with measures of kidney function that leave a degree of risk stratification using commonly available markers . Kidney Disease: Improving Global Outcomes conducted a survey and sponsored a controversies conference to supply a transparent understanding to both the nephrology and no nephrology communities of the evidence base for the definition and classification recommended by renal disorder Quality Outcome Initiative develop global consensus for the adoption of an easy definition and arrangement , and (3) identify a collaborative research agenda and plan which may improve the evidence base and facilitate implementation of the definition and classification of CKD..

**Keyword:**

Data mining , chronic kidney disease, Orange ,K-Nearest Neighbour, naive bayes

**Introduction :**

Data mining is study of algorithms for locating patterns in large data sets. It is an integral a neighborhood of recent industry, where data from its operations and customers are mined for gaining business insight. It is also important in modern scientific endeavours. Data mining is an interdisciplinary topic involving, databases, machine learning and algorithms. The course will cover the basics of knowledge mining. It will explain the essential algorithms like data pre-processing, association rules, classification, clustering, sequence mining and visualization. It will also explain implementations in open source software. Finally, case studies on industrial problems are going to be demonstrated. Data mining is that the analysis of knowledge for relationships that haven't previously been discovered or known. A term coined for a new discipline lying at the interfaceof database technology, machine learning, patternrecognition, statistics and visualization.The key element in much more elaborate process called„Knowledge Discovery in Databases”.The efficient extraction of previously unknown patternsin very large data bases. Data mining is that the analysis of (often large) observational datasets to hunt out unsuspected relationships and to summarize thedata in novel ways during which are both understandable and useful to the data owner (Hand, Mannila, Smyth). The main objective specialise in this research is to seek out an optimal model and test the power of classification algorithms with state of the art parties in global health care domain. A number of experiment have been conducted using orange for comparission to prediact of Data mining technique the Kidney Disease 2000 record using different attribute.

**Orange:**

Orange is a open source Data Visualization and analysis tool and it is a machine learning software and visual front-end. Its graphical interface builds upon the cross-platform framework. Orange is an open-source software package released under GPL. Orange may be a scriptable environment for quick prototyping of the newest algorithms and testing patterns. It is a gaggle of python-based modules that exist within the core library. It implements some functionalities that execution time isn't essential, which is completed in Python.

**What is chronic kidney disease (CKD)?**

Chronic renal disorder includes conditions that damage your kidneys and reduce their ability to stay you healthy by doing the roles listed. If renal disorder gets worse, wastes can build to high levels in your blood and cause you to feel sick. you'll develop complications like signal , anemia (low blood count), weak bones, poor nutritional health and nerve damage. Also, renal disorder increases your risk of getting heart and vessel disease. These problems may happen slowly over an extended period of your time . Chronic kidney disease could also be caused by diabetes, high vital sign and other disorders. Early detection and treatment can often keep chronic renal disorder from getting worse. When renal disorder progresses, it's going to eventually cause renal failure , which needs dialysis or a kidney transplant to take care of life.

**Methodology:**

Here we use 2 classification techniques i.e naïve baiyes & KNN classifeier to predict the presence of chronic kidney disease in human body.The dat set collected & applied by 2 different classification techniques which is been evaluated on the basis of accuracy,precision,F -measure & recall. Sample diagram of predictive data mining.

**Proposed approach:**

**Data set for CKD**

**Data mining Classification**

**Naïve baiyes SVM Random forest KNN**

**Performance evaluation**

**Which is the best approach????????**

**Fig: 1.1-proposed system**

**The working of the architecture is as follows:**

The dataset of CKD patients have been collected & with the help of Orange data mining tool we pass the data, into the classifier named naïve baiyes & KNN .In this project we collected the data set from UCI machine learning repository as the input for prediction which consists of attributes & values. The resultant value is to find the accuracy that how many patientients are suffered from CKD within a particular time period. In order to ameliorate the prediction two evaluation parameters are taken as classifiers. The fact-finding outcome is retrieved, which shows best classifier between the two.

**Data set:**

Used dataset is collected form multiple PathoLabs, Hospitals. This dataset contains 400 instances with 25 attributes.The attributes during this KFT dataset age, blood pressure(bp), specific gravity(sg), albumin(al), sugar(su), red blood cells(bc), pus cell(pc), pus cell clumps(pcc), bacteria(ba), blood glucose random(bgr), blood urea(bu), serum creatinine(sc), sodium(sod), potassium(pot), haemoglobin(hemo), packed cell volume(pcv), white blood cell count(wc), red blood cell count(rc), hypertension(htn), diabetes mellitus(dm), coronary artery disease(cad), appetite(appet), pedal edema(pe), anaemia(ane), class.This dataset consists of renal affected disease information. This is binary classification, as we've used two classes for predicting CKD and NOT CKD.

Orange tool is open source application which provides user friendly tool.Chronic kidney disease patients are need more care before they diagnose their issues, doctors are need to first accurate eastimate by viewing their previous reports .So medical science invented many scientific techniques & different methods to identify the disease properly before it’s severity.

**Evaluation parameters:**

Here’s the data mining parameters are metioned bellow:

**A: Sensitivity**(True positive rate): Estimate the proportion of ailing people that are precisely recognized from data set.

**B: Specificity**(Actual negatives) : Detect the exact proportion of healthy peo ple from dataset.

**C: Precision:** This is also called positive Predictive ones , which is defined as the average probability of pertinent sampling.

**D: Recall:** It is defined as average probability of complete sampling.

**E: Accuracy**( Rate of correct classifications)

**F: Confusion Matrix:**

It displays the no. of correct & incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is represented in the form of n-by-n ,where n is the no. of classes .The accuracy of each classification algorithms can be calculated from that.

Sensitivity = number of true positives

Number of true positives + number of false negatives

Specificity = number of true negatives

Number of true negatives + number of false positives

Precision = number of true positives

Number of true positives + number of false positives

Recall = number of true positives

Number of true positives + number of false negatives

Accuracy = number of correctly classified samples

Total number of samples

**G: Naïve Bayes:** Simply it is a probabilistic classifier drew on Bayes’ theorem with strong (naïve)independence presumption amidst the features. This is highly flexible & needs a number of variables in a learning algorithm. The comparative study from CKD dataset using orange tool in cross validation sampling keeping no.of folds 10, the classification accuracy is 94.8%, precision is 95.4% & recall is 94.8.

**H: Support Vector machine:**

Support vector machine (SVM) is a powerful, state-of -the-art algorithm based on linear & non-linear regression. Oracle data-miming implements SVM for binary & multiclass classification. The advantage of the SVM is that ,by use of the so called “Kernel trick”, the distance b/w a molecule & the hyper plane can be calculated in a transformed(non-linear feature space, lacking of the explicit transformation of the original descriptors. The radial basis function kernel (Gaussian kernel) which is the most commonly used was applied to this study. For our data sets CKD comparative analysis of precession, recall & accuracy is respectively.

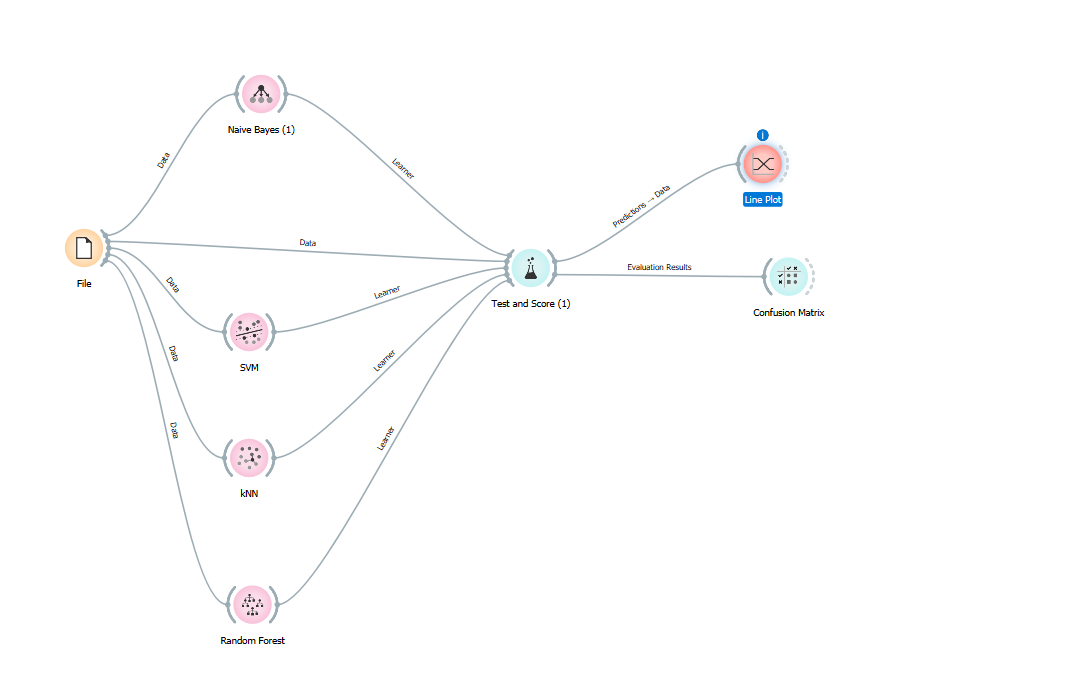
**I: Random forest:**

RF classification tree is an ensemble of unpruned classification tree they give good performance in practical problem. If data sets are not noisy, they work excellent & it is not subject to over fitting .It works fast, generally it’s more accurate than other tree base algorithms. Random forest predictions are based on no. of trees & each tree are train within isolation. We have 3 main choices to performed random tree are constructed. For our data sets CKD, comparative analysis of precession & recall & accuracy is respectively by analysing orange.

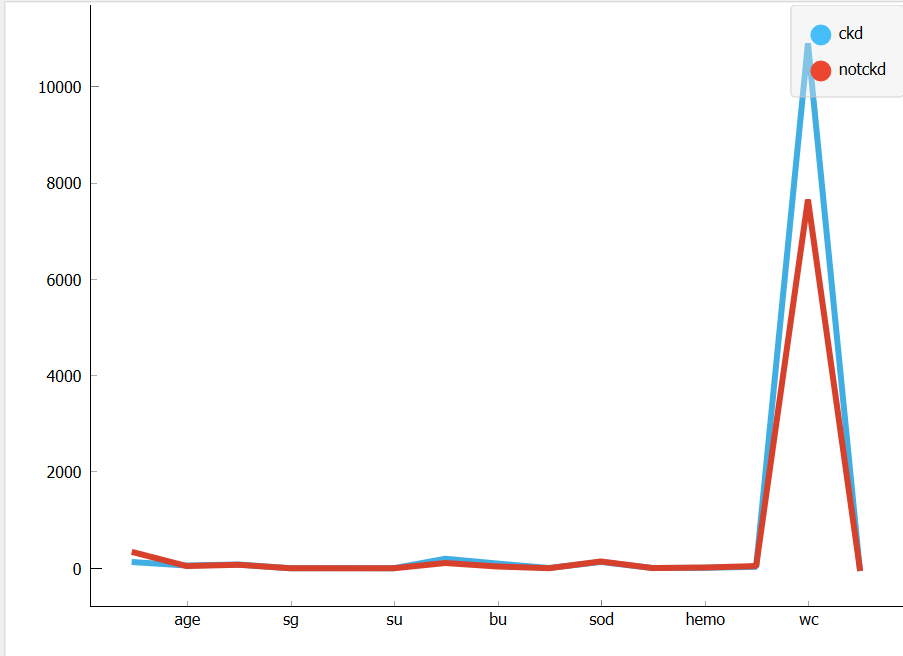
**J: KNN (K-nearest neighbour):**

KNN is an advanced method for classification which find the group of object k’s in trading data sets & closed into test values & in the event that equivalent class is partaken in different of K-closest neighbours, at the point per-neighbour loads of as class are included or the subsequent weighted aggregate is utilized as the probability score of that class as for the test record. Its accuracy is depending open k’s values & use nearest neighbour classifier.Sekecting K’s can be complete experimentally & where a no. about shapes taken out from the training are set KNN.Our datasets kidney disease comparative analysis of precession ,accuracy & recall is respectively.

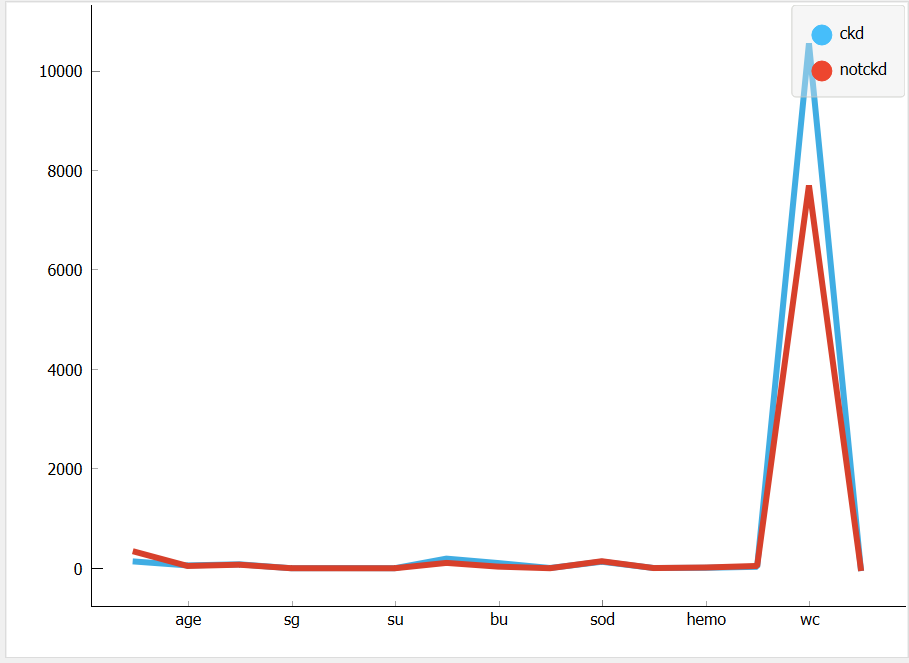
**Work flow in orange tool:**

****

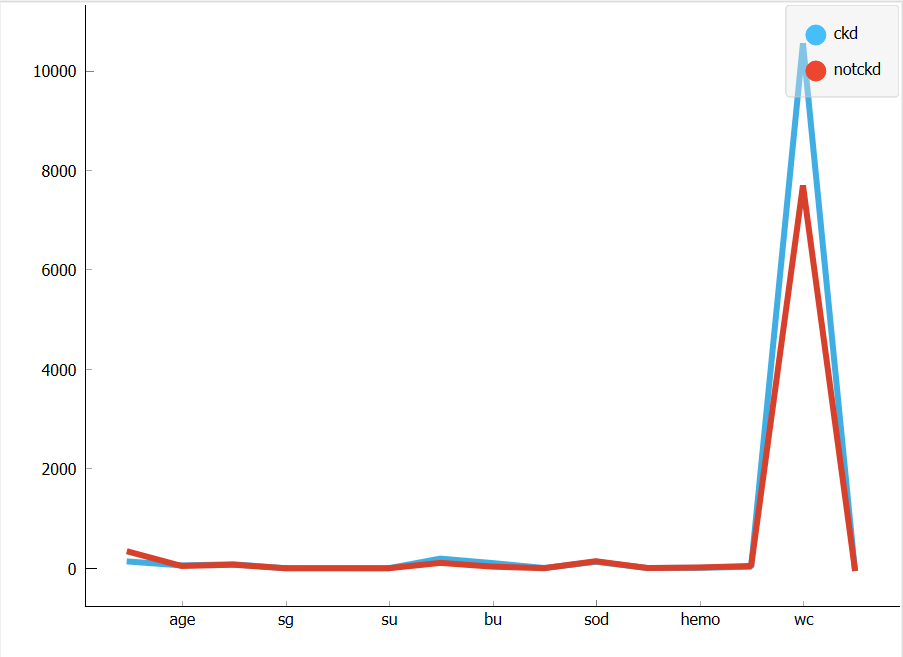
**Fig 1: work flow In orange tool**



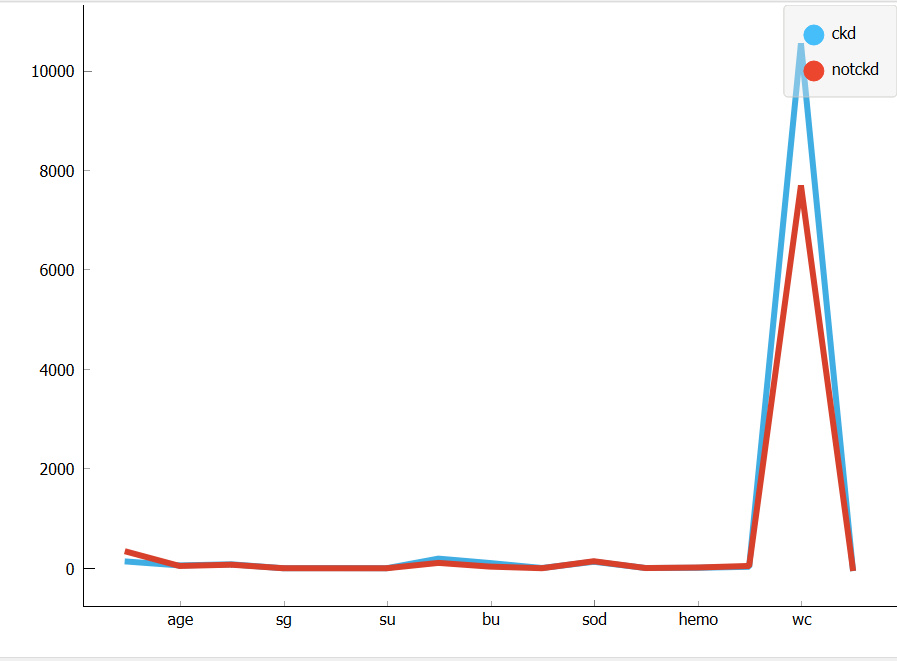
**Fig 2: Line plot using KNN-algorithm**



**Fig 3: Line plot using Naïve baiyes algorithm**



**Fig 4: Line plot using SVM algorithm**



**Fig 5:Lineplot using Random forest algorithm**

**Result analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| Classification  Algorithm | Precession | recall | accuracy |
| KNN | **0.943** | **0.932** | **0.932** |
| Naïve bayes | **0.954** | **0.948** | **0.948** |
| SVM | **0.998** | **0.998** | **0.998** |
| Random forest | **1.000** | **1.000** | **1.000** |

**Table 1: classification Algorithms with orange tool result**

**DISCUSSION:**

Orange tool is open source application which provides user friendly tool.Chronic kidney disease patients are need more care before they diagnose their issues, doctors are need to first accurate eastimate by viewing their previous reports .So medical science invented many scientific techniques & different methods to identify the disease properly before it’s severity.

Data mining techniques help to figure out best classification techniques to help the predict kidney diseases its patterns & behavior.Classification algorithms are help to find out data mining methods to unclassified cases or label class & visualisation tools like boxplot,line plot,shieve dagram etc. help us to to visualize he predicted data in graphical way.

As per my result in orange tool I am implementing 4 algorithms of classification & our result says that Random forest is more accurate as compared to other 3 algorithms.This algorithms are used in computer version & also in robotics.For our datasets kidney disease comparative analysis of ckd & non-ckd value are 250 & 150 respectively.

Socio-statistic & life organize based devision approach give client inclinations towards the products dependent on their statistics & life organise patterns..Ten times cross approval is utilizedto assess the presentation of the classifiers.The after effects of random-forest classifier show improved exactness in arranging occasions of significant portions.It also delivered better outcomes in foreseeing & controlling the clients of various life sections towards item inclinations contrasted with other 3 methodologies.

Random forest are proven that this algorithm are calculate more accuracy then other we work on in future to make decision on it & make ML to get prediction in early basses just data sets & make all possible prediction on it & the predictions help to identify kidney disease in an accurate way.

**Conclusion:**

Data mining for health care is evaluating the medical treatments and ensures detection of fraud and abuse. Comparative analysis concerning accuracy,precession and recall ,orange is the best overall performance compared to other tools. In future, more disease dataset can be used for classification methods. Data Mining is a potentially powerful tool for companies; however, more research is needed to measure the benefits of Data Mining.

**References:**

1. Fraser SD, Roderick PJ, Aitken G, et al. Chronic kidney disease, albuminuria and socioeconomic status in the Health Surveys for England 2009 and 2010. J Public Health (Oxf) 2014;36(4):577–586. [PubMed] [Google Scholar]
2. Bixia Gao, ... Shougang Zhuang, in Chronic Renal Disease (Second Edition), 2020
3. KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Anemia in Chronic Kidney Disease. Am.J.Kidney Dis. 2006;47:S11–S145. [PubMed] [Google Scholar]
4. McClellan W, Aronoff SL, Bolton WK, et al. The prevalence of anemia in patients with chronic kidney disease. Curr.Med.Res.Opin. 2004;20:1501–1510. [PubMed] [Google Scholar]
5. Gal-Moscovici A, Sprague SM. Bone health in chronic kidney disease-mineral and bone disease. Adv.Chronic.Kidney Dis. 2007;14:27–36. [PubMed] [Google Scholar]
6. El-Kishawi AM, El-Nahas AM. Renal osteodystrophy: review of the disease and its treatment. Saudi.J.Kidney Dis.Transpl. 2006;17:373–382. [PubMed] [Google Scholar]
7. A. Becks, J.-C. Toebermann, in Computer Aided Chemical Engineering, 2002

**External Links:-**

**For dataset-**[**https://www.kaggle.com/mansoordaku/ckdisease**](https://www.kaggle.com/mansoordaku/ckdisease)

[https://en.wikipedia.org/wiki/Orange -(software)](https://en.wikipedia.org/wiki/Orange%20%20-(software))