Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches

BY

Minhaz Uddin Emon
ID:171-15-1294

Md. Al Mahmud Imran
ID: 171-15-1243

Rakibul Islam
ID: 171-15-1263

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ohidujjaman
Senior Lecturer
Department of CSE
Daffodil International University

Co Supervised By

Amit Chakraborty Chhoton
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2021

APPROVAL

This project titled "**Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches"**, submitted by Minhaz Uddin Emon, Md. Al Mahmud Imran and Rakibul Hassan to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15-01-2021.

BOARD OF EXAMINERS

Professor Dr. Touhid Bhuiyan — **Chairman**
Professor & Head
Department of CSE
Daffodil International University

Dr. S. M. Aminul Haque — **Internal Examiner**
Associate Professor & Associate Head
Department of CSE
Daffodil International University

Ohidujjaman — **Internal Examiner**
Senior Lecturer
Department of CSE
Daffodil International University

Dr. Mohammad Sorif Uddin — **External Examiner**
Professor
Department of Computer Science & Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ohidujjaman, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

Ohidujjaman
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

Amit Chakraborty Chhoton
Lecturer
Department of CSE
Daffodil International University

Submitted by

Minhaz Uddin Emon
ID: -171-15-1294
Department of CSE
Daffodil International University

Md. Al Mahmud Imran
ID: -171-15-1243
Department of CSE
Daffodil International University

Rakibul Islam
ID: -171-15-1263
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing that makes it possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Ohidujjaman**, **Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of machine learning, data mining to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice,reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan,** Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire coursemate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Machine learning and data mining play a vital role in health care and also medical information and detection, Now a day machine learning techniques use awareness of some major health risks such as diabetic prediction, brain tumor detection, covid 19 detections, and many more. The kidney is the most important organ of our body and if it has any problem then the impact is more dangerous to our body. Chronic kidney disease (CKD), otherwise referred to as renal disease. Chronic kidney disease requires disorders that damage and reduce the capacity of our kidneys to keep us healthy. So, we need to be concerned about kidney disease to our very primary stage. We take a few attributes to measure our analysis about chronic kidney disease and this attribute is one of the major occurrences of chronic kidney disease. Therefore 8 machine learning classifier are used to measure analysis using weka tools namely: Naive Bayes(NB), Logistic Regression(LG), Multilayer Perceptron(MLP), Stochastic Gradient Descent(SGD), Adaptive Boosting(Adaboost), Bagging, Decision Tree(DT), Random Forest(RF) classifier are used. We feature extraction of all attributes using principal component analysis(PCA). We gain the highest accuracy from the Random Forest(RF) and it is 99% and ROC(receiver operating characteristic) curve value is also highest from other algorithms.

TABLE OF CONTENTS

TABLE LIST

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Nowadays machine learning is an important field for research. In different areas, different statistical and machine learning algorithms are implemented. Such as marketing, health, and medical problems, weather forecasting, the study of socioeconomic behavior, etc we can use machine learning in those fields. In the medical field, many diseases can be detected by machine or predicted by machine by the application of machine learning algorithms. Chronic diseases pose an important threat to the global health agenda of the 21st century. The rising prevalence of chronic diseases such as chronic kidney disease has serious consequences for health and economic output in developing countries. The rapid increase in common risk factors, especially among the poor, such as diabetes, hypertension, and obesity, would result in even greater and deeper burdens that developing nations are not prepared to cope with. There has been a lack of exposure to chronic diseases, chronic kidney disease in particular, primarily due to the focus of the global health community on infectious diseases, and lack of consciousness. There is a vital need to concentrate on it and the adoption of more inclusive, cost-effective, and preventive chronic disease strategies by developed countries.

There are many hospitals which store the data of chronic kidney disease patients in their database. Through analyzing these data, various patterns can be found that will be helpful for decision-making. Using data mining techniques on these data, it is possible to discover many kinds of knowledge and use this knowledge to predict the disease. There is a massive amount of people who are affected by chronic kidney disease. In Bangladesh maximum are not concerned about the disease. As a result, the affected people by the disease are increasing day by day. If people could detect or predict if they are affected or going to be affected then it could be controlled. People could take the necessary steps to not be affected. For predicting the disease, predictive can be used. There are several methods that can be

used like, classification, regression, categorization. Among those, many people think classification is the best. After accomplishing the research, it will help to predict Chronic Kidney disease. People will be aware of the disease and their health condition.

The main objectives of the research are to predict the disease using a machine learning algorithm, warning whether a person at risk of the disease or not, and comparing the outcomes of various machine learning algorithms to determine which algorithm provides the best performance.

1.1 Background

1.1.1 How kidney works

As for the internal structure of the kidney, the renal cortex and the renal cortex are considered the outer zone. The medulla is called the inner field. The most distinctive are the renal pyramids and renal papillae. The medulla's features are characterized by the connective tissue's renal columns. Extensions are emitted by the medulla from the cortex downwards. Collection Ducts carrying urine formed for excretion by nephrons to the kidney calyx They are bundled with papillae. The primary working and operating unit of the kidney Nephrons, guy. Really they're the kidney filtering unit. Each of the two kidneys contains Around a million nephron-based filtering units. Each nephron operates in 2 steps: Blood collects the required material into the glomerulus and then the renal tubules. Blood, guy. The artery of the kidney carries blood to the kidney. This large blood vessel is touching Nephrons in smaller blood vessels. When the blood hits the nephron, it passes out. Often known as the Malpighian body in the renal corpuscle. The Renal Corpuscle includes The glomerulus and the capsule of Bowman. The glomerulus is a small blood cluster. Ships. It filters blood and consumes larger molecules including blood cells and proteins It flows through the blood vessel and the remaining smaller molecules called capsular urine. The Bowman capsule into the renal tubules. The renal tubules end up in the set Ducts, guy. Each renal tubule consisted of a PCT (Proximal convoluted tubule) DCT (Distal convoluted tubule) and Henle. A blood vessel runs through each of the tubules. As the

diluted fluid travels down the channel, the PCT absorbs water, sodium and glucose. Back in the blood, Loop of Henle absorbs more potassium, chloride and sodium. Blood then DCT extracts extra acid from the blood. At this time the fluid has been filled with Urea, a by-product of protein synthesis. There's a nephron at the end of it. Collecting duct where the fluids are drained out of the nephrons. Calyces are collecting the remaining fluids And the waste to the bladder. Here, excess liquid and waste is urine. In the end, the Filtered blood flows out of the kidney into the renal vein. If the filter rate is Glomerulus steadily decreases the amount of radioactive content in the blood that tends to rise. Which ends up in the form of chronic kidney disease.

1.1.2 **Symptom of Chronic Kidney Disease**

Many patients with chronic kidney disease (CKD) will not show complications and they usually do not cause complications until they enter an advanced stage.

Symptom have seen in:

- Loss of weight and reduced appetite
- Swollen knees, thighs or hands – due to water accumulation (oedema)
- Breath shortness
- Fatigue
- Blood in the urine (urine)
- Increased demand for pee – particularly at night
- Difficulty to sleep (insomnia)
- Skin itchy
- Muscle cramping
- Feel sick
- Headache
- Male erectile dysfunction

1.1.3 Stage of Chronic Kidney Disease

To help increase the quality of treatment for people with kidney disease, the National Kidney Foundation (NKF) has developed a checklist to help physicians recognize each stage of kidney disease. NKF has categorized kidney failure into five stages. When a doctor knows what level of kidney failure a person has, he or she will give the right care, since each stage needs various testing and therapies. It depends on the Glomerular filtration rate (GFR) measure

First stage - when GFR measure is more than 90

Second stage - when GFR measure is 60-89

Third stage - when GFR measure is 45-59

Fourth stage - when GFR measure is 30-44

Fifth stage - when GFR measure is 15-29

1.1.4 Chronic Kidney Disease Test

- Blood Test : The main examination for kidney failure is a blood check. The test tests the volume in your blood of a waste product called creatinine.
- Urine Test : Check the concentrations in the urine of compounds called albumin and creatinine, known as albumin: creatinine ratio, or ACR.
- Other Test : An ultrasound scan, MRI scan, or CT scan to see what the kidneys look like and to check if there are any blockages.

1.2 Research Motivation

Chronic Kidney Disease is a life threatening disease, if we have proper knowledge about it and how we prevent this then we will save our life from affecting this disease. Since there is no early indication of this disease, the only estimation of the test outcome of this disease is The possibility of CKD can be revealed in blood and urine. Early identification, however,

will aid Patients will benefit from this disorder. Early detection is based on the forecast for the Test effect. This is my justification for working on CKD prediction using the computer. Classifiers for Literacy

1.3 Problem Description

In this paper, the dataset is used from UCI machine learning library dataset for Chronic Kidney Disease. Creating the best predictive model of accuracy, precision, recall, f-measure and learning rate among the chosen ones Models by applying different classifiers for machine learning and evaluating their outcomes after Preprocessing and converting data.

1.4 Research Objective

- The main objectives of the research are to predict the disease using a machine learning algorithm, warning whether a person at risk of the disease or not, and comparing the outcomes of various machine learning algorithms to determine which algorithm provides the best performance
- To allow the property perspective on the used dataset to be properly mitigated.
- To find a better statistical machine-learning algorithm in order to find a lower FN count more accurately.

1.5 Research Scope

Machine learning algorithms focused on classification will predict or not chronic kidney disease. Using machine learning methods, we find greater performance. In the future, we will work as neural networks in deep learning to make decisions.

1.6 Research Organization

The paper residue is structured as follows:

- The related works are described in chapter 2
- The entire methodology is shown in chapter 3.

- The findings are discussed in chapter 4.
- The conclusion and future work are discussed in chapter 5.

CHAPTER 2
RELATED WORK

Amirgaliyev et. al. [1] predicted the results of the patient of chronic kidney disease dataset. Medical history, physical exams, and laboratory studies were the basis of the collected dataset. They developed models with this dataset using the support vector machine algorithm for linear kernel SVM classifiers that were evaluated to find the best ratings for sensitivity, specificity, and accuracy metrics. They demonstrated that 93.1 percent accuracy was the best.

Charleonnan et. al. [2] predicted chronic kidney disease using clinical data. Together with K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and call tree classifiers for predicted CKD, four machine learning strategies were explored. Five times the average precision of the four classifiers was carried out. From the experimental results, it can be seen that with 98.3 percent, the SVM classifier provides the highest accuracy than the others, while the average accuracy of 96.55 percent, 94.8 percent, and 98.1 percent can be provided by Logistic, Decision Tree, and KNN.

Sinha et. al. [3] Chronic Kidney Disease was predicted using SVM and KNN classifiers. The help of vector machine SVM and KNN classifier on the prediction of chronic kidney disease (CKD) accuracy, precision, and executed time. The efficiency of the KNN classifier was found to be higher than SVM in the experimental results. The SVM (Support Vector Machine) accuracy was 73.75% and the K-Nearest Neighbour (KNN) accuracy was 78.75%.

Sharma et. al. [4] evaluated of various classification techniques for Chronic Kidney Disease (CKD) Diagnosis. There used 2400 instances and 24 attributes in the dataset used for the analysis. By applying them to the CKD results, the authors assessed twelve classification techniques. To measure efficacy, the results of the prediction were compared with the real medical outcomes of the subject by candidate strategies. Prognostic accuracy, precision, specificity, and sensitivity are the diverse metrics used for performance analysis. The findings show that the decision-tree achieved 98.6 percent accuracy, 97.20 sensitivity, one precision, and one specificity.

Khan et. al. [5] uesd Machine Learning (ML) techniques for Chronic Kidney Disease(CKD) Prophecy. A dataset that was taken from the 400 Instances UCI ML repository. NBTree, J48, Support Vector Machine, Logistic Regression, Multi-layer Perceptron and Naïve Bayes were the ML algorithms. The results of the techniques listed were contrasted with the characterization of the most accurate technique that brings high precision to the classification of patients with CKD and NOT CKD. The experimental results used accuracy 96.50% for LR, 95.75% for NB, 97.25% for MLP, 98.25% for SVM, 97.75% for J48, 98.75% for NBTree.

Polat et. al. [6] predicted the result of the diagnosis of CKD. To scale back the dimension of the Chronic Kidney Disease dataset, two important kinds of feature selection techniques were selected, namely, wrapper and filter approaches. The classifier subset evaluator with the greedy stepwise program and the wrapper subset evaluator with the simplest First program was used inside the wrapper method. Correlation function selection subset evaluator with the greedy stepwise program and filtered sub-set evaluator with the simplest inside the filter approach. They used the main software. The results showed that the Support Vector Machine classifier features a higher accuracy rate (98.5 percent) in the diagnosis of CKD compared to other selected methods by using a filtered subset evaluator with the simplest First Software feature selection process.

Radha et. al. [7] predicted the classification algorithms had been considered for the diagnosis of chronic kidney disease. On completely distinct algorithms such as Naive Bayes, Decision Tree, K Nearest Neighbor, and Support Vector Machine, the expected results were carried out. The expected outcome reveals that the K-Nearest Neighbour algorithm offers a better result than the other classification algorithms and generates an accuracy of 98 percent.

Serpen et. al. [8] predicted the result of diagnosis rule extraction from patient data for Chronic Kidney Disease (CKD). To formulate a bunch of designation rules for CKD, the C4.5 decision tree rule was applied to the patient data. For a total patient count of 400, they properly classified 393 patients and incorrectly classified seven cases. The 3-fold cross-validation C4.5 algorithm achieved a predicted accuracy of 98.25%.

Ahmad et. al. [9] the results predicted whether patients with kidney disease have entered a phase of chronic kidney disease or not. The methodology of this study consisted of 2 main phases: modeling of classification and development of the system. Modeling of classification consists of collecting information, preparing data, grouping data, classification, extracting rules. It was expected that this device would help the doctor in deciding on the chronic condition of patients with kidney disease with good precision. This system's accuracy was 98.34 percent.

Hosseinzadeh et. al. [10] diagnostic expected model for chronic nephrosis in net of things platform. The classification methods used consisted of a decision tree (J48), strategies for SVM, Naïve Bayes(NB), and MLP. The experimental results show that, compared to Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Naïve Bayes classifiers, the applied dataset with the expected elected options produces 97 percent accuracy, 99 percent sensitivity, and 95 percent specificity through the application of decision tree (J48) classifier. Also, the planned feature set will improve the execution time as compared to different datasets with totally different options.

Comparative analysis from existing method

TABLE 1: SOME COMPARATIVE ANALYSIS OF PREVIOUS MODEL AND OUR MODEL

Author	Model	Accuracy
Amirgaliyev [1]	SVM	93.1 %
Charleonnan [2]	SVM,KNN & LR	98.3%,96.55% & 94.8%
Sinha [3]	SVM & KNN	73.75% & 78.75%
Sharma [4]	Decision-tree	98.6%
Khan [5]	NB,LR,MLP,J48,SVM & NBTree	95.75%,96.50%,97.25%,97.75%, 98.25% & 98.75%
Polat [6]	SVM	98.5%
Radha [7]	SVM & KNN	98% better than KNN
Serpen [8]	Decision tree	98.25%
Ahmad [9]	Classification modeling	98.34%
Hosseinzadeh [10]	SVM, Decision tree (J48), MLP, and NB strategies	97%
Our Model	Random forest classifier	99%

CHAPTER 3
METHODOLOGY

This chapter is divided into a few parts. Let's discuss about this

3.1 Data Description

In this paper, we use 400 chronic kidney disease datasets from the UCI machine learning repository. In this dataset, there are 25 attributes, 24 are predictive variables and 1 attribute is decision class. Now we discuss briefly our all attributes.

[1] Age: This attribute represents the age in years of a person. It is a numeric value and it is a predictive variable.

[2] Bp: This attribute means blood pressure and it is a numeric value and it is a predictive variable.

[3] Sg: This attribute means specific gravity and its value is numeric. It is a predictive variable.

[4] Al: This attribute means albumin. its value is numeric. It is a predictive variable.

[5] Su: This attribute means sugar. its value is numeric. It is a predictive variable

[6] Rbc: This attribute means red blood cells. its value is nominal. It is a predictive variable.

[7] Pc: This attribute means pus cell. its value is nominal. It is a predictive variable.

[8] Pcc: This attribute means pus cell clumps. its value is nominal. It is a predictive variable.

[9] Ba: This attribute means bacteria. its value is nominal. It is a predictive variable.

[10] Bgr: This attribute means blood glucose random. its value is numeric. It is a predictive variable.

[11] Bu: This attribute means blood urea. its value is numeric. It is a predictive variable.

[12] Sc: This attribute means serum creatinine. its value is numeric. It is a predictive variable.

[13] Sod: This attribute means sodium. its value is numeric. It is a predictive variable.

[14] Pot: This attribute means potassium. its value is numeric. It is a predictive variable.

[15] Hemo: This attribute means hemoglobin. its value is numeric. It is a predictive variable.

[16] Pcv: This attribute means packed cell volume. its value is numeric. It is a predictive variable.

[17] Wc: This attribute means white blood cell count. its value is numeric. It is a predictive variable.

[18] Rc: This attribute means red blood cell count. its value is numeric. It is a predictive variable.

[19] Htn: This attribute means hypertension. its value is nominal. It is a predictive variable.

[20] Dm: This attribute means diabetes mellitus. its value is nominal. It is a predictive variable.

[21] Cad: This attribute means coronary artery disease. its value is nominal. It is a predictive variable.

[22] Appet: This attribute means appetite. its value is nominal. It is a predictive variable.

[23] Pe: This attribute means pedal edema. its value is nominal. It is a predictive variable.

[24] Ane: This attribute means anemia. its value is nominal. It is a predictive variable.

[25] Class: It is the response attribute. A person having chronic kidney disease or not having chronic kidney disease are given here. And this variable is a nominal type.

3.2 Algorithm Description

3.2.1 Naive Bayes:

A set of Bayes' Theorem-based classification algorithms are Naive Bayes classifiers. This is not just a single algorithm, but an algorithm family where a common concept is shared by all of them.

Bayes' theorem is stated mathematically as the following equation,

$$□(□ / □) = \frac{□(□ / □)□(□)}{□(□)}$$

where A and B are events, P(A) is the probability of event A and P(B) is the probability of event B.

Now, for the dataset, applying Bayes' theorem,

$$\square(\square \ / \square) = \frac{\square(\square \ / \square)\square(\square)}{\square(\square)}$$

where, y = class variable and X = dependent feature vector of size n.

Here,

$$\square = (\square_1, \square_2, \square_3, \ldots\ldots, \square_\square)$$

3.2.2 Logistic Regression:

Linear Regression analysis can't be applied if the relationship is nonlinear. In that case the Logistic Regression can be used. The equation ofLinear Regression is,

$$\square = \square_0 + \square_1\square_1 \ + \square_2\square_2 \ + \ldots\ldots + \square_\square\square_\square$$

Here, y = variable and $\square_1, \square_2 \ , \ldots\ldots, \square_\square$ = the predictor variables, α_0 = Intercept,

$\alpha_1, \square_2, \ldots\ldots, \square_\square$ = coefficient.

Applying sigmoid function on the equation, we get the logistic function,

$$\square(\square) = \frac{1}{1+\square^{\square_0+\square_1\square_1 \ +\square_2\square_2 \ +\ldots\ldots+\square_\square\square_\square}}$$

3.2.3 MLP:

The multi-layer perceptron MLP classifier is a feed-forward artificial neural network model that maps input data sets to a compatible output range. For any feed-forward ANN, the term MLP is used ambiguously, often loosely, often explicitly to refer to networks

consisting of multiple layers of perceptrons. MLP has one or even more intermediate layers or hidden layers.

The $□^{□h}$ output node error in the data point n can be observed, which is described by the following equation, where d and c indicate the real and predicted values, both.

$$□_□(□) = □_□(□) − □_□(□)$$

3.2.4 SGD Classifier:

Stochastic Gradient Descent (SGD) is a very effective approach to fitting convex loss functions with linear classifiers and regressors. It needs a number of hyper parameters.

The algorithm reviews the training examples. For each example it need to updates the model parameters according to the update rule which is,

$$□ ← □ − \frac{□}{□}\sum_{□=1}^{□} □□_□(□)$$

Here η is the learning rate, w is weight vector which lies in the x,y plane, $□_□$ is the loss function of i-th observation.

3.2.5 Adaboost:

AdaBoost refers to the Adaptive Boosting classifier. A set of weak classifiers are merged into a strong one using this technique. Here the equation represents the usage of the AdaBoost algorithm for classification.

$$□(□) = □□□□(\sum_{□=2}^{□} □_□□_□(□))$$

Here, the $□^{□h}$ is a weak classifier which is represented by $□_□$ and $\theta_□$ which refers to the corresponding weight.

3.2.6 Bagging Classifier:

The bagging classifier is a meta-estimator that is ensembled. It fits the random subsets of the original dataset with each base classifier. Then aggregates their predictions to form a final prediction. Usually, it is used as a way to reduce a black-box estimator's variance.

Predictions for unseen samples x' can be made after training by combining the predictions on x' from all the individual regression trees:

So, predicting function $\hat{□} = \frac{1}{□}\sum_{□=1}^{□} □_□(□')$

Where B is the number of trees which is a free parameter.

3.2.7 Decision Tree Classifier:

Tree-like structures are used in Decision Tree classification. The root nodes means conditions and child nodes means the class label. Branches of the root nodes means effects of the conditions. The Entropy E(S) can be represented as

$$□(□) = \sum_{□=1}^{□} □_□ □□□_2 □_□$$

Where the probability of the jth class is $□_□$.

3.2.8 Random Forest Classifier:

It is a learning algorithm based on an ensemble tree. A selection of decision trees from the randomly chosen training set subset is the Random Forest Classifier. To evaluate the final class it takes votes from the different decision trees.

Using the mean squared error which is MSE while using the Random Forest Algorithm to fix regression problems.

$□□□ = \frac{1}{□}\sum_{□=1}^{□} (□_□ - □_□)$

Here N = number of data points, fi = value that the model returns and yi = the actual value of the data.

For Gini index,

$$\text{Gini} = 1 - \sum_{i} (p_i)^2$$

Here, pi means the relative frequency of the class

We can also use entropy to determine how nodes branch in a decision tree.

Here E is entropy, is the class of ith.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

3.3 Proposed Model

The procedures for execution are illustrated in this section. Python and the Scikit-learn library have been used to carry out the research.

In figure 1, generate a diagram for the procedure of the proposed model.

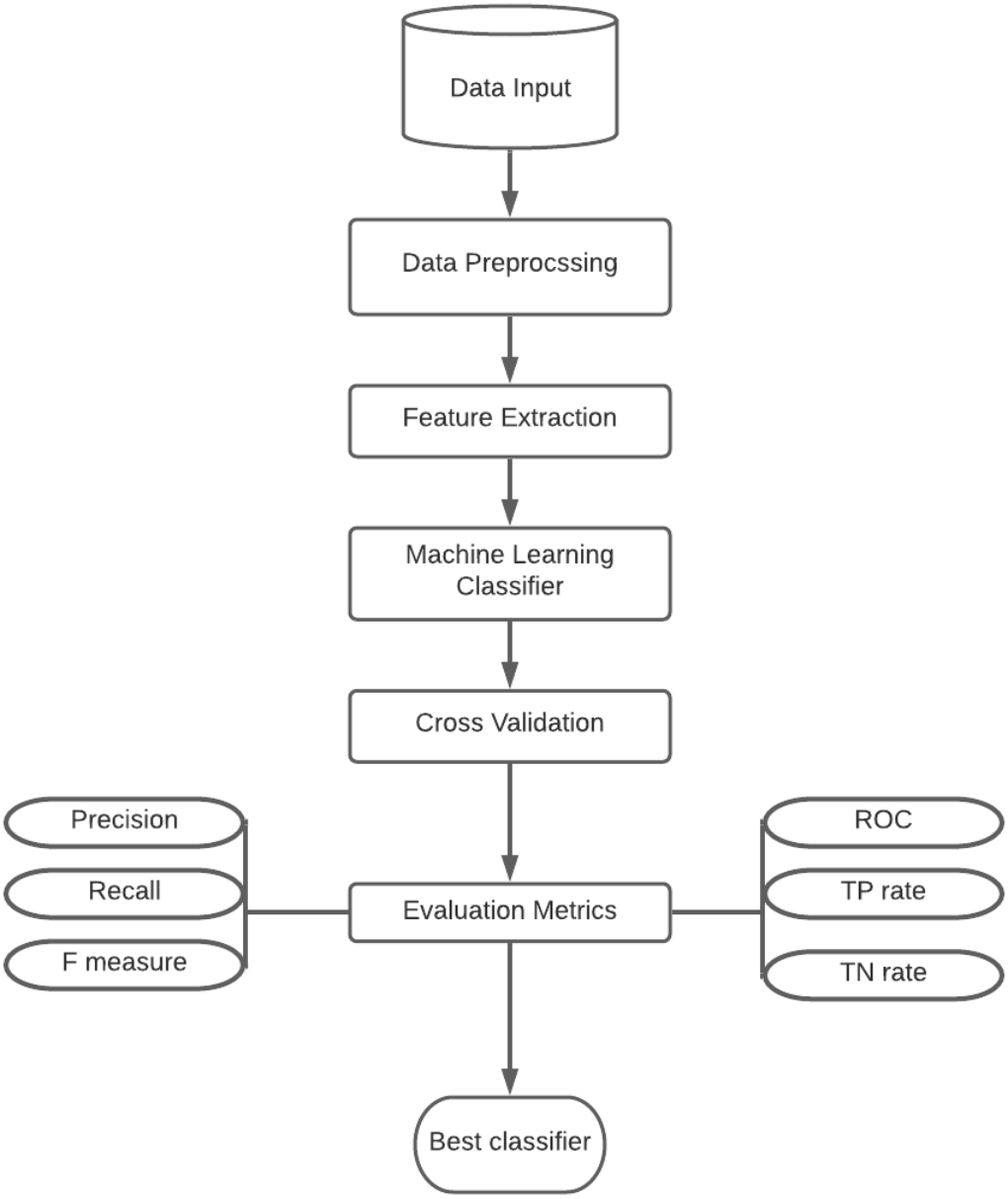


Figure 1: Implementation Procedure of Chronic Kidney Disease

3.3.1 Input Data

In this research total 400 patient data collected to predict an analysis and find out performance. This dataset collected from UCI machine learning repository

3.3.2 Data Prepossessing

In this dataset, total 25 attributes are used, 24 are predictive variables and 1 is response variable. In predictive 24 attributes some attributes are nominal some attributes are numerical. Therefore convert nominal attributes into numerical attributes using a mapping function. In this dataset, the nominal value contains rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane this attribute we convert this a numeric value using a mapping function. And now our data set is full of numeric value. Moreover splitting the dataset into 80% for training and 20% for testing use.

3.3.3 Feature Extraction

In this section, from fig 2 shown feature extraction using principal component analysis. In principle component analysis we colour into two types red and green. The red colour represent chronic kidney disease(CKD) and the green colour represent no chronic kidney disease(NOCKD)

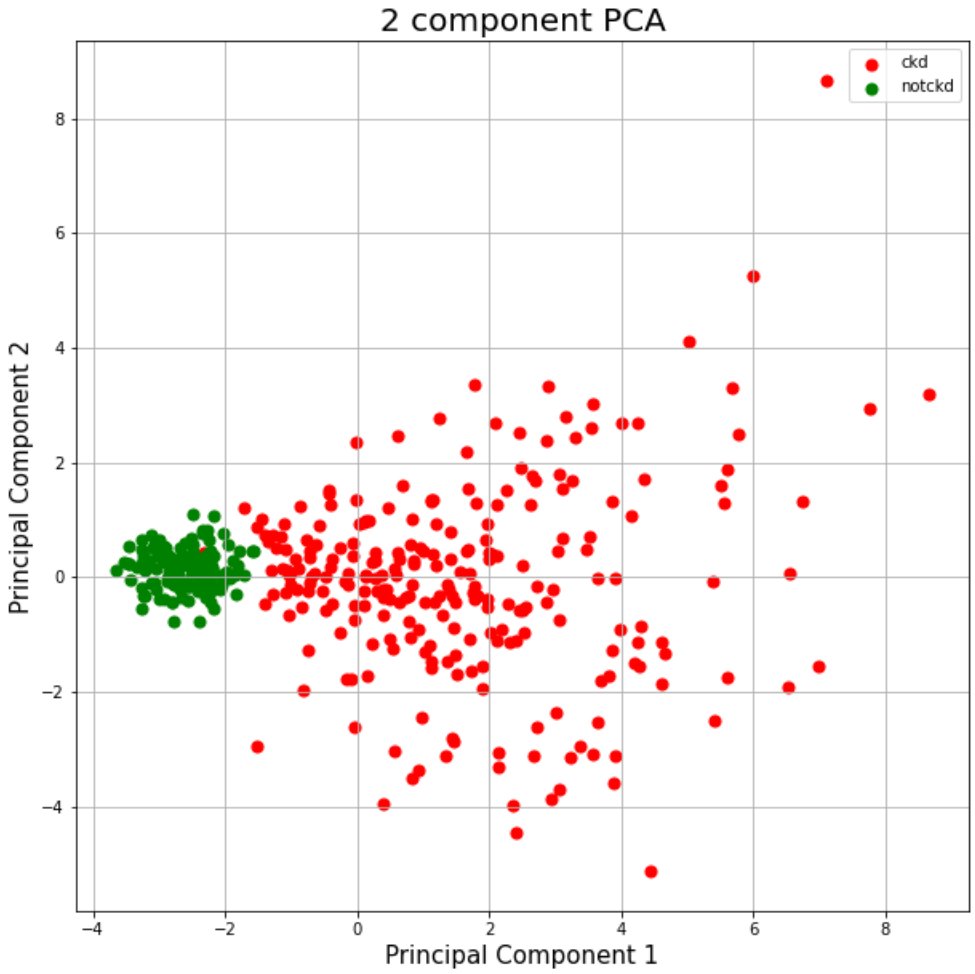


Figure 2: Feature Extraction using Principal Component Analysis

In figure 3, correlation metrics are shown using target attribute to feature attribute

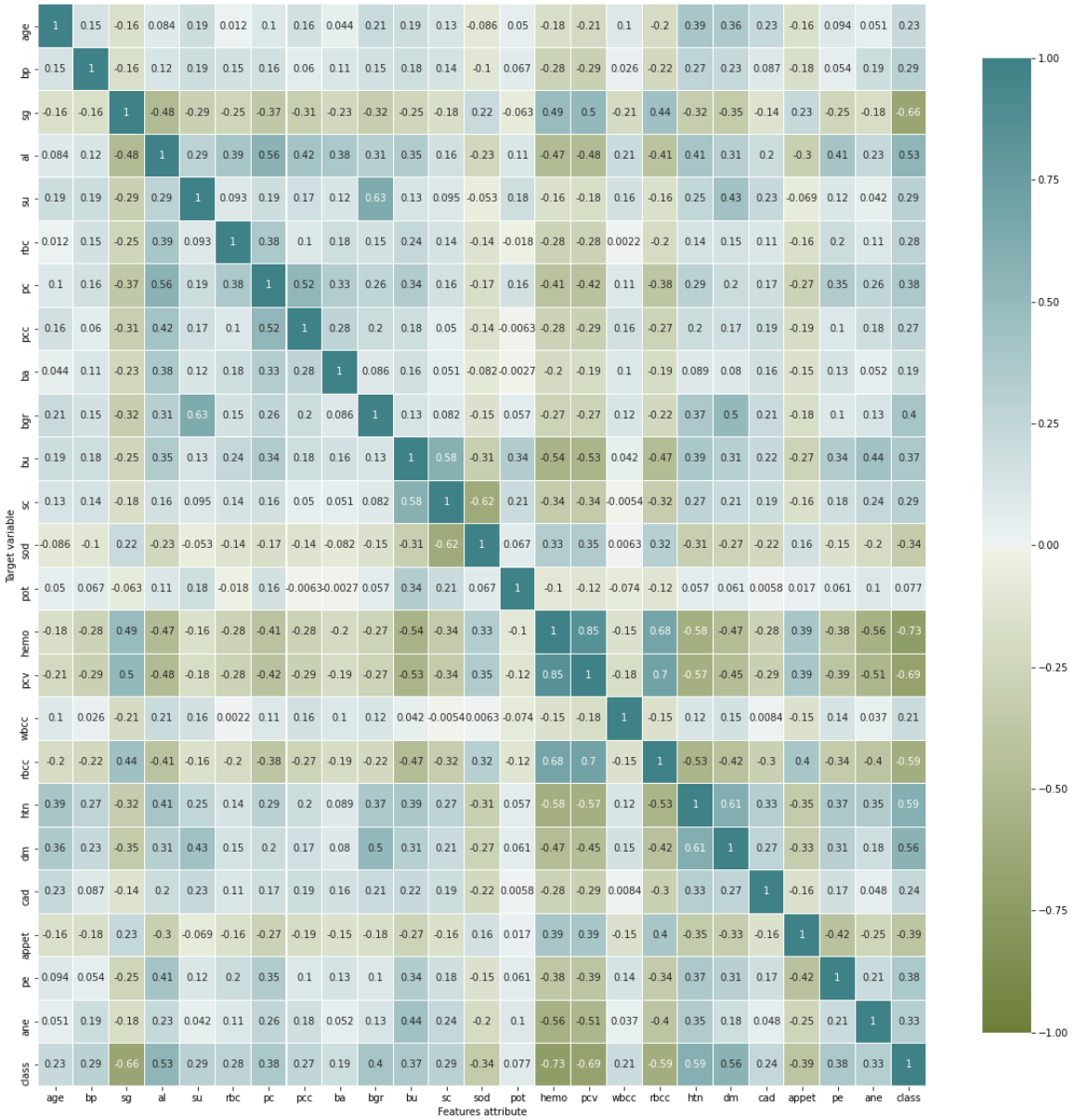


Figure 3: Correlation Metrics among Target to Feature Attribute

3.3.4 Machine Learning Classifier

We use 8 machine learning algorithms to evaluate the chronic kidney disease(CKD) or not chronic kidney disease(NO CKD) performance. This algorithms are: Logistic Regression(LG), Naive Bayes(NB), Multilayer Perceptron(MLP), Stochastic Gradient

Descent(SGD), Adaptive Boosting(Adaboost), Bagging, Decision Tree(DT), and Random Forest(RF).

3.3.5 Cross-Validation

Cross-validation is a technique to test machine learning models by training multiple machine learning techniques on the available input data subsets and evaluating them on the complementary data subset. To detect overfitting, use cross-validation, eg, filling to generalize a sequence. Comparing and selecting a model or a given predictive modeling problem is commonly used in applied machine learning because it is simple to understand, easy to implement, and results in ability estimates that usually have a lower bias than other approaches. As our data set is small amounts so we use k-fold cross-validation.

The data is split into k subsets in K Fold cross-validation. Now the holdout solution is repeated k times such that every time the test set/validation is used as one of the k subsets and all the other k-1 subsets are combined to form a training set. In all k experiments, the error estimate is summed for the maximum efficacy of our model. As can be shown, each data point is checked exactly once and is trained in a k-1 times series. As we use most of the fitting data, this decreases bias considerably and also considerably reduces variance, as most data are also used in validation sets. Another drawback of this approach is that the interchange of training sets and evaluation sets. K = 5 or 10 is normally optimal as a general rule and an observational proof, but nothing is set and any meaning can be taken.

We use k(10) fold cross-validation to find out our evaluation metrics.

3.3.6 Evaluation Metrics

In this section, we find out some important results for our study. Here we briefly describe those.

Precision

Precision or positive predictive value = $\frac{\square\square\square\square\ \square\square\square\square\square\square\square\square}{\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square}$

Recall:

$$\text{Recall} = \frac{\square\square\square\square\ \square\square\square\square\square\square\square\square}{\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square}$$

F Measure

$$\text{F-1 score} = \frac{2 * \square\square\square\square\ \square\square\square\square\square\square\square\square}{2 * \square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square}$$

Accuracy:

$$\text{Acc} = \frac{\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\ \square\square\square\square\square\square\square\square}{\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square + \square\square\square\square\square\ \square\square\square\square\square\square\square\square}$$

3.3.7 Best Model

The degree of precision of the data set is also calculated to generate accuracy of various types of algorithms and the best one can be achieved by using a Random Forest classification.

3.3.8 Confusion Metrics

This figure 4 represents the confusion metrics structure.

1. TP means true positive = when a label is predicted correctly, we predicted it is CKD and it is the same.
2. FP means false positive = when a label predicted falsely, we predicted it is CKD but it is not CKD.
3. FN means false-negative = when a predicted label is missing, we predicted it is not CKD but it is CKD.
4. TN means true negative = when a label is correctly predicted by another label, we predict it is not CKD and in the dataset, it is not CKD.

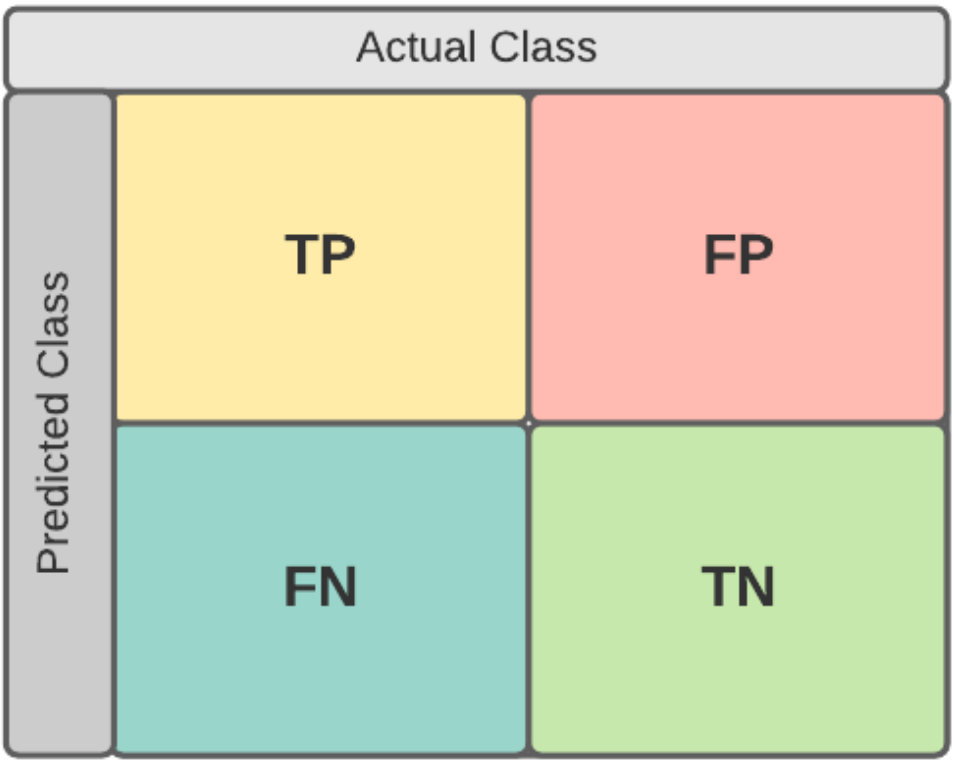


Figure 4: Confusion Metrics Structure

CHAPTER 4

RESULT ANALYSIS

4.1 Confusion Metrics Analysis

4.1.1 Confusion Metrics Analysis using Cross Validation

The confusion metrics analysis are shown in table II for cross validation, therefore confusion metrics analysis is an important part of this research. To achieve good accuracy predicted positive, predicted negative is a more important attribute. In this paper, Random Forest gives the best outcome from any other classifiers.

Table II: Confusion matrix for cross-validation

Model	Label	Predicted Positive	Predicted Negative
Naive Bayes	NO CKD CKD	229 1	21 149
Logistic	NO CKD CKD	239 7	11 143
MLP	NO CKD CKD	241 1	9 149
SGD	NO CKD CKD	243 0	7 150
AdaBoost	NO CKD CKD	243 8	7 142
Bagging	NO CKD CKD	246 7	4 143
Decision tree	NO CKD CKD	249 9	1 141
Random forest	NO CKD CKD	249 3	1 147

4.1.2 Confusion Metrics Analysis using 30% Test Data

The confusion metrics analysis are shown in table II for 30% test data, therefore confusion metrics analysis is an important part of this research. To achieve good accuracy predicted positive, predicted negative is a more important attribute. In this paper, Random Forest gives the best outcome from any other classifiers.

TABLE III: CONFUSION MATRIX FOR 30% TEST DATA

Model	Label	Predicted Positive	Predicted Negative
Naive Bayes	NO CKD CKD	50 10	0 60
Logistic	NO CKD CKD	44 7	6 63
MLP	NO CKD CKD	47 17	4 53
SGD	NO CKD CKD	0 0	50 70
AdaBoost	NO CKD CKD	50 2	0 68
Bagging	NO CKD CKD	50 1	0 69
Decision tree	NO CKD CKD	48 3	2 67
Random forest	NO CKD CKD	49 0	1 70

4.2 Evaluation Metrics Analysis

4.2.1 Evaluation Metrics Analysis for Cross Validation

The perform analysis shown in table IV using k fold cross validation from different machine learning classifiers namely: Logistic Regression(LG), Naive Bayes(NB), Multilayer Perceptron(MLP), Stochastic Gradient Descent(SGD), Adaptive Boosting(Adaboost), Bagging, Decision Tree(DT), Random Forest(RF). Moreover the

highest value from the random forest(RF) classifier and its accuracy is 99%. The second highest accuracy achieved from MLP, SGD, and Decision Tree classifiers and value are 98%. The lowest accuracy we get from the Naive Bayes(NB) classifier and value is 95%. Second lowest accuracy we gain from Logistic Regression classifier and it is 96%.

TABLE IV: EVALUATION METRICS FOR CROSS-VALIDATION

Model	Accuracy	Label	Precision	Recall	f1-score	ROC	TP rate	FP rate
Naive Bayes	95	NO CKD CKD	0.99 0.87	0.91 0.99	0.95 0.93	0.99	0.91 0.99	0.00 0.08
Logistic	96	NO CKD CKD	0.97 0.92	0.95 0.95	0.96 0.94	0.98	0.95 0.95	0.04 0.04
MLP	98	NO CKD CKD	0.99 0.94	0.96 0.99	0.98 0.96	0.99	0.96 0.99	0.00 0.03
SGD	98	NO CKD CKD	1.00 0.95	0.97 1.00	0.98 0.97	0.98	0.97 1.00	0.00 0.02
Adaboost	96	NO CKD CKD	0.96 0.95	0.97 0.94	0.97 0.95	0.99	0.97 0.94	0.05 0.02
Bagging	97	NO CKD CKD	0.97 0.97	0.98 0.95	0.97 0.96	0.98	0.98 0.95	0.04 0.01
Decision Tree	98	NO CKD CKD	0.96 0.99	0.99 0.94	0.98 0.96	0.97	0.99 0.94	0.06 0.00
Random forest	99	NO CKD CKD	0.98 0.99	0.99 0.98	0.99 0.98	0.98	0.99 0.98	0.02 0.00

4.2 Evaluation Metrics Analysis for 30% Test Data

The perform analysis shown in table IV using k fold cross validation from different machine learning classifiers namely: Logistic Regression(LG), Naive Bayes(NB),

Multilayer Perceptron(MLP), Stochastic Gradient Descent(SGD), Adaptive Boosting(Adaboost), Bagging, Decision Tree(DT), Random Forest(RF). Moreover the highest value from the random forest(RF) classifier and its accuracy is 99% Bagging classifier also gives 99% accuracy but its AUC value is lower than Random Forest classifier.. The second highest accuracy achieved from Adaboost 98%. The lowest accuracy we get from the SGD classifier and value is 58%.

TABLE V: EVALUATION METRICS FOR 30% TEST DATA

Model	Accuracy (%)	Label	Precision (%)	Recall(%)	f1-score (%)	AUC (%)
Naive Bayes	91	NO CKD CKD	83 100	100 86	91 92	95
Logistic	89	NO CKD CKD	89 91	88 90	87 91	96
MLP	83	NO CKD CKD	73 93	92 76	81 83	86
SGD	58	NO CKD CKD	0 58	0 100	0 74	50
Adaboost	98	NO CKD CKD	96 100	100 97	98 99	99
bagging	99	NO CKD CKD	98 100	100 99	99 99	99
Decision Tree	95	NO CKD CKD	94 97	96 96	95 96	95
Random forest	99	NO CKD CKD	100 99	98 100	99 98	100

4.3 ROC Analysis

A ROC (receiver operating characteristic) curve is a graphical representation that shows a binary classifier device's diagnostic capability as its threshold of discrimination is different. The ROC curve assembles the TPR(true positive rate) from various threshold settings against the FPR (false positive rate).

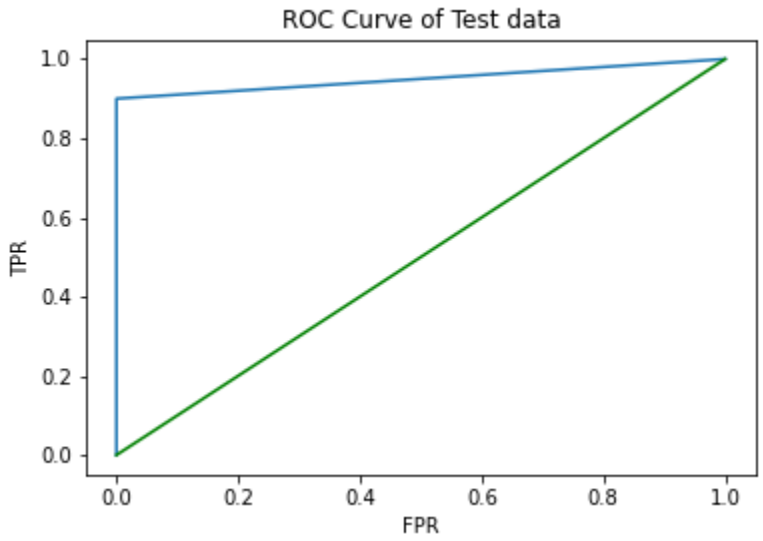


Figure 5: ROC curve for Naive Bayes classifier

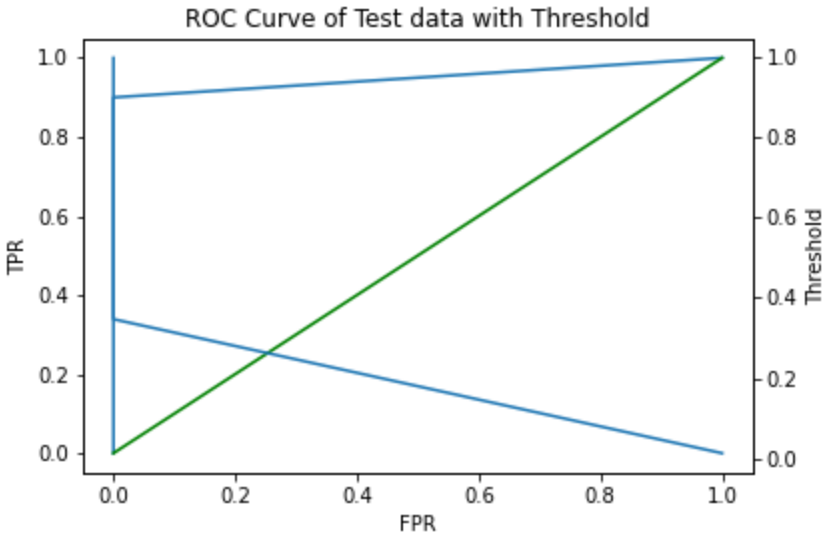


Figure 6: Threshold for Naive Bayes classifier

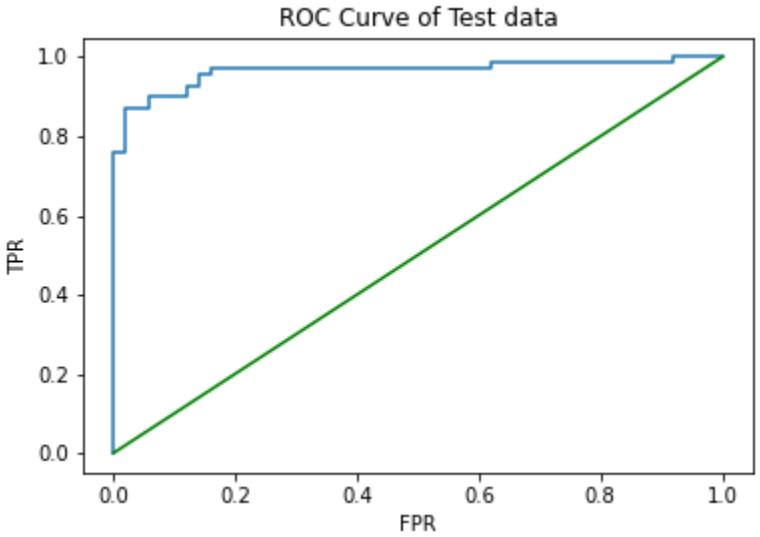


Figure 7: ROC curve for LR classifier

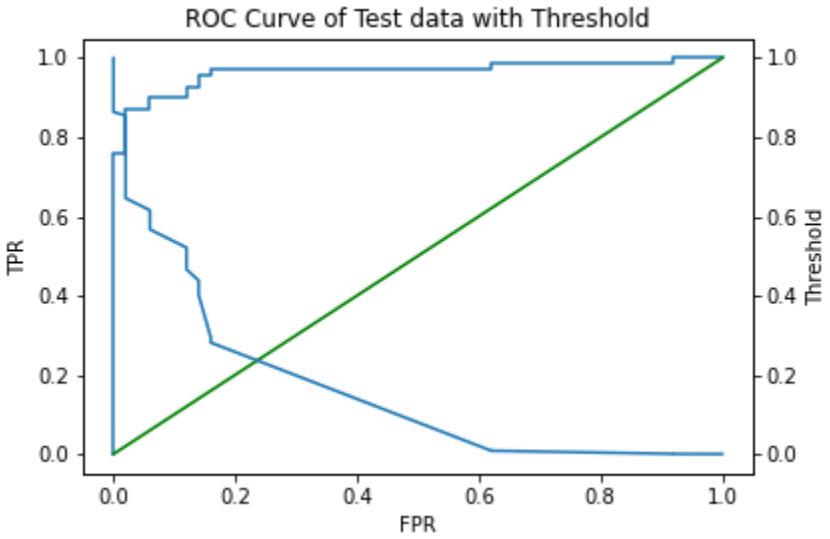


Figure 8: Threshold for LR classifier

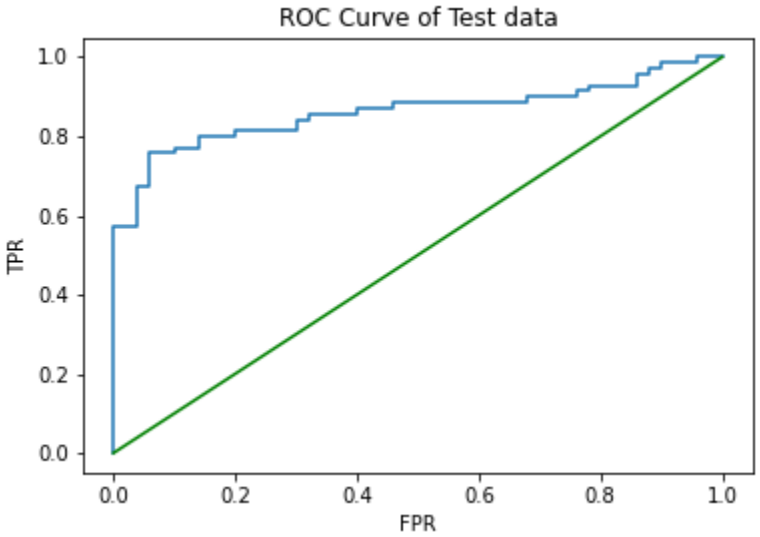


Figure 9: ROC curve for MLP classifier

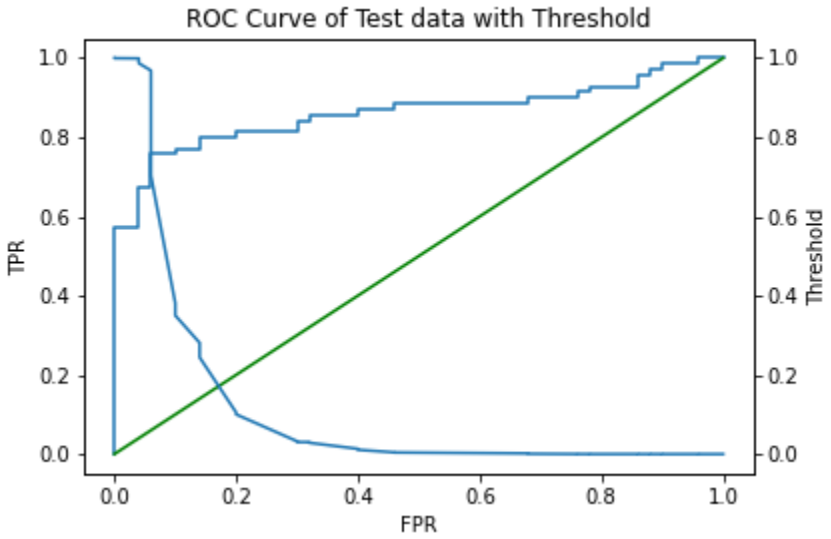


Figure 10: Threshold for MLP classifier

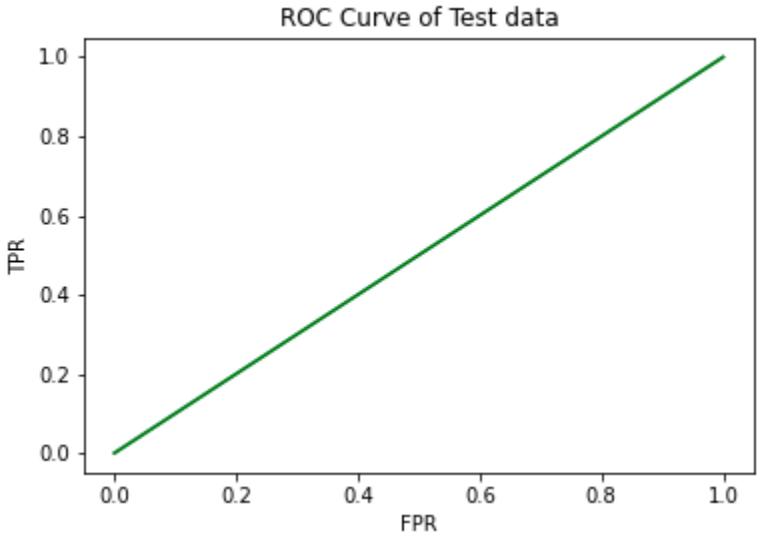


Figure 11: ROC curve for SGD classifier

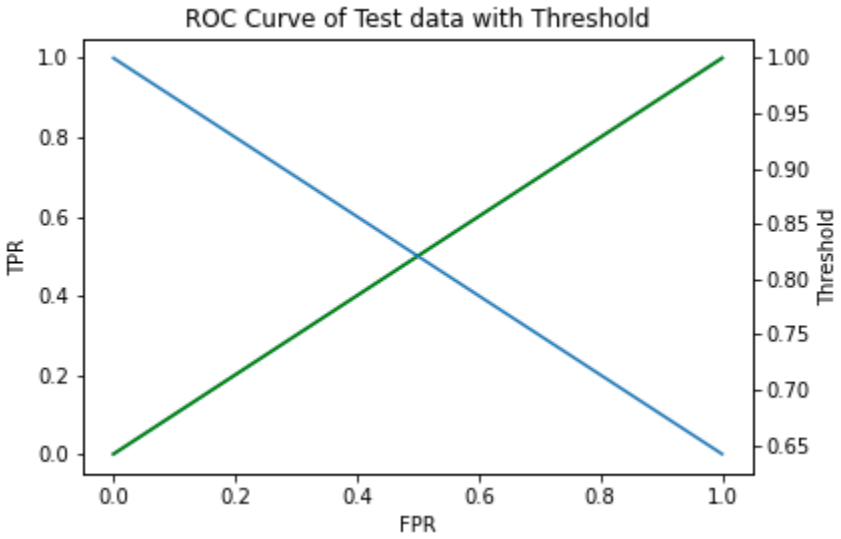


Figure 12: Threshold for SGD classifier

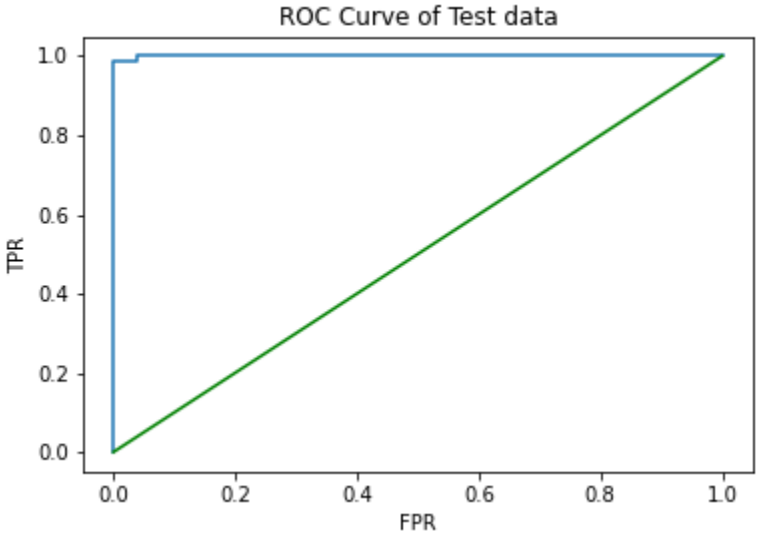


Figure 13: ROC curve for AdaBoost classifier

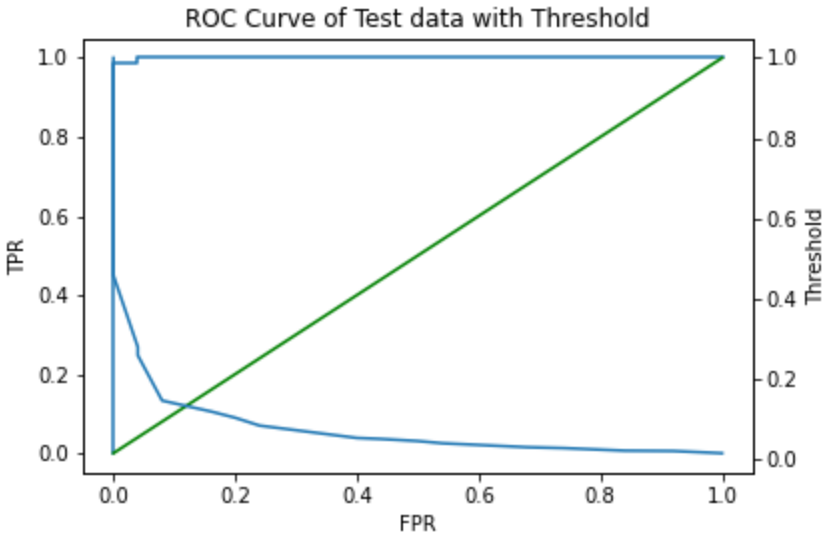


Figure 14: Threshold for Adaboost classifier

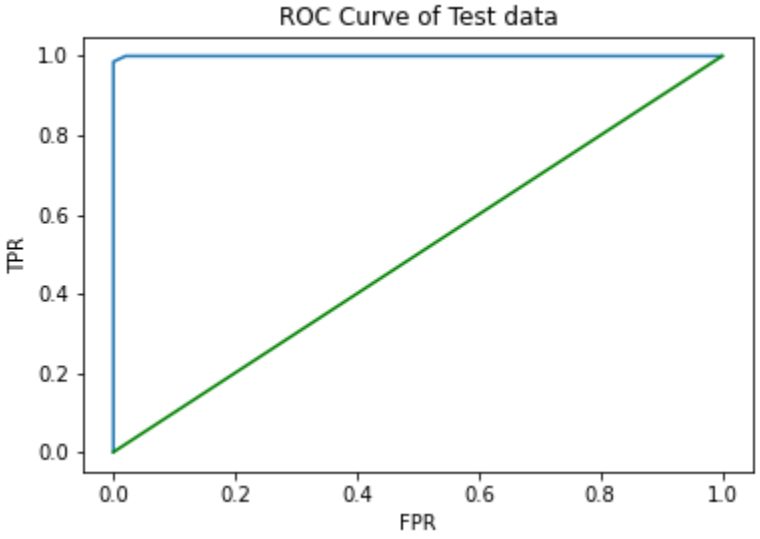


Figure 15: ROC curve for Bagging classifier

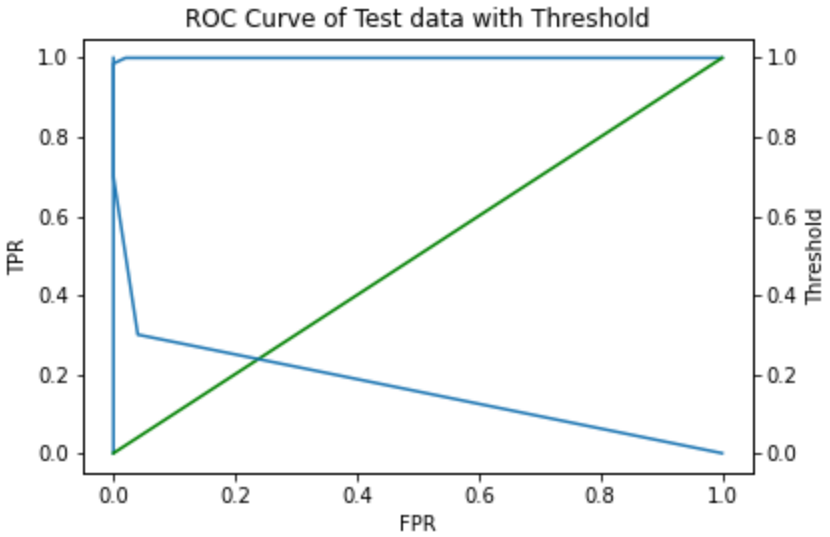


Figure 16: Threshold for Bagging classifier

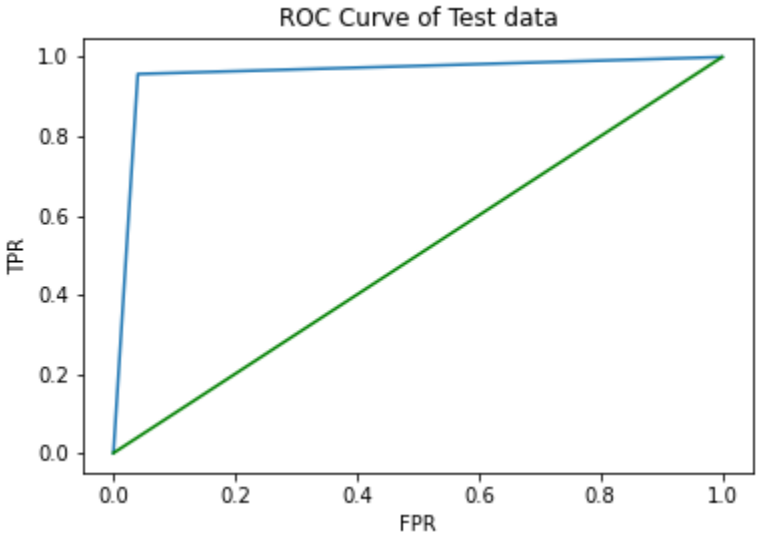


Figure 17: ROC curve for Decision Tree classifier

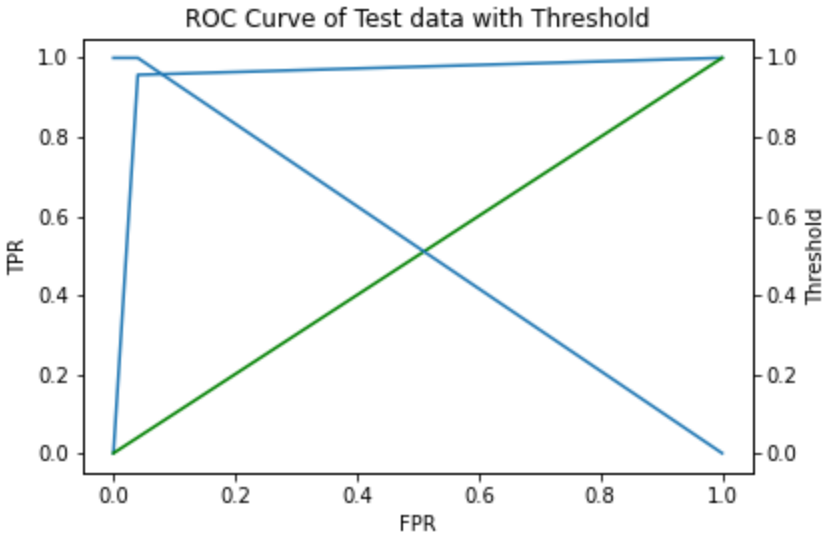


Figure 18: Threshold for Decision Tree classifier

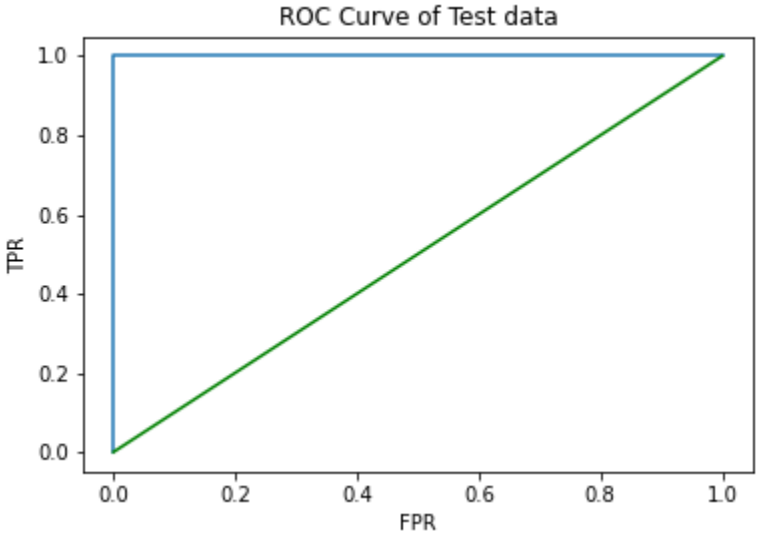


Figure 19: ROC curve for Random Forest classifier

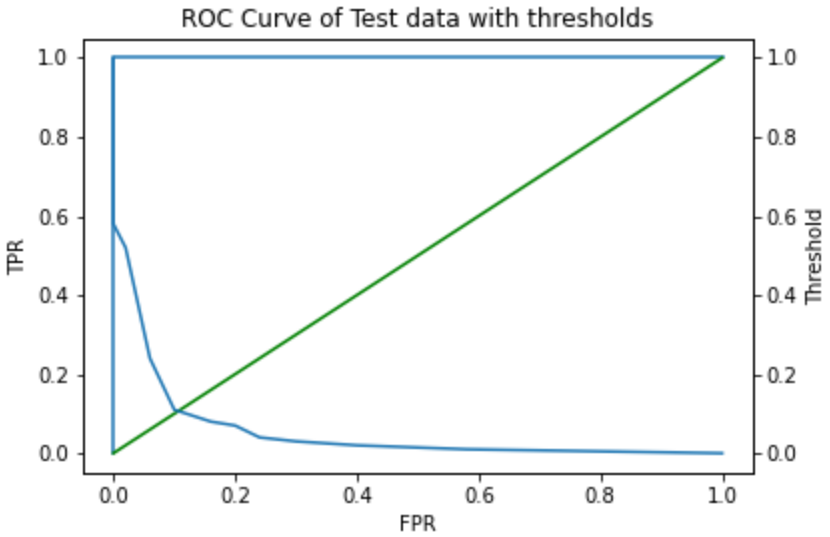


Figure 20: Threshold for Random Forest classifier

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

Chronic kidney disease(CKD) describes the significant loss of kidney function, also called chronic kidney failure.The waste and excess blood fluids are filtered from the blood and then excreted into the urine by the kidneys. When chronic kidney disease comes to an advanced stage, harmful amounts of fluid, electrolytes, and waste can build up in the body. The risk of end-stage renal disease (ESRD) and cardiovascular disease is increased by CKD, and other risk factors for heart disease are also present in individuals with CKD, such as elevated blood lipids. In individuals with CKD, the most common cause of death is cardiovascular disease rather than kidney failure. Kidney disease or chronic kidney disease is not a minor issue for a person's health. It will be a death cause for a person, so everyone needs to be concerned about his health preventing this in the very early stage of his or her life.

In this paper, several machine learning classifiers have been used to find out best accuracy, ROC, precision, recall and f measure. However random forest gives the 99% accuracy and its ROC value is 100.

5.2 Future Work

Data is everything for a data mining project is everything. More data, more secret patterns, and more dimensions will be provided with more data. There were 400 records in this dataset, which is not enough for a better prediction. More information can be applied to this dataset in the future and the the most significant variety of data would be a plus point for further analysis. Moreover deep learning will be used and image processing or relevant technique will be implemented in this work.

REFERENCE

[1] Schmidt, M. I., Duncan, B. B., Sharrett, A. R., Lindberg, G., Savage, P. J., Offenbacher, S., & ARIC investigators. (1999). Markers of inflammation and prediction of diabetes mellitus in adults (Atherosclerosis Risk in Communities study): a cohort study. The Lancet, 353(9165), 1649-1652.

[2] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning meth Vineet Maheshwari. (2019, January 05). Retrieved from https://medium.com/datadriveninvestor/ensemble-learning-9e5924fd6567

[3] Your Kidneys & How They Work. Retrieved from https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-theywork last accessed on November 05, 2020 at 13.05.

[4] How Your Kidneys Work. Retrieved from https://www.kidney.org/kidneydisease/howkidneyswrk last accessed on November 05, 2020 at 13.30.

[5] How They Work. Retrieved from https://www.webmd.com/ato-z-guides/function-kidneys last accessed on November 05, 2020 at 14.02

[6] Wikipedia contributors. Chronic kidney disease. In Wikipedia, The Free Encyclopedia https://en.wikipedia.org/w/index.php last accessed on December 05, 2020 at 08.05.

[7] Wikipedia contributors. Coronary arteries. In Wikipedia, The Free Encyclopedia from https://en.wikipedia.org/w/index.php last accessed on November 05, 2020 at 15.11.

[8] Amirgaliyev, Y., Shamiluulu, S., & Serek, A. (2018, October). Analysis of chronic kidney disease dataset by applying machine learning methods. In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE.

[9] Charleonnan, A., Fufaung, T., Niyomwong, T., Chokchueypattanakit, W., Suwannawach, S., & Ninchawee, N. (2016, October). Predictive analytics for chronic kidney disease using machine learning techniques. In 2016 Management and Innovation Technology International Conference (MITicon) (pp. MIT-80). IEEE.

[10] Sinha, P., & Sinha, P. (2015). Comparative study of chronic kidney disease prediction using KNN and SVM. International Journal of Engineering Research and Technology, 4(12), 608-12.

[11] Sharma, S., Sharma, V., & Sharma, A. (2016). Performance based evaluation of various machine learning classification techniques for chronic kidney disease diagnosis. arXiv preprint arXiv:1606.09581.

[12] Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. (2020). An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy. IEEE Access, 8, 55012-55022.

[13] Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. Journal of medical systems, 41(4), 55.

[14] Radha, N., & Ramya, S. (2015). Performance analysis of machine learning algorithms for predicting chronic kidney disease. Int. J. Comput. Sci. Eng. Open Access, 3, 72-76.
[15] Serpen, A. A. (2016). Diagnosis rule extraction from patient data for chronic kidney disease using machine learning. International Journal of Biomedical and Clinical Engineering (IJBCE), 5(2), 64-72.
[16] Ahmad, M., Tundjungsari, V., Widianti, D., Amalia, P., & Rachmawati, U. A. (2017, November). Diagnostic decision support system of chronic kidney disease using support vector machine. In 2017 Second International Conference on Informatics and Computing (ICIC) (pp. 1-4). IEEE.
[17] Hosseinzadeh, M., Koohpayehzadeh, J., Bali, A. O., Asghari, P., Souri, A., Mazaherinezhad, A., ... & Rawassizadeh, R. (2020). A diagnostic prediction model for chronic kidney disease in internet of things platform. Multimedia Tools and Applications, 1-18.
[18] Baldi, P., Brunak, S., & Bach, F. (2001). Bioinformatics: the machine learning approach. MIT press.
[19] Herbst, J. (2000, May). A machine learning approach to workflow management. In European conference on machine learning (pp. 183-194). Springer, Berlin, Heidelberg
[20] Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics, 26(9), 1169-1175.

APPENDICES

Abbreviation

CKD = Chronic Kidney Disease

ROC = Receiver Operating Characteristic

NB = Naive Bayes

MLP = Multilayer Perceptron

SGD = Stochastic Gradient Descent

AdaBoost = Adaptive Boosting

RFC = Request For Comments

DTC = Decision Tree Classifier

Appendix: Research Reflections

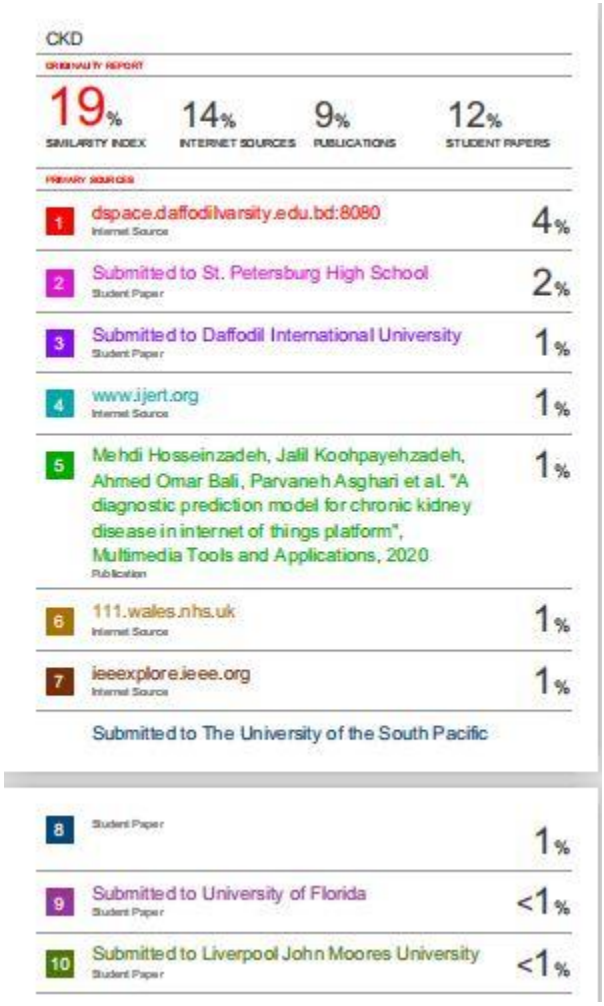
At the beginning of this research work, we had very little experience with machine learning algorithms about detection and prediction. Our supervisor was very kind and helpful to us. He gave us valuable guidance and helped us a lot. In this working time of research, we learned much new information, new techniques, use of new algorithms, and worked with different methods. At the time when we start working with this, we face many problems but gradually we become more and more familiar with these algorithms.

Finally, by doing the research we have gained courage and been inspired to do more research in the future.

International Publication: Scopus Index

Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches: Minhaz Uddin Emon, Al Mahmud Imran, Rakibul Islam, and Ohidujjaman: Accepted in 6th International Conference on Inventive Computatio n Technologies ICICT 2021, India, from January 20th - 22nd 2021.

PLAGIARISM REPORT



CKD

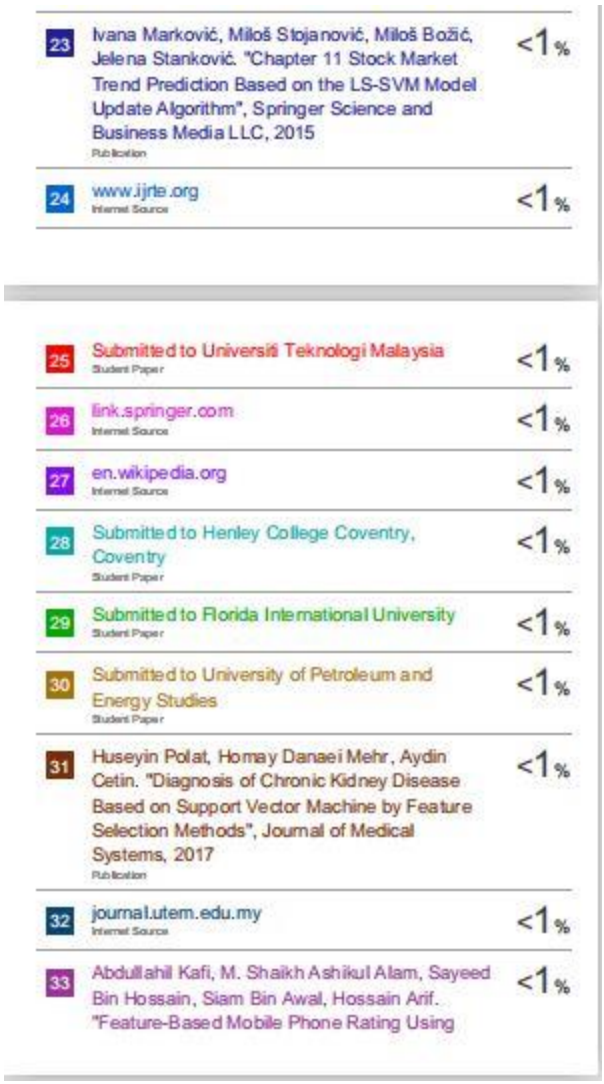
ORIGINALITY REPORT

19%	14%	9%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

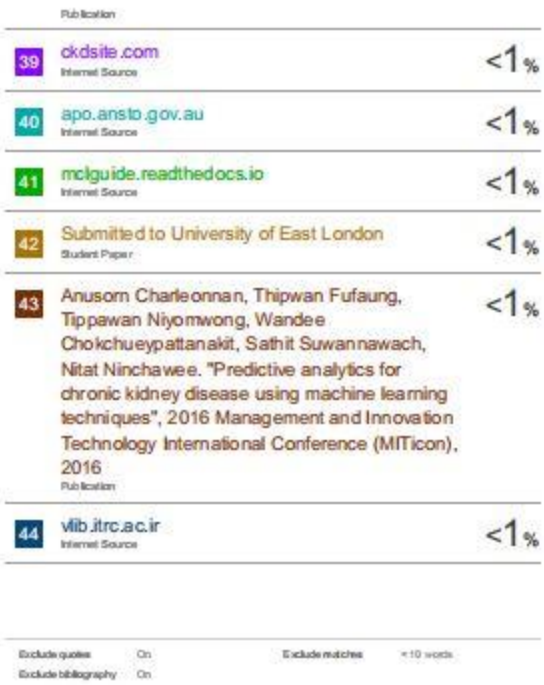
1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to St. Petersburg High School Student Paper	2%
3	Submitted to Daffodil International University Student Paper	1%
4	www.ijert.org Internet Source	1%
5	Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Ahmed Omar Bali, Parvaneh Asghari et al. "A diagnostic prediction model for chronic kidney disease in internet of things platform", Multimedia Tools and Applications, 2020 Publication	1%
6	111.wales.nhs.uk Internet Source	1%
7	ieeexplore.ieee.org Internet Source	1%
	Submitted to The University of the South Pacific	
8	Student Paper	1%
9	Submitted to University of Florida Student Paper	<1%
10	Submitted to Liverpool John Moores University Student Paper	<1%

11	"Advances in Artificial Intelligence and Data Engineering", Springer Science and Business Media LLC, 2021 Publication	<1%
12	Submitted to University of Warwick Student Paper	<1%
13	Submitted to Ain Shams University Student Paper	<1%
14	www.geeksforgeeks.org Internet Source	<1%
15	Alexander Arman Serpen. "Diagnosis Rule Extraction from Patient Data for Chronic Kidney Disease Using Machine Learning", International Journal of Biomedical and Clinical Engineering, 2016 Publication	<1%
16	library.iugaza.edu.ps Internet Source	<1%
	ijprjournals.com	
17	Internet Source	<1%
18	"Proceedings of International Joint Conference on Computational Intelligence", Springer Science and Business Media LLC, 2020 Publication	<1%
19	Submitted to Florida Atlantic University Student Paper	<1%
20	"Chronic Disease Prediction Using Effective Feature Selection", International Journal of Recent Technology and Engineering, 2019 Publication	<1%
21	Submitted to Institute of Health & Management Pty Ltd Student Paper	<1%
22	Submitted to Coventry University Student Paper	<1%



23	Ivana Marković, Miloš Stojanović, Miloš Božić, Jelena Stanković. "Chapter 11 Stock Market Trend Prediction Based on the LS-SVM Model Update Algorithm", Springer Science and Business Media LLC, 2015 Publication	<1%
24	www.ijrte.org Internet Source	<1%
25	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
26	link.springer.com Internet Source	<1%
27	en.wikipedia.org Internet Source	<1%
28	Submitted to Henley College Coventry, Coventry Student Paper	<1%
29	Submitted to Florida International University Student Paper	<1%
30	Submitted to University of Petroleum and Energy Studies Student Paper	<1%
31	Huseyin Polat, Homay Danaei Mehr, Aydin Cetin. "Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods", Journal of Medical Systems, 2017 Publication	<1%
32	journal.utem.edu.my Internet Source	<1%
33	Abdullahil Kafi, M. Shaikh Ashikul Alam, Sayeed Bin Hossain, Siam Bin Awal, Hossain Arif. "Feature-Based Mobile Phone Rating Using	<1%

	Sentiment Analysis and Machine Learning Approaches", 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019 Publication	
34	Fuzhe Ma, Tao Sun, Lingyun Liu, Hongyu Jing. "Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network", Future Generation Computer Systems, 2020 Publication	<1%
35	www.eurekaselect.com Internet Source	<1%
36	dspace.up.edu.ps Internet Source	<1%
37	"Computational Vision and Bio-Inspired Computing", Springer Science and Business Media LLC, 2020 Publication	<1%
38	Mubarik Ahmad, Vitri Tundjungsari, Dini Widianti, Peny Amalia, Ummi Azizah Rachmawati. "Diagnostic decision support system of chronic kidney disease using support vector machine", 2017 Second International Conference on Informatics and Computing (ICIC), 2017 Publication	<1%
39	ckdsite.com Internet Source	<1%
40	apo.ansto.gov.au Internet Source	<1%
41	mclguide.readthedocs.io Internet Source	<1%
42	Submitted to University of East London Student Paper	<1%



	Publication	
39	ckdsite.com Internet Source	<1%
40	apo.ansto.gov.au Internet Source	<1%
41	mclguide.readthedocs.io Internet Source	<1%
42	Submitted to University of East London Student Paper	<1%
43	Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee. "Predictive analytics for chronic kidney disease using machine learning techniques", 2016 Management and Innovation Technology International Conference (MITicon), 2016 Publication	<1%
44	vlib.itrc.ac.ir Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		