Performance Investigation of Different Machine Learning Algorithms in Predicting Chronic Kidney Disease

by

Md. Fahim Shikder (160021061)

Rezuanur Rahman Dip (160021065)

Ragib Ahsan (160021083)

A Dissertation Submitted to the Academic Faculty for Partial Completion of the Requirements for the Degree of

BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh

March 2021

CERTIFICATE OF APPROVAL

The thesis titled "Performance Investigation of Different Machine Learning Algorithms in Predicting Chronic Kidney Disease" submitted by Md. Fahim Shikder (160021061), Rezuanur Rahman Dip (160021065), and Ragib Ahsan (160021083) has been found as satisfactory and accepted as partial fulfillment of the requirement for the degree of Bachelor of Science in Electrical and Electronic Engineering on 10th March, 2021.

Approved by:

(Signature of the Supervisor)
Dr. Md. Ashraful Hoque
Professor
Department of Electrical and Electronic Engineering
Islamic University of Technology

(Signature of the Co-Supervisor)
Fahim Faisal
Assistant Professor
Department of Electrical and Electronic Engineering
Islamic University of Technology

(Signature of the Co-Supervisor)
Mirza Muntasir Nishat
Lecturer
Department of Electrical and Electronic Engineering
Islamic University of Technology

Table of Contents

Dedication

We dedicate this work to our families, friends and everyone who has been by our side, who has been a major source of relief and inspiration in ups and downs, in joys and sorrows of our lives. We wish them health, happiness and fulfillment on their own journeys.

Acknowledgements

First, we are wholeheartedly thankful to The Almighty Allah for giving us the patience and strength to conduct this thesis work seamlessly. We are deeply grateful to our honorable supervisor, ***Prof. Dr. Md. Ashraful Hoque***, Dean of Faculty of Engineering, IUT for giving us the opportunity to work under him. He has showed us the proper direction about how to conduct a research work and advised us to gather a strong basic for our work. He has always inspired us to do more and more innovative works. Without his support and advice, we would have lost track and get disoriented at the very beginning.

We would like to express our deepest gratitude to our co-supervisor,***Mr. Mirza Muntasir Nishat,*** Lecturer, Department of EEE, IUT, for sincerelygiving us his time and effort whenever it was necessary. He introduced us to the field of research when we were fully unaware of the area. His constant guidelines and insights have been extremely helpful to us. He always motivated us when we got stuck with different problems and managed to find us a solution every time. Without his clear direction, we would have an extremely hard time to complete this research.

We are also sincerely thankful to our co-supervisor ***Mr. Fahim Faisal,*** Assistant Professor, Department of EEE, IUT, for his proper directions, time and opinion that helped us to a great extent to conduct our thesis work. Throughout the whole time, he helped us with valuable insights from his experience on the field of research that made us have a clear idea about how to start and progress though a thesis work.

Abstract

This paper implies an investigative approach of studying the performance of different boosting algorithm in predicting chronic kidney diseases more accurately. In recent years chronic kidney disease (CKD) has reached a global prevalence as high as 11–13% with the majority in stage 3 which can lead to end stage renal disease (ESRD) if not detected early. Different boosting machine learning algorithms has been proven to be an effective tool to detect CKD while it's still in one of its initial stages. A dataset containing 400 instances and 25 attributes from the University of California, Irvine (UCI) repository has been exploited to train and test the model classifier. Four different data frames and correlation heatmap were constructed by four different strategies to begin the operation of the classifiers. Eleven machine learning algorithms were studied and their performance parameters like confusion matrix and accuracy were analyzed. Furthermore, a broad comparative investigation was conducted through the simulation of precision, sensitivity, F1 score, ROC-AUC of each algorithm.

CHAPTER 1

INTRODUCTION

1.1 Chronic Kidney Disease: A Brief Study

The heterogeneous disorders of kidney generally referred as Chronic kidney disease[43]. These disorders include conditions that damage our kidneys and decrease their ability to keep our body healthy by doing its job listed. Kidney is an essential organ of human body and chronic kidney diseases have become a prime issue in health sector because it is one of the main reasons of mortality. The main job of kidney is to remove the wastes produced by other organs from our body. When kidneys are failed to complete these functions for a long time that leads the ultimate chronic kidney malfunction. So, we can say that chronic kidney disease is the steady diminution of kidney functionalities over a period of time.

Like other diseases, this kind of kidney disease does not show that much of symptoms generally at the preliminary stage. It takes a long time to express its symptoms and aftereffects. But when it starts to show symptoms it is already late. The main reason is at the beginning it does not that much serious symptoms. It shows some symptoms like tiredness, dizziness, loss of taste etc. This kind of symptoms are not being taken care of that much and mostly overlooked. But in the long run it turns into other complicated kidney disease like glomerulus malfunction, kidney infection etc which can consequent to end stage renal failure [44]. Consequently, patients can be steered clear of procedures like kidney transplants and dialysis which cannot offer concrete safety. Apart

from this type of kidney diseases it also causes severe malfunctions to other body parts comprehending high risk of heart disease, high blood pressure, bone disease etc [1].

The only remedy of this type of chronic kidney disease is the early detection of the symptoms. But in most of the cases it is not possible to uphold the detection due lack of seriousness as well as there are not that much of quality diagnosis that can help to detect ckd at an early stage.

In spite of, the kidney disease should be predicted as early as possible, due to lack of concern and less amounts of symptoms most of the time chronic kidney disease is not diagnosed before mid-level or critical stage. About 96% people having kidney malfunction are not aware of having chronic kidney disease [2]. In many countries diagnostic centres collect huge amount of data for various purposes and those databases can be used to predict one's chronic kidney disease at any stages. As early and mid-stage kidney disease requires mostly blood and urine test [3], various dataset of blood test and urine test can be used to predict if someone has CKD or not. A lot of lives can be saved and severe complexity can be avoided if we can detect kidney diseases at an early state.

1.2 What is Machine Learning?

With the growth of Artificial intelligence machine learning has become one of the most effective and important tools in the engineering field. Machine learning is an applied form of artificial intelligence. It provides computerized system the ability to learn and develop from previous experience without being explicitly programmed [45].

Machine learning helps the computers to learn itself without human interference or help. It also gives the computers the ability to take actions according to the previous experience. To provide the previous experience, it requires a model that consists of huge datasets. These datasets contain the information of a specific event that helps the model to train itself. These datasets are the most valuable resource for a machine learning model. Based on the data the model takes all the steps that helps the user to get the perfect output.

Depending on the type of the model there are mainly three types of machine learning approaches.

i) Supervised Machine Learning.

ii) Unsupervised Machine Learning.

iii) Reinforcement Machine Learning.

Supervised learning uses the model that already contains dataset with previous experience while unsupervised learning uses the dataset that is a completely new. In other words, supervised model guides the input to the output while unsupervised system walks to the output without any guideline. Reinforcement learning is the approach that uses the model that interacts with a dynamic model where a particular task is fixed that has to be completed by the computer.

1.3 Machine Learning in Healthcare

Now-a-days machine learning is being used in most of the areas of our life. From daily email checking to launching a rocket to the outer space has the application of machine learning. Healthcare sector is not exclusive in the list of filed where machine learning is being applied. Eventually it’s one of the largest applied areas of machine learning.

With the development of technologies, machine learning is used widely in healthcare sector and is helping patients and clinicians in every other way. The most common healthcare use cases for machine learning are automating medical billing, clinical decision support and the development of clinical care guidelines. Clinical support and guidelines are helping us to detect diseases accurately and taking cautions on the basis of the accurate diagnosis. Machine learning can be the new and effective technology to test medical situations. It is all about exploration and extraction of huge datasets.

Chronic kidney disease remedy requires early-stage detection. Using machine learning can really help health professionals to diagnosis chronic kidney disease without the symptom barrier. Researchers all over the world has already started using data-science and machine learning algorithms to predict CKD at early and mid-stages to help medical professionals to provide better cure to the mass people before CKD get critical position and increase other fatal diseases.

Machine learning is one of the most noteworthy and effective technology in medical industry now a days to diagnose and predict different types of disease and their stages. As machine learning is all about exploration of huge dataset and their patterns, features, modes etc. The huge amount data set of diagnoses of different diseases can be fed into different machine learning algorithms. This implementation of algorithms in medical databases can help medical professional significantly to take constructive decisions on diseases, help them to obstruct errors and provide mass people a healthy life.

In this research, we investigated eleven Machine Learning algorithms which showed competent results in predicting Chronic kidney disease.

CHAPTER 2

Literature Review

Nowadays machine learning (ML) is one of the most noteworthy and effective technologies in the medical industry to diagnose and predict different types of diseases and their stages [8-9]. The huge amount of data set of diagnoses of different diseases can be fed into different machine learning algorithms to explore their patterns and features. This implementation of algorithms in medical databases can help medical professionals significantly to take constructive decisions on diseases, aid them to lessen human-made errors, and eventually ensure a healthy life for mass people.

2.1 Related Works

In recent times as machine learning techniques are being popular in medical sectors for diagnosing; chronic kidney disease is also in the queue to be predicted by dint of machine learning algorithms.

1. Engin AVCI et al. applied some classifier algorithms i.e.K-star, SVM, J48 etc. using dataset extracted from UCI and compared them by means of accuracy, sensitivity and parameters. According to their work the J48 classifier had 99% accuracy [4].

2. Gunarathne W.H.S.D. et al. from SLIIT also applied some other classification algorithm on the same database and evaluated their performance and accuracy. The algorithms they have used are Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Logistic Regression and Multiclass Neural Network. They have concluded that Multiclass Decision

Forest algorithm shows better performance than the other algorithms with the accuracy of 99.1% [5].

3. S.DilliArasu et al. have execute a research to manage the dataset. Since dataset can have missing values and that can degrade the accuracy of the result, so before applying the algorithms we have to pre-process the dataset by filling up the missing values. They proposed the WAELI algorithm to predict the missing value by applying single and multiple value imputation. The final value is produced by calculating the weighted average of each model [6].

4. L.JerlinRubibni et al. from Alagappa University have carried out a research on comparison of different types of analysis of early-stage prediction of CKD by using Multilayer perception, Radial basis functions and Logistic regression [7]. Their dataset was also extracted from the UCI machine learning repository. They concluded as the Multilayer perception has better accuracy than other neural networks.

5. Parul Sinha et al. has conducted an experiment to compare the performance of SVM (support vector machine) and KNN (K-nearest neighbor) classifiers on the dataset of UCI. According to them both classifiers provide promising output on the prediction of CKD. But they accomplished KNN over SVM by means of the precision of both classifiers [8].

6. Another work on diagnosis of CKD was conducted by HuseyinPolat et al. They have improvised on the feature selection of the dataset before applying the algorithm. They applied the filter, wrapper and embedded feature selection methods on the dataset and then passed them through the SVM algorithm. According to their work the filter schema subset evaluation was achieved the best outcome which was 98.5% [9].

7. A different approach of featuring of datasets was conducted by Nusrat Tazin et al. from Northern University Bangladesh. They approached Root Mean Squared Error, Mean Absolute Error and Receiver Operating Characteristic curve to pre-process the data. After featuring the dataset, they applied Naïve Bays, Decision Tree, SVM and KNN algorithm. According to their comparison in every featured dataset the Decision Tree provides the best output which is about 98%-99% [10].

8. To improve the balance in the dataset Pinar Yildirim has carried out a research to predict CKD on imbalance data by multilayer perception. In this research the main focus was to manage the imbalance dataset by sampling it using various sampling methods like Under sampling, Oversampling, Resampling, Spread sub sampling and SMOTE. For sampling calculation, he proposed the multilayer perception method which is a neural network approach. According to his work, resample method has performed better than other sampling algorithms [11].

9. Devika R et al. have focused on Naïve bays, KNN and Random forest algorithms in their research to predict CKD. Among these classifiers, Random forest classifier has performed better with 99% accuracy [12].

10. AbdulhamitSubas et al. has also claimed Random forest algorithm is better than other machine learning algorithms to predict CKD [13]. According to their findings they have claimed Random forest algorithm has 100% accuracy to predict CKD than other algorithms like KNN, Naïve Bays, SVM etc.

11. I.A. Pasadana et al. has also carried out a research to predict CKD using different types of decision tree algorithm. They have applied DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomForest, RandomTree, REPtree and Simple Cart

algorithms to predict CKD. They have also denouement the Random Forest algorithm having the highest accuracy among all the decision tree algorithm which is about 100% [14].

12. Boosting algorithm is another method to increase result accuracy other than data pre-processing. MerveDoğruyolBaşar et al. [15] has accomplished applying AdaBoost algorithm to increase the accuracy of their result in their researche.

13. AmanahFebrianIndriani et al.[16] has accomplished applying AdaBoost algorithm to increase the accuracy of their result in their respective researches. AmanahFebrianIndriani et al. has implied PSO algorithms to optimize their result more precisely. After applying AdaBoost and PSO feature selection algorithm combinedly they were able to increase their average accuracy by 36.20% [16].

14. Adeola Ogunleye et al has applied XGBoost algorithm to enhance their result accuracy. Their proposed model has achieved 100% accuracy after applying XGBoost algorithm [17].

2.2 Research Objective and Outline

The main objective of our research work is to ease the process to an extent. Considering the present constraints in this field and the importance of detection of Chronic kidney disease, we had some objectives before we kicked off our undergraduate research work. The objectives are:

1. Studying about Chronic Kidney Disease (CKD): The goal was to understand the disease, how it functions in our body, what are the symptoms, how it spread it in our body, what is the stages that are crucial for this disease and the importance of early detection of this disease.

2. Applying Machine Learning Algorithms: Our main object is to apply several machine learning algorithms in to a dataset that contains different attributes about CKD. We have applied 11 machine learning algorithms and compared them on the basis of the outcome that provides by the supervised machine learning algorithms.

3. Building computer aided diagnosis system for efficient prediction of CKD: In addition to keeping performance parameters satisfactory, we worked hard on building a computer aided diagnosis system that can predict CKD precisely from any input data. The method is supervised by a large dataset that already contains previous attributes on CKD and their outcome.

Chapter 1 provides an introduction to our dissertation. Chapter 2 is comprised of a Literature review that includes the current knowledge, including substantive findings, and input on a particular subject theoretical and methodological. Another part of chapter 2 contains details about the dataset we have dragged from UCI respiratory. Chapter 3 talks about the details on all the 11 machine learning algorithms we have applied to fulfil our purpose. The methodology of the work such as the Data Description, Data Pre-processing, Feature Scaling, Data Frames and correlation heatmaps, Hyper Parameter Tuning are all described in chapter 4. In chapter 6 we have discussed the results and analysis of our model. Here we have discussed the confusion matrices, Comparison of accuracies between different algorithms, Precision, sensitivity, F1 score, AUC-ROC curves for both tuned and without tuned states. Chapter 6 contains the

conclusion of the current work as well as the information about a future query to improve the performance of the methods proposed.

2.3 Data Set and Attributes

Dataset is the requisite material for applying machine learning algorithms. In this study we have applied our algorithms on Chronic Kidney Disease dataset of UCI machine learning repository [18]. UCI dataset repository is one of the vast used and reliable datasets for applying machine learning algorithms. The dataset we have used contains 400 instances and 25 attributes. The attributes consist of age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema and anemia.

Chapter-3

Study of Machine Learning Algorithms

Machine learning is a subset of Artificial Intelligence that studies algorithms that works and adapts to different scenarios through experience. The experience is gained by training with a collection of data which is called training data. After going through the training, machine learning algorithms predict or classify data without explicitly programmed to do so.

In this paper, 10 supervised classification learning algorithms are chosen to identify Chronic Kidney Diseases and their results are briefly compared under different criterions.

3.1. Logistic Regression

Logistic regression aims to model the probabilities for classification problems with two possible outcomes. It's an extensive feature of the linear regression model for classification problems. Logistic regression transforms its output values using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

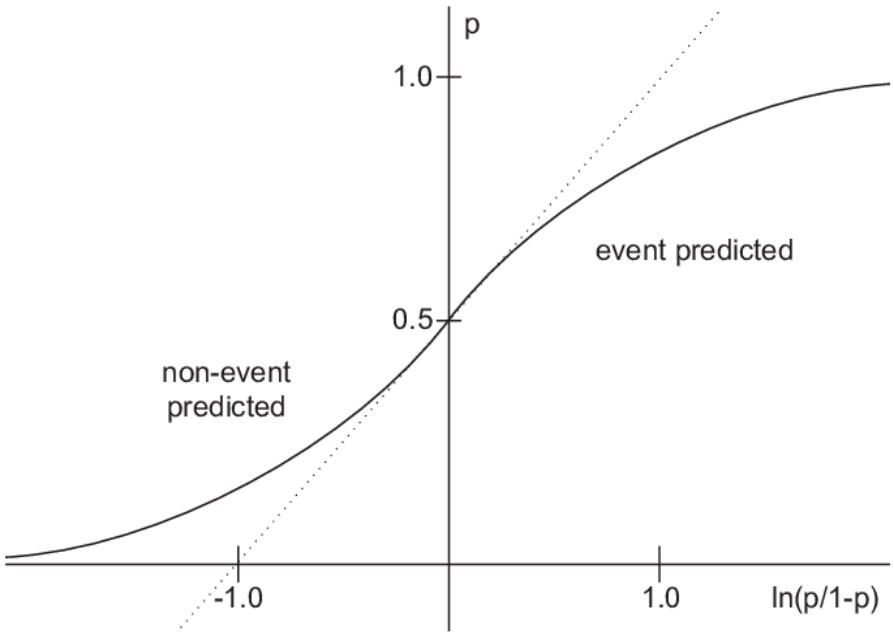
Logistic Regression is a statistical classification model which estimates the probability of an event existing within a certain class. Despite having “regression” in its name, logistic regression is a widely used binary classifier. A threshold is set to predict in which class does a data belong, which is called a decision boundary. This classification probability is calculated by the logistic function which is actually a sigmoid function [22-23]. A linear model is included in the logistic function like below:

$$\hat{p}_\theta = h_\theta(x) = \sigma(x^T\theta)$$

Here,

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

A basic illustration of logistic regression is shown below:



3.2. K-Nearest Neighbors

K-Nearest Neighbors is one of the simplest and most used supervised machine learning algorithms. K-NN algorithm presumes the similarity between the new data and available cases and place the new case into the category that is most similar to the available categories. Technically it doesn't train any dataset, instead an observation is predicted to fall under that class which have the largest proportion of k-nearest neighbors around it.

The value of K is chosen in such a way that minimum number of errors are encountered while making accurate predictions. Distance is considered to be a metric to determine similarity i.e., the closet datapoint around the point under observation can be considered most similar to the data point [24-25]. There are a large variety of distance metrics like:

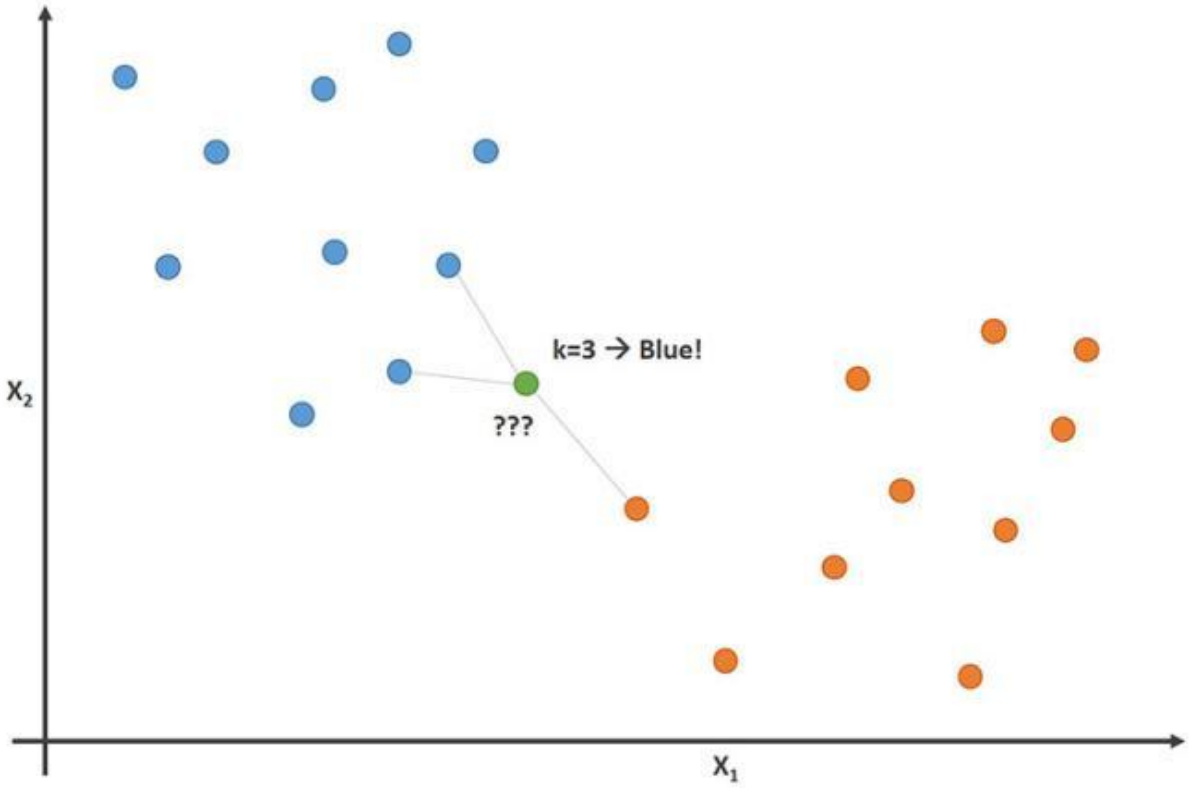
Euclidean distance,

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (x_i^2 - y_i^2)}$$

Manhattan distance,

$$d_{man\,hattan} = \sum_{i=1}^{n} |x_i - y_i|$$

Following figure shows how the K-NN algorithm works:



3.3. Support Vector Machine

Support Vector Machine is one of the most robust algorithms based on the statistical learning framework which offers solution for both regression and classification problems. The task of the

support vector machine algorithm is to find a hyperplane in an N-dimensional space where N being the number of features that distinctly classifies the data points. The objective is to find a hyperplane which has the maximum margin i.e., maximum distance between the data points of each class. Using the kernel trick, SVM can classify both linear and non-linear datasets. The datasets are separated by a *(n-1)* hyperplane, where every data point is considered to be a *n*-dimensional vector. For a two-dimensional space, hyperplane is a line separating a plane in two parts [24][26]. A support vector classifier can be defined by the following terms:

$$f(x) = \beta_o + \sum_{i \epsilon S} \alpha_i \, K(x_i, x_{i'})$$

Here,

β_o = Bias

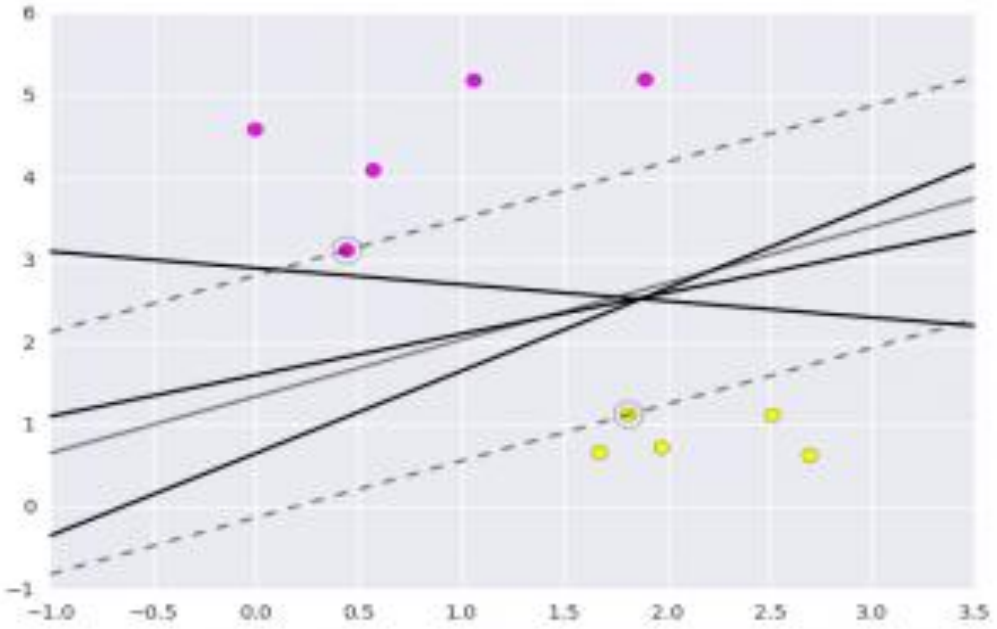
S= Set of observations

α= Model parameters that has to be learned

For linear kernel,

$$K(x_i, x_{i'}) = \sum_{j=i}^{p} x_{ij} \, x_{i'j}$$

Following figure shows a support vector machine simulation:



3.4. Decision Tree

Decision Tree is another supervised learning algorithm whose goal is to that can train a model to classify a target variable by learning simple chained decision rules from previous input variables. The variables are split recursively based on a set of impurity criteria until some stopping criteria is reached. The decision tree model looks much like an upside-down tree where the first decision rule resides at the top and subsequent decision rules spreads below like branches of the tree. For predicting a class label for an instance of record we begin from the root of the tree [24]. We compare the values of the root feature with the record's feature. On the basis of comparison, we

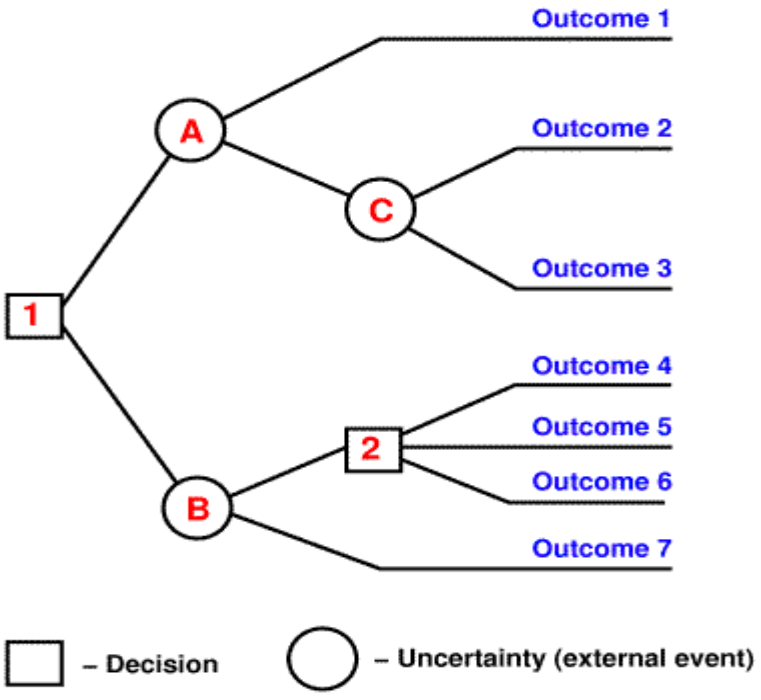
follow the branch corresponding to that value and jump to the next node. The next node is selected depending on which node can provide the maximum gain values. Among many impurity measurement systems, Gini impurity is selected for the used model [27-29].

$$G(t) = 1 - \sum_{i=1}^{c} p_i^2$$

Here,

G(t)= Gini impurity at node t

p_i = Proportion of observation at class c of node t



3.5. Random Forest

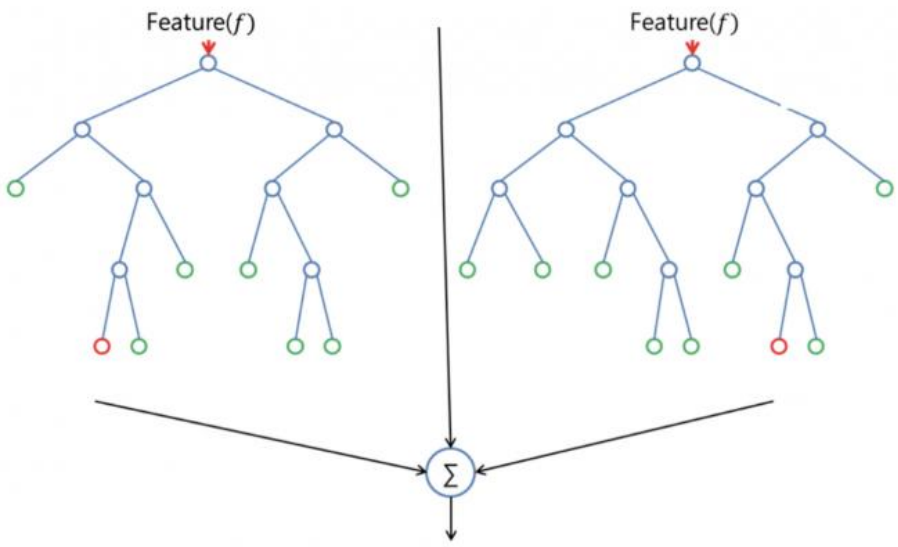
Random Forest is a learning algorithm for both regression and classification which operates by creating multiple decision trees at training time and providing output class of individual trees. Random forest operates maintaining the idea that a large number of relatively uncorrelated models operating as a committee will outperform any of the individual constituent models. It generates decent prediction results even without hyper parameter tuning. This model does a small tweak that utilizes the de-correlated tree by building a multitude of decision trees on bootstrapped samples from training data, this process is known as bagging [30]. During bootstrapping, it filters a few numbers of feature columns out of all feature columns. Bootstrap modeling decreases the variance and increases the bias [29]. It has an effective method for estimating missing data. Predictions of unknows inputs after training can be written as:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Where, B= Optimal number of trees

Also, uncertainty of the prediction can be written as:

$$\sigma = \sqrt{\frac{\sum_{b-1}^{B}(f_b(x') - \hat{f})^2}{B-1}}$$



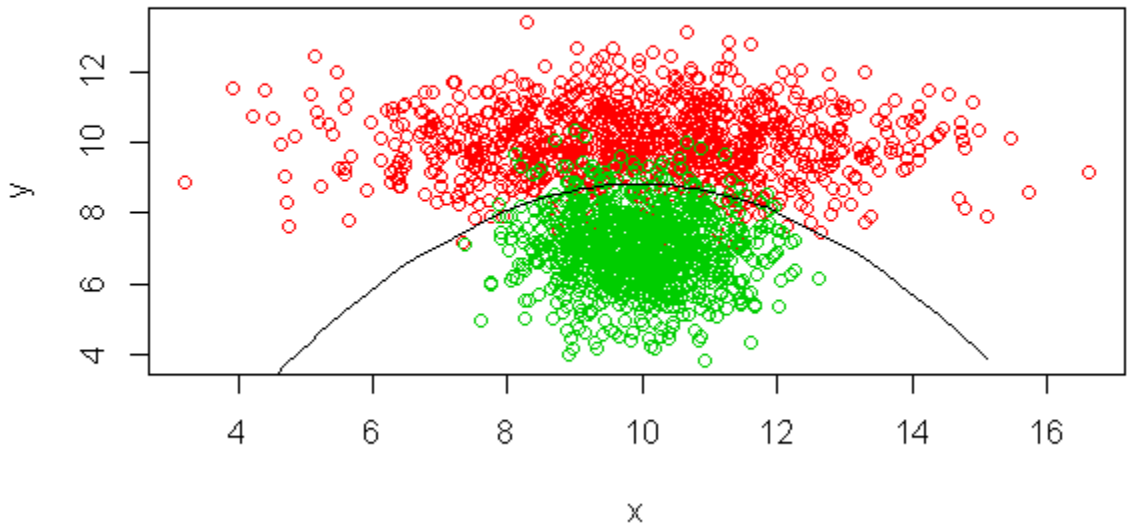
3.6. Naive Bayes

Naive Bayes is a supervised algorithm which imposes independence of features while classifying data. This classifier assumes that the presence of a particular feature in a class is not related to the presence of any other attribute. This algorithm works extremely fast relative to the other classification algorithms. This model is an effective tool for datasets which have a high number of input features. It considers all the features available including some of the features that have weak effects on the final prediction. This algorithm is easily scalable and it is a widely used algorithm for real-world applications. The probabilistic model of Naïve Bayes algorithm can be written as:

$$P(A\backslash B) = \frac{P(B\backslash A)*P(A)}{P(B)} = \frac{P(A\cap B)}{P(B)}$$

Here, A and B are two independent events [29][31].

A simulation sample of Naive Bayes algortihm:



3.7. AdaBoost

AdaBoost is short of *Adaptive Boosting* which is an adaptive machine learning algorithm which aims to convert a set of weak data features into strong ones. It is used to boost the performance of decision trees and this is based on binary classification. This is an iterative algorithm where in each iteration the dataset is divided into two regions and the features used in one iteration will be

given less weightage and the misclassified data will be given higher weightage in the next iteration. This is performed by building a model from the training data, then creating another model that attempts to correct the errors from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added. After all the iterations are finished, they are combined with the corresponding weights to form up a strong classifier which predicts the classes of the unseen data [32-33]. The classifier output can be written as:

$$F_T(x) = \sum_{t=1}^{T} f_t\,(x)$$

Here, f_t = A weak feature

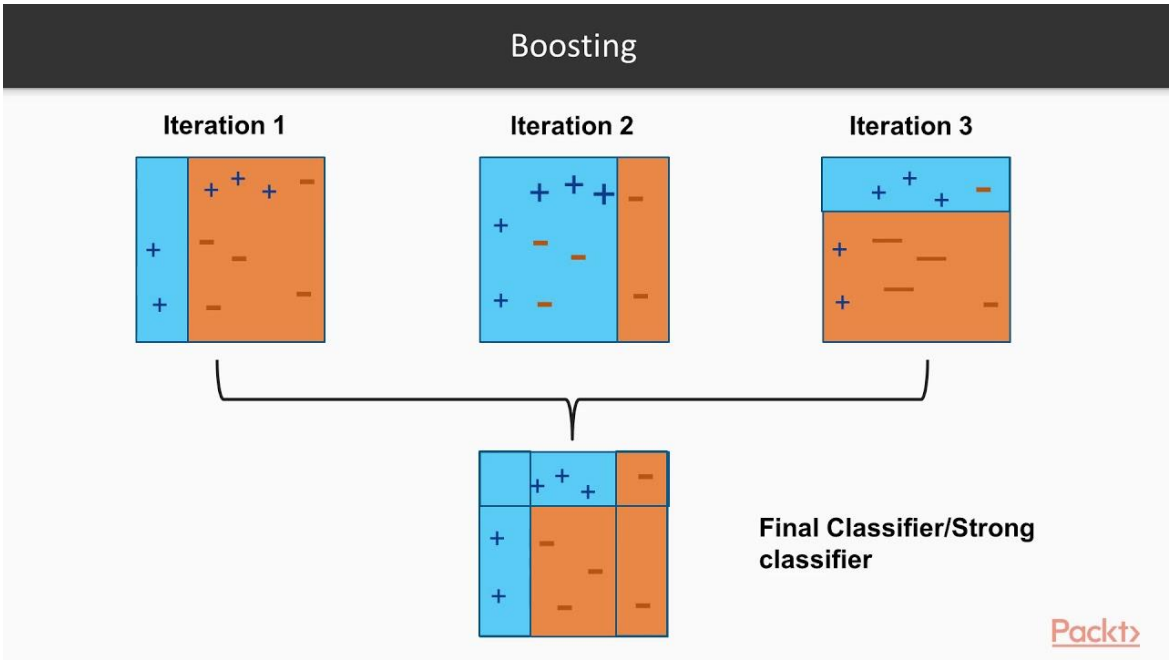
After adaptive boosting, the error function is calculated the following way:

$$E_t = \sum_{i} E[F_{t-1}(x_i) + \alpha_t h(x_i)]$$

Here,

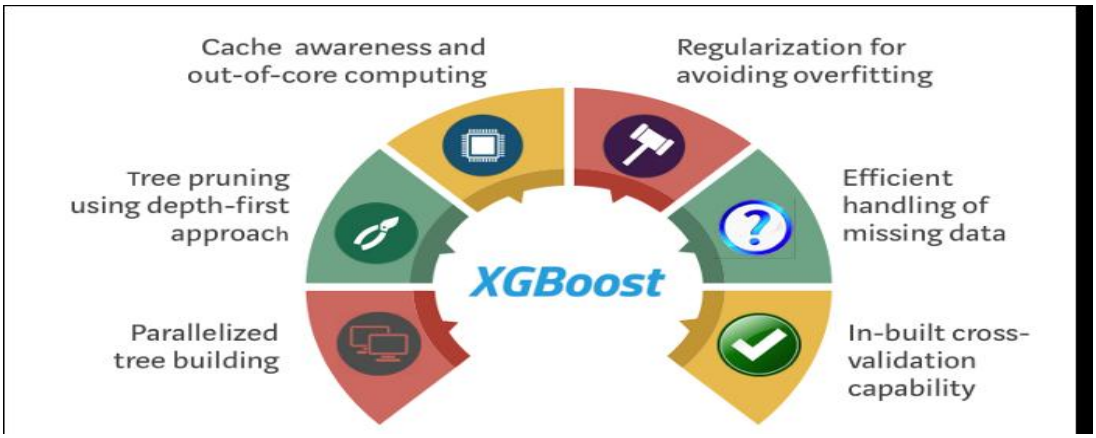
$F_{t-1}(x)$ = Boosted classifier

α_t = Co-efficient assigned to the weak classifier



3.8. XGBoost

XGBoost is an ensemble machine learning algorithm which that uses a framework called gradient boosting.XGBoost focuses on computational speed and model performance and it offers a number of advanced features. When it comes to small to medium structured or tabular datasets, this algorithm is one of the most efficient ones to perform regression, classification, ranking, predictions. It performs a diverse range of tasks assigned to it by boosting the weak features in the datasets using gradient descent architecture [34]. XGBoost provides a parallel tree boosting named GBDT, GBM that solve many data science problems in a fast and accurate way.



3.9. Multilayer Perceptron

Multilayer perceptron (MLP) is a feedforward artificial neural network which consists of multiple perceptron organized into layers. It contains nodes of at least three layers named input node, hidden layer and output node. This network uses a non-linear activation function which maps weighted inputs to each neuron outputs. In this paper we used sigmoid functions as the activation functions.

$$y(v_i) = \tanh(v_i)$$

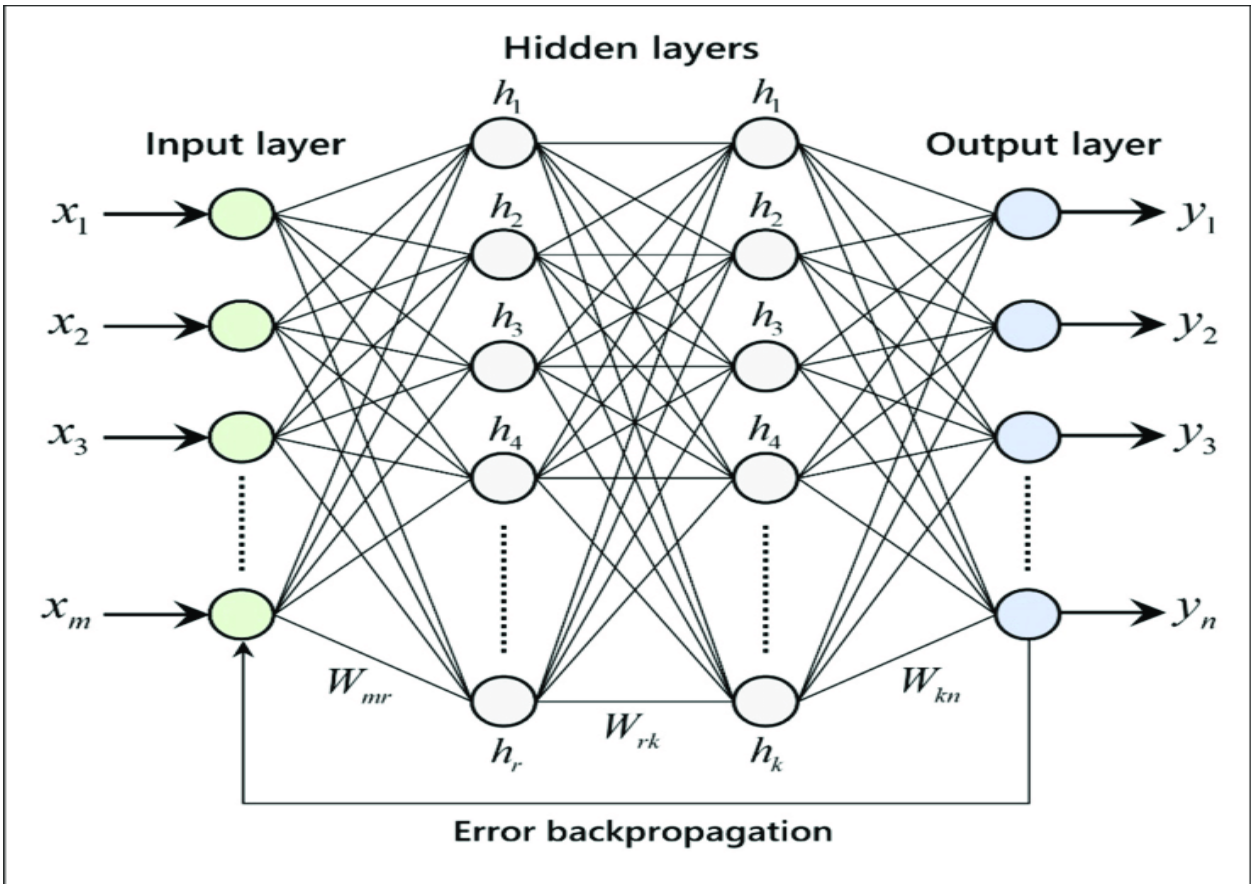
$$y(v_i) = (1 + e^{-v_i})^{-1}$$

The range of the first hyperbolic tangent is -1 to 1 and the second hyperbolic tangent is a logistic function [35]. Training involves adjusting the parameters or the weights and biases of the model in order to minimize error. Backpropagation is used to make those weigh and bias adjustments

relative to the error. The error itself can be measured in a variety of ways, including by root mean squared error (RMSE).

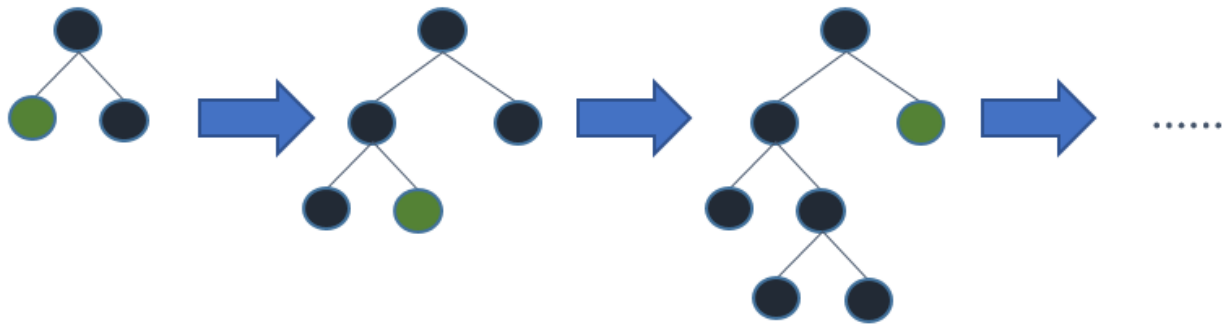
Learning in the perceptron is carried out by back propagation and minimized error function at the output node *j* after performing gradient descent can be written as:

$$\varepsilon(n) = \frac{1}{2}\sum_{j} e_j^2(n)$$



3.10. LightGBM

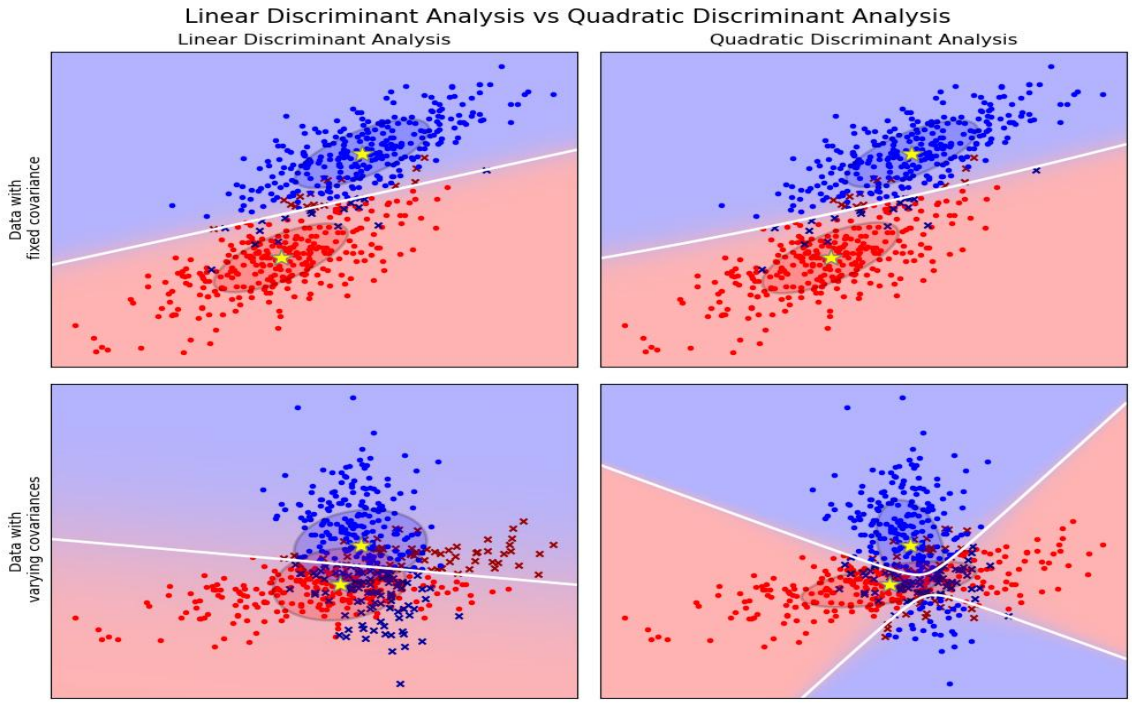
LightGBM is another algorithm that uses gradient boosting as a framework and decision tree as an algorithm to perform classification, ranking and other machine learning tasks. Instead of level-wise splitting, LightGBM uses leaf wise split approach in a binary tree that speeds up the training of data and reduces memory usage. LightGBM also uses histogram-based algorithms that store continuous features into discrete bins. This speeds up training and reduces memory usage. This algorithm has better compatibility with large datasets than other boosting algorithms [36-37].



Leaf-wise tree growth

3.11. Quadratic Discriminant Analysis

Quadratic Discriminant Analysis is a statistical classifier which disjoints two or more classes of data by using quadratic decision surface. This classifier is used on those cases where there exists a difference between the covariance matrices [38].


Linear Discriminant Analysis vs Quadratic Discriminant Analysis
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances

Chapter-4

Methodology

4.1 Data Description

In this study, we have applied our algorithms on the Chronic Kidney Disease dataset of UCI machine learning repository containing 400 instances and 25 attributes [16]. UCI dataset repository is one of the vastly used and reliable datasets for applying machine learning algorithms. The information of all the 25 attributes is tabulated in Table I.

TABLE I. ATTRIBUTE INFORMATION [18]

Sl. No.	**Attribute**	**Information**
1.	Age *(age)*	Discrete Integer Values
2.	Blood pressure *(bp)*	Discrete Integer Values
3.	Specific Gravity *(sg)*	Nominal Values

4.	Albumin *(al)*	Nominal Values
5.	Sugar *(su*)	Nominal Values
6.	Red blood cells *(rbc)*	Nominal Values
7.	Pus cell *(pc)*	Nominal Values
8.	Pus cells clumps *(pcc)*	Nominal Values
9.	Bacteria *(ba)*	Nominal Values
10.	Blood Glucose Random *(bgr)*	Numerical Values (mgs/dl)
11.	Blood Urea *(bu)*	Numerical Values (mgs/dl)
12.	Serum creatinine *(sc)*	Numerical Values
13.	Sodium *(sod)*	Numerical Values (mEq/L)
14.	Potassium *(pot)*	Numerical Values (mEq/L)
15.	Hemoglobin	Numerical Values

	(hemo)	(gms)
16.	Packed Cell Volume *(pcv)*	Numerical Values
17.	White blood cell count *(wc)*	Discrete Integer Values
18.	Red blood cell count *(rc)*	Numeric Values
19.	Hypertension *(htn)*	Nominal Values
20.	Diabetes Mellitus *(dm)*	Nominal Values
21.	Coronary Artery Disease *(cad)*	Nominal Values
22.	Appetite *(appet)*	Nominal Values
23.	Pedal Edema *(pe)*	Nominal Values
24.	Anemia *(ane)*	Nominal Values
25.	Classification	Nominal Values

	(class)	

4.2 Data Preprocessing

Data preprocessing is an instrumental part of data mining. In the data preprocessing step of machine learning, the data of the large datasets get transformed and encoded so that the machine can easily compile it i.e., convert the data, images into zeros (0) and ones (1). There are mainly two types of data in a dataset. Which are:

- Numerical data: Numerical data are the type of data which can be expressed by numbers. For example: age, birth year, any kind of quantity etc.
- Categorical data: Data whose values are extracted from a defined set of values. For example: Monday, Thursday, Boolean expressions (true, false)

Both kind of data are pushed through some steps of data preprocessing. The steps are:

A. Data Quality Assessment: Machine learning model run poorly on poor quality data, they do not provide accurate predictions, classification, regression results on poor data. So, the quality of the data will possess significant impact on the results that different machine learning models will provide. In this step, the following type of data will be filtered and furnished:
    a. Missing values
    b. Duplicate values
    c. Inconsistent values

B. Feature Assemble: In this step, the random, unorganized data will be put into perspective to build some patterns and categorize the data. This will minimize the memory storage requirement and fasten up the machine learning models.

C. Feature Sampling: In machines learning, we often need to deal with a very large dataset consisting hundreds or thousands of instances. The information a dataset carries is directly proportional to its size. But working with such huge datasets are money and time consuming. Instead, we can get satisfactory results by taking a portion of the dataset into the analysis which can save both time and money. This is called feature sampling. This enables the machine learning model to learn more quickly and accurately.

D. Reduction of dimensionality:Dimensionality reduction refers to the reduction of number of input variables in a dataset. Dimensionality reduction is used in areas that deal with large numbers of observation, such as digital signal processing, speech recognition and bioinformatics.

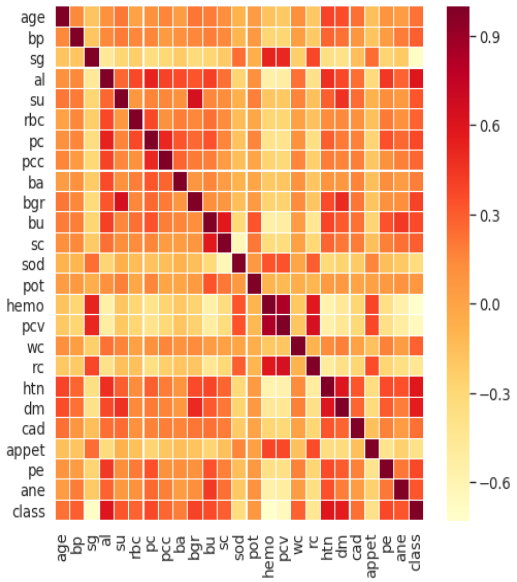
E. Feature Encoding: Sometimes there are some categorical entries in the dataset like true/false, yes/no etc. But machine learning models can only work with numerical entries. For this reason, it is necessary to transform the categorical values into numerical values. This process is called feature encoding.

To implement machine learning algorithms proficiently, data has to be accurate and well-organized. Initially, this data set was comprised of several missing values that did not facilitate to develop a satisfactory outcome. Hence, four different data frames were constructed to implement the algorithms and observe the results. Firstly, the missing values were filled with mean values of corresponding attributes. After that, the missing values were

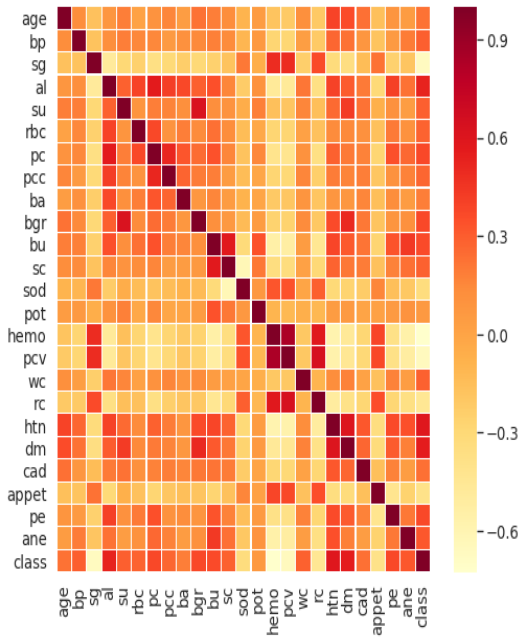
also filled with median and mode values and obtained 400 entries in total. Lastly, we dropped all the null values in which we were left with 158 entries.

4.3 Data Frames & Correlation Heatmaps:

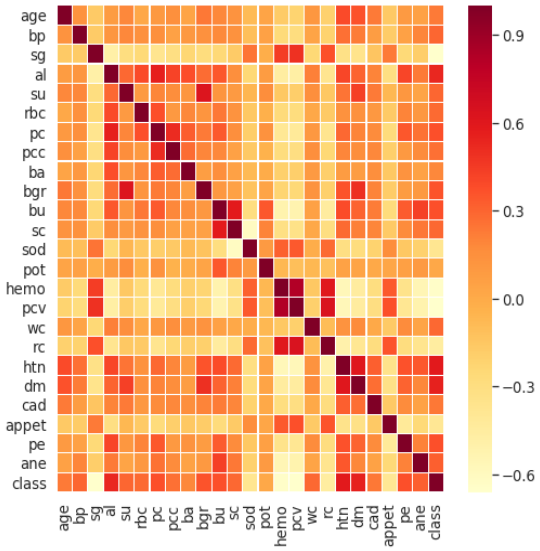
Mean:



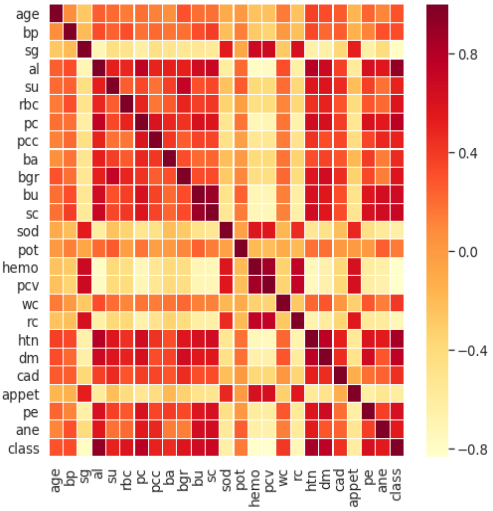
Median:


age
bp
sg
al
su
rbc
pc
pcc
ba
bgr
bu
sc
sod
pot
hemo
pcv
wc
rc
htn
dm
cad
appet
pe
ane
class
0.9
0.6
0.3
0.0
−0.3
−0.6

Mode:


age
bp
sg
al
su
rbc
pc
pcc
ba
bgr
bu
sc
sod
pot
hemo
pcv
wc
rc
htn
dm
cad
appet
pe
ane
class
0.9
0.6
0.3
0.0
−0.3
−0.6

No Null:


age
bp
sg
al
su
rbc
pc
pcc
ba
bgr
bu
sc
sod
pot
hemo
pcv
wc
rc
htn
dm
cad
appet
pe
ane
class
0.8
0.4
0.0
−0.4
−0.8

4.4 Hyper Parameter Tuning:

Eleven machine learning algorithms were implemented to predict patients with chronic kidney diseases or without chronic kidney diseases. First of all, algorithms were implemented with default hyper parameter setting. Accuracy along with other performance metrics were calculated. For acquiring better prediction hyper parameter tuning of the machine learning algorithms is essential. There are two popular ways of hyper parameter tuning:

- RandomizedsearchCV
- GridsearchCV

RandomizedsearchCV: RandomizedsearchCV is a strategy of tuning hyperparameters where random combinations of the hyperparameters are implemented to find the best fit for the model. The selection of parameters is completely random. RandomizedsearchCV is very effective in case of many parameters to try and the training time is a concern. [39]

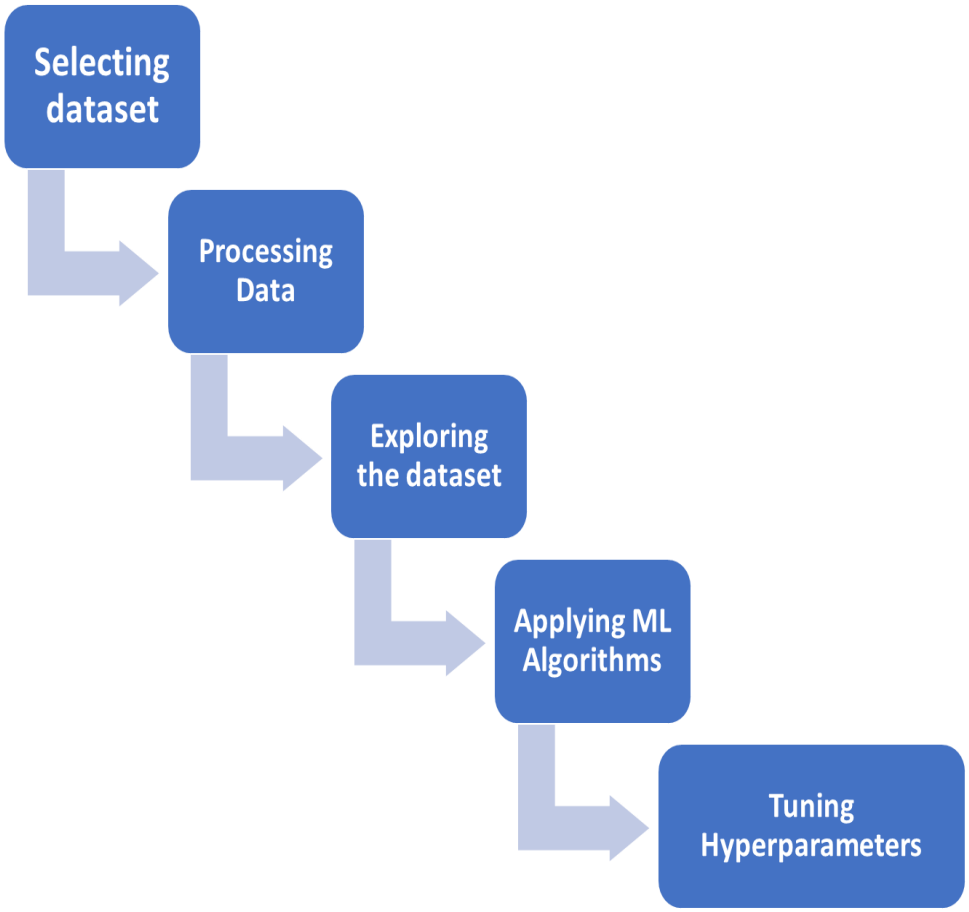
GridsearchCV: GridsearchCV is an effective technique for finding the parameters in supervised learning algorithms and making the model more generalized. In GridsearchCV, all possible combinations of the parameters of interest are tried. Then, the best set of parameters are chosen. [39]

GridsearchCV performs well in case of a small number of hyperparameters. On the other hand, if the number of parameters is high and calculation time is a concern, it is better choice to use the RandomzedsearchCV. [39]

In this research hyperparameters were tuned with RandomizedsearchCV technique. Then performance metrics were calculated accordingly to compare with the default hyperparameters.

4.5 Methodology Flow Chart:

The entire methodology can be illustrated by the following flow chart:



Chapter-5

Results Analysis

5.1 Confusion Matrices

Confusion Matrix is a performance metric for machine learning classification models where output can be two or more classes. It consists of a table with four combinations of predicted and actual values. It is useful for calculating other performance metrics like Recall, Precision, Specificity, Accuracy and AUC-ROC Curve. [40]

	Predicted	
Actual	TF	FP
	FN	TP

True Positive: Predicted positive incident is actually positive.Model predicted that a patient has CKD which he/she actually has.

True Negative: Predicted negative incident is actually negative. Model predicted that a patient has no CKD which he/she does not have.

False Positive: (Type 1 Error): Predicted positive incident is actually negative. Model predicted that a patient has CKD which he/she does not have.

False Negative: (Type 2 Error): Predicted negative incident is actually positive. Model predicted that a patient does not have CKD which he/she actually has.

Confusion Matrices of eleven algorithms in detecting CKD from our research is given below:

TABLE

CONFUSION MATRICES FOR DIFFERENT ALGORITHMS

Confusion Matrix KNN			Predicted	
			False	True
Mean	Actual	False	143	7
		True	27	223
Median	Actual	False	143	7
		True	35	215
Mode	Actual	False	144	6
		True	31	219
No Null	Actual	False	115	0
		True	9	34

Confusion Matrix Logistic Regression			Predicted	
			False	True
Mean	Actual	False	148	2
		True	4	246
Median	Actual	False	147	3
		True	2	248

Mode	Actual	False	150	0
		True	6	244
No Null	Actual	False	115	0
		True	1	42

Confusion Matrix Decision Tree			Predicted	
			False	True
Mean	Actual	False	145	5
		True	11	239
Median	Actual	False	144	6
		True	7	243
Mode	Actual	False	145	5
		True	5	245
No Null	Actual	False	115	0
		True	3	40

Confusion Matrix Random Forest			Predicted	
			False	True
Mean	Actual	False	150	0
		True	1	249
Median	Actual	False	150	0
		True	1	249
Mode	Actual	False	150	0
		True	8	242

No Null	Actual	False	115	0
		True	2	41

Confusion Matrix SVC			Predicted	
			False	True
Mean	Actual	False	144	6
		True	7	243
Median	Actual	False	143	7
		True	5	245
Mode	Actual	False	144	6
		True	11	239
No Null	Actual	False	115	0
		True	1	42

Confusion Matrix Naïve Bayes			Predicted	
			False	True
Mean	Actual	False	150	0
		True	17	233
Median	Actual	False	150	0
		True	14	236
Mode	Actual	False	150	0
		True	14	236
No Null	Actual	False	115	0
		True	15	28

Confusion Matrix AdaBoost			Predicted	
			False	True
Mean	Actual	False	150	0
		True	2	248
Median	Actual	False	150	0
		True	1	249
Mode	Actual	False	150	0
		True	1	249
No Null	Actual	False	115	0
		True	1	42

Confusion Matrix XGBoost			Predicted	
			False	True
Mean	Actual	False	149	1
		True	1	249
Median	Actual	False	149	1
		True	1	249
Mode	Actual	False	150	0
		True	2	248
No Null	Actual	False	115	0

		True	1	42

Confusion Matrix MLP			Predicted	
			False	True
Mean	Actual	False	135	15
		True	21	229
Median	Actual	False	121	29
		True	20	230
Mode	Actual	False	141	9
		True	44	206
No Null	Actual	False	115	0
		True	9	34

Confusion Matrix LightGBM			Predicted	
			False	True
Mean	Actual	False	150	0
		True	1	249
Median	Actual	False	150	0
		True	1	249
Mode	Actual	False	150	0
		True	2	248
No Null	Actual	False	115	0
		True	1	42

Confusion Matrix QDA			Predicted	
			False	True
Mean	Actual	False	150	0
		True	26	224
Median	Actual	False	150	0
		True	23	227
Mode	Actual	False	150	0
		True	24	226
No Null	Actual	False	115	0
		True	43	0

5.2 Comparison of Accuracies among Different Algorithms

Accuracy is the most common and intuitive performance metric which is the ratio of correctly predicted incidents to the total incidents. The equation of accuracy for classification problem is given below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad [41]$$

Comparison of accuracies among different algorithms for both default hyperparameters and tuned hyper parameters is given below:

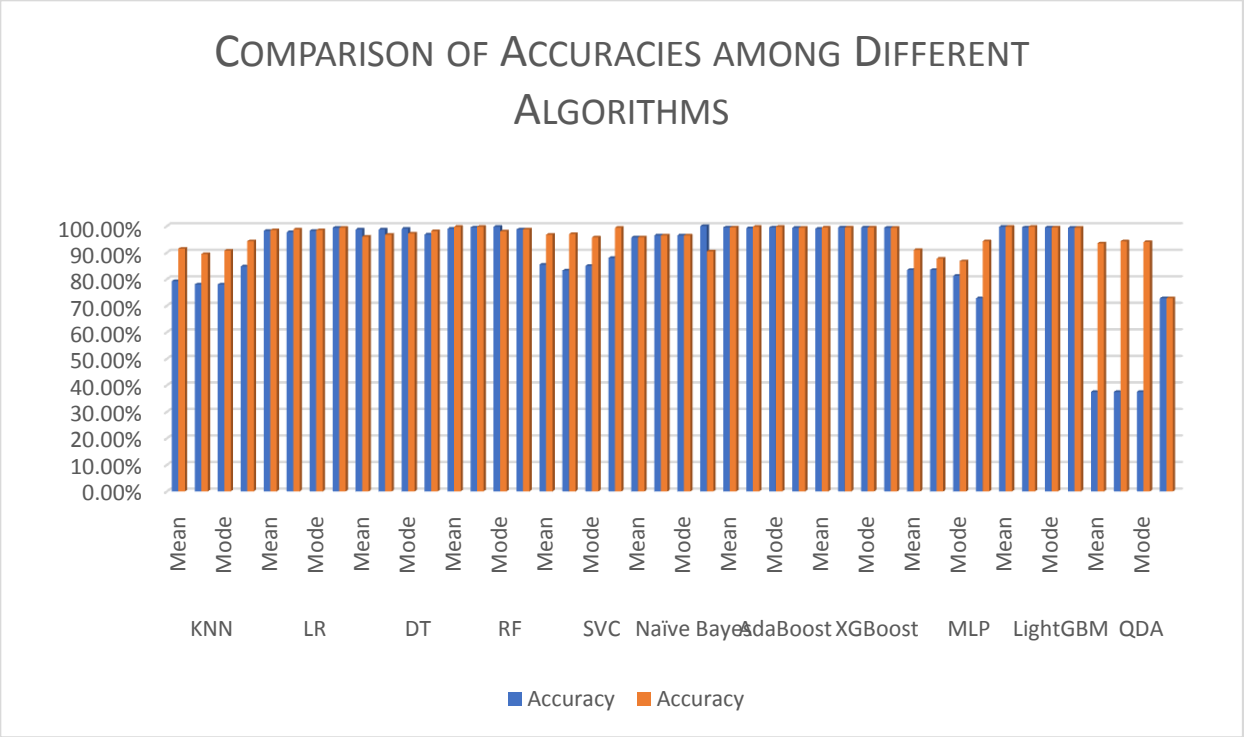
TABLE

COMPARISON OF ACCURACIES AMONG DIFFERENT ALGORITHMS

Algorithm	Null Filling Method	Accuracy (Without Tuning)	Accuracy (With Tuning)
KNN	Mean	79.25%	91.5%
	Median	78%	89.5%
	Mode	78%	90.75%
	Dropping Null	84.81%	94.3%
Logistic Regression	Mean	98.25%	98.5%
	Median	97.75%	98.75%

	Mode	98.25%	98.5%
	Dropping Null	99.36%	99.36%
Decision Tree	Mean	98.75%	96%
	Median	98.75%	96.75%
	Mode	99%	97.25%
	Dropping Null	96.83%	98.10%
Random Forest	Mean	99%	99.75%
	Median	99.5%	99.75%
	Mode	99.75%	98%
	Dropping Null	98.73%	98.73%
SVC	Mean	85.5%	96.75%
	Median	83.25%	97%
	Mode	85%	95.75%
	Dropping Null	87.97%	99.36%
Naïve Bayes	Mean	95.75%	95.75%
	Median	96.5%	96.5%
	Mode	96.5%	96.5%
	Dropping Null	100%	90.50%
AdaBoost	Mean	99.5%	99.5%
	Median	99.25%	99.75%
	Mode	99.5%	99.75%
	Dropping Null	99.36%	99.36%
XGBoost	Mean	99%	99.5%
	Median	99.5%	99.5%
	Mode	99.5%	99.5%

	Dropping Null	99.36%	99.36%
MLP	Mean	83.5%	91%
	Median	83.5%	87.75%
	Mode	81.25%	86.75%
	Dropping Null	72.78%	94.3%
LightGBM	Mean	99.75%	99.75%
	Median	99.5%	99.75%
	Mode	99.5%	99.5%
	Dropping Null	99.36%	99.36%
QDA	Mean	37.5%	93.5%
	Median	37.5%	94.25%
	Mode	37.5%	94%
	Dropped Null	72.78%	72.78%



5.3 Comparison of Precisions among Different Algorithms

Higher accuracy doesn't always mean that the model is best. Accuracy can be a great metric only if the dataset is symmetric which means false positive incidents and false negative incidents carry same value. But this is not true for diseases detecting models. So, other metrices like precision were evaluated observe the performance of the model. Precision is the ratio of correctly predicted positive incidents to the total predicted positive incidents. High precision means low false positive rate. [41]

$$\text{Precision} = \frac{TP}{TP+FP} \quad [41]$$

Comparison of precisions among different algorithms for both default hyperparameters and tuned hyper parameters is given below:

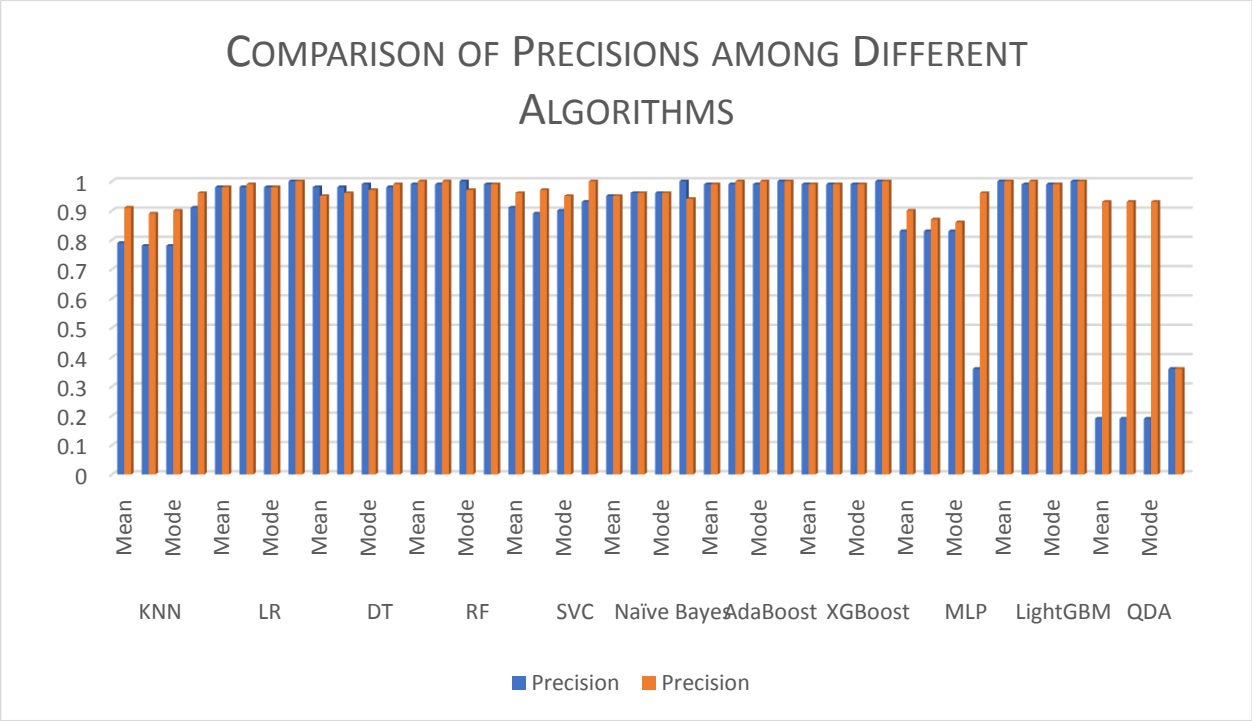
TABLE

COMPARISON OF PRECISIONS AMONG DIFFERENT ALGORITHMS

Algorithm	Null Filling Method	Precision (Without Tuning)	Precision (With Tuning)
KNN	Mean	0.79	0.91
	Median	0.78	0.89
	Mode	0.78	0.90
	Dropping Null	0.91	0.96
Logistic Regression	Mean	0.98	0.98
	Median	0.98	0.99
	Mode	0.98	0.98
	Dropping Null	1.0	1.0
Decision Tree	Mean	0.98	0.95
	Median	0.98	0.96
	Mode	0.99	0.97
	Dropping Null	0.98	0.99
Random Forest	Mean	0.99	1.0
	Median	0.99	1.0
	Mode	1.0	0.97

	Dropping Null	0.99	0.99
SVC	Mean	0.91	0.96
	Median	0.89	0.97
	Mode	0.90	0.95
	Dropping Null	0.93	1.0
Naïve Bayes	Mean	0.95	0.95
	Median	0.96	0.96
	Mode	0.96	0.96
	Dropping Null	1.0	0.94
AdaBoost	Mean	0.99	0.99
	Median	0.99	1.0
	Mode	0.99	1.0
	Dropping Null	1.0	1.0
XGBoost	Mean	0.99	0.99
	Median	0.99	0.99
	Mode	0.99	0.99
	Dropping Null	1.0	1.0
MLP	Mean	0.83	0.90
	Median	0.83	0.87
	Mode	0.83	0.86
	Dropping Null	0.36	0.96
LightGBM	Mean	1.0	1.0
	Median	0.99	1.0
	Mode	0.99	0.99
	Dropping Null	1.0	1.0

QDA	Mean	0.19	0.93
	Median	0.19	0.93
	Mode	0.19	0.93
	Dropped Null	0.36	0.36



5.4 Comparison of Recalls among Different Algorithms

Recall is the ratio of predicted positive incidents which are correct to the all positive incidents. Recall answers the question of all the patients who truly have CKD, how many did the model detect.

$$\text{Recall} = \frac{TP}{TP+FN} \quad [41]$$

Comparison of recalls among different algorithms for both default hyperparameters and tuned hyper parameters is given below:

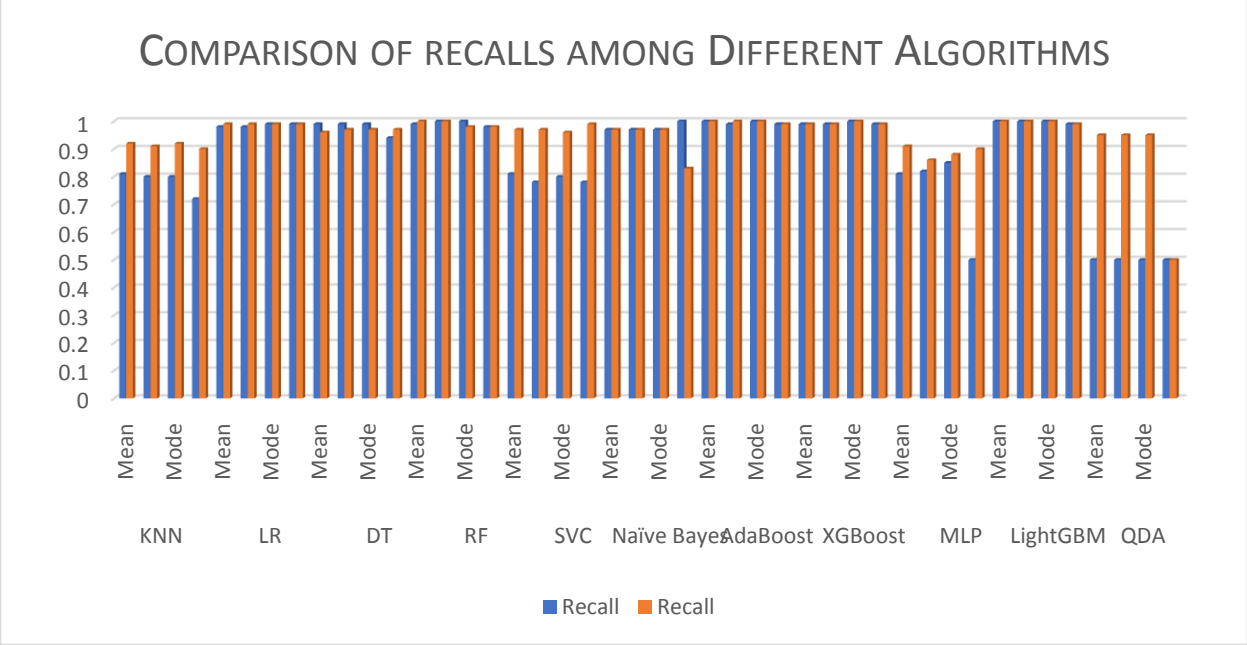
TABLE

COMPARISON OF RECALLS AMONG DIFFERENT ALGORITHMS

Algorithm	Null Filling Method	Recall (Without Tuning)	Recall (With Tuning)
KNN	Mean	0.81	0.92
	Median	0.80	0.91
	Mode	0.80	0.92
	Dropping Null	0.72	0.90
Logistic Regression	Mean	0.98	0.99
	Median	0.98	0.99

	Mode	0.99	0.99
	Dropping Null	0.99	0.99
Decision Tree	Mean	0.99	0.96
	Median	0.99	0.97
	Mode	0.99	0.97
	Dropping Null	0.94	0.97
Random Forest	Mean	0.99	1.0
	Median	1.0	1.0
	Mode	1.0	0.98
	Dropping Null	0.98	0.98
SVC	Mean	0.81	0.97
	Median	0.78	0.97
	Mode	0.80	0.96
	Dropping Null	0.78	0.99
Naïve Bayes	Mean	0.97	0.97
	Median	0.97	0.97
	Mode	0.97	0.97
	Dropping Null	1.0	0.83
AdaBoost	Mean	1.0	1.0
	Median	0.99	1.0
	Mode	1.0	1.0
	Dropping Null	0.99	0.99
XGBoost	Mean	0.99	0.99
	Median	0.99	0.99
	Mode	1.0	1.0

	Dropping Null	0.99	0.99
MLP	Mean	0.81	0.91
	Median	0.82	0.86
	Mode	0.85	0.88
	Dropping Null	0.50	0.90
LightGBM	Mean	1.0	1.0
	Median	1.0	1.0
	Mode	1.0	1.0
	Dropping Null	0.99	0.99
QDA	Mean	0.50	0.95
	Median	0.50	0.95
	Mode	0.50	0.95
	Dropped Null	0.50	0.50



5.5 Comparison of F1-scores among Different Algorithms

F1 Score means the weighted average of Precision and Recall. This performance metric takes both false positives and false negatives into calculation. F1-score gives more insights than accuracy if the model deals with an uneven class distribution [41]. Accuracy is the best performance parameter if false positives and false negatives have similar weight. But this is not true in diseases detection models. So, Precision and Recall are taken into account for our model. Therefore, F1-scores for eleven algorithms are calculated.

Comparison of F1-scores among different algorithms for both default hyperparameters and tuned hyper parameters is given below:

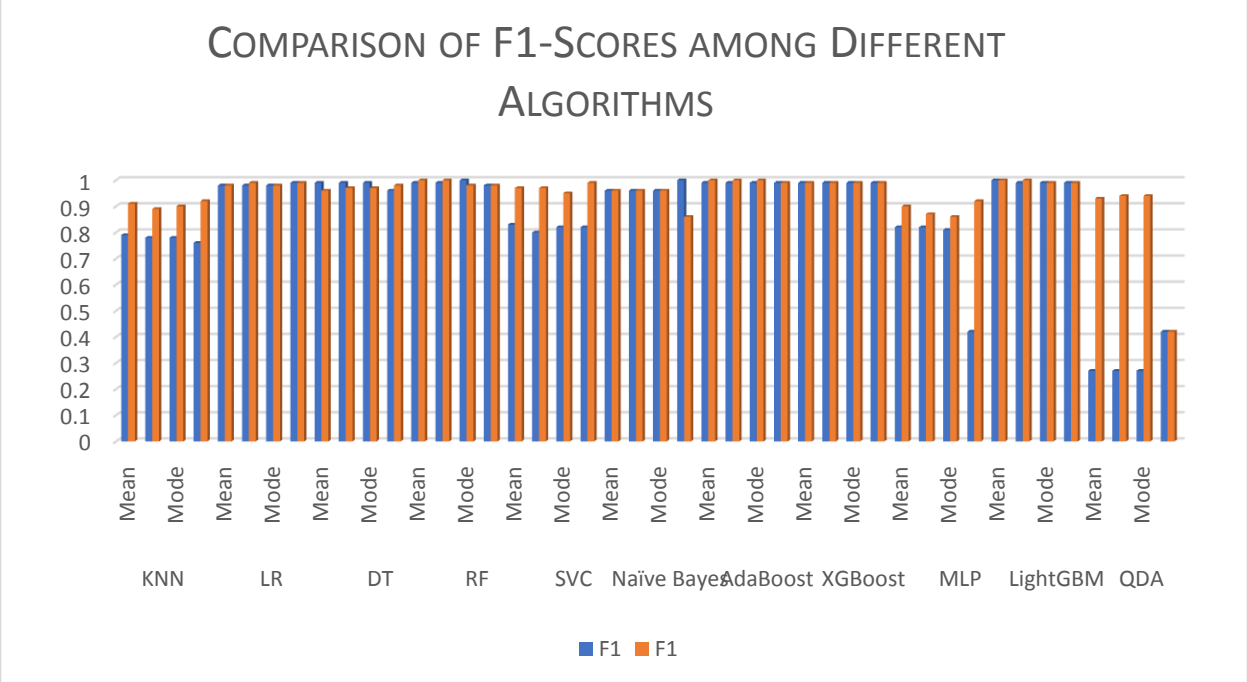
TABLE

COMPARISON OF F1-SCORES AMONG DIFFERENT ALGORITHMS

Algorithm	Null Filling Method	F1 (Without Tuning)	F1 (With Tuning)
KNN	Mean	0.79	0.91
	Median	0.78	0.89
	Mode	0.78	0.90
	Dropping Null	0.76	0.92

Logistic Regression	Mean	0.98	0.98
	Median	0.98	0.99
	Mode	0.98	0.98
	Dropping Null	0.99	0.99
Decision Tree	Mean	0.99	0.96
	Median	0.99	0.97
	Mode	0.99	0.97
	Dropping Null	0.96	0.98
Random Forest	Mean	0.99	1.0
	Median	0.99	1.0
	Mode	1.0	0.98
	Dropping Null	0.98	0.98
SVC	Mean	0.83	0.97
	Median	0.80	0.97
	Mode	0.82	0.95
	Dropping Null	0.82	0.99
Naïve Bayes	Mean	0.96	0.96
	Median	0.96	0.96
	Mode	0.96	0.96
	Dropping Null	1.0	0.86
AdaBoost	Mean	0.99	1.0
	Median	0.99	1.0
	Mode	0.99	1.0
	Dropping Null	0.99	0.99
	Mean	0.99	0.99

XGBoost	Median	0.99	0.99
	Mode	0.99	0.99
	Dropping Null	0.99	0.99
MLP	Mean	0.82	0.90
	Median	0.82	0.87
	Mode	0.81	0.86
	Dropping Null	0.42	0.92
LightGBM	Mean	1.0	1.0
	Median	0.99	1.0
	Mode	0.99	0.99
	Dropping Null	0.99	0.99
QDA	Mean	0.27	0.93
	Median	0.27	0.94
	Mode	0.27	0.94
	Dropped Null	0.42	0.42



5.6 Comparison of AUC-ROC among Different Algorithms

Area Under the curve of receiver Operating Characteristics is one of the most important performance metrics for checking classification models' performance. ROC is a curve of probability and AUC shows the degree of separability. It represents the capability of the model in distinguishing between classes. Higher value of the AUC means the model is better at predicting. On the other hand, when a model has AUC near to the 0, it will give poor performance in predicting [42].

Comparison of AUC-ROC among different algorithms for both default hyperparameters and tuned hyper parameters is given below

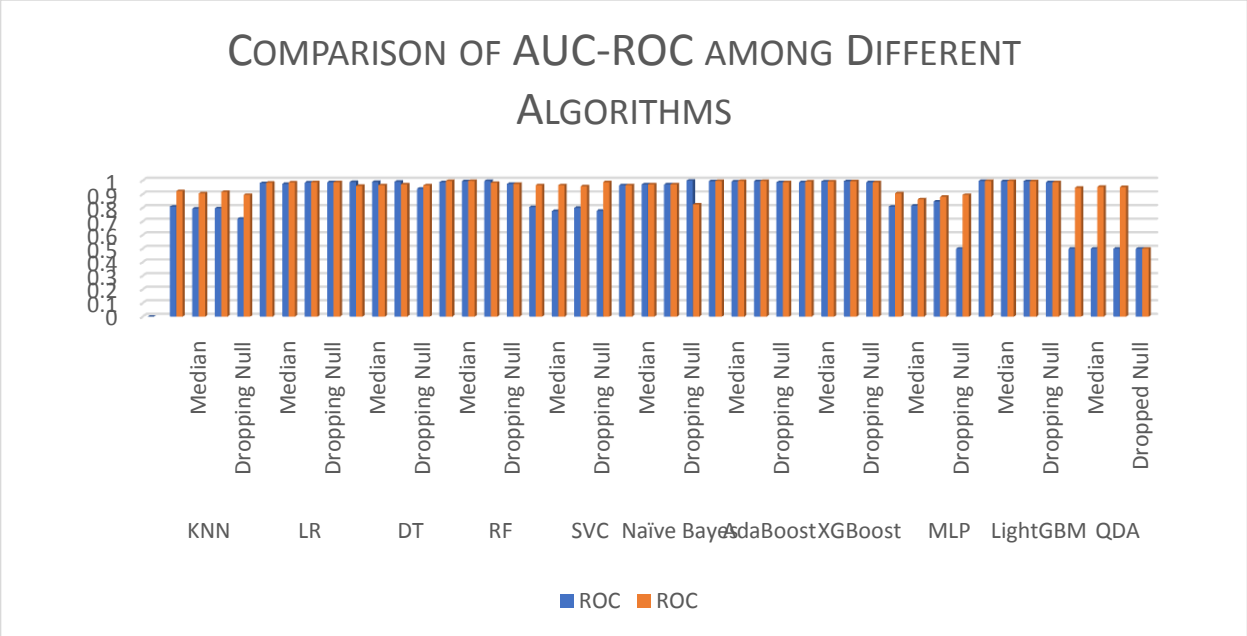
TABLE

COMPARISON OF AUC-ROC AMONG DIFFERENT ALGORITHMS

Algorithm	Null Filling Method	ROC (Without Tuning)	ROC (With Tuning)
KNN	Mean	0.809	0.922
	Median	0.795	0.906
	Mode	0.797	0.917
	Dropping Null	0.720	0.895
Logistic Regression	Mean	0.982	0.985

	Median	0.976	0.986
	Mode	0.986	0.988
	Dropping Null	0.988	0.988
Decision Tree	Mean	0.99	0.961
	Median	0.99	0.966
	Mode	0.992	0.973
	Dropping Null	0.941	0.965
Random Forest	Mean	0.989	0.998
	Median	0.996	0.998
	Mode	0.998	0.984
	Dropping Null	0.976	0.976
SVC	Mean	0.806	0.966
	Median	0.776	0.966
	Mode	0.80	0.958
	Dropping Null	0.779	0.988
Naïve Bayes	Mean	0.966	0.966
	Median	0.972	0.972
	Mode	0.972	0.972
	Dropping Null	1.0	0.825
AdaBoost	Mean	0.996	0.998
	Median	0.994	0.998
	Mode	0.996	0.998
	Dropping Null	0.988	0.988
XGBoost	Mean	0.988	0.994
	Median	0.994	0.994

	Mode	0.996	0.996
	Dropping Null	0.988	0.988
MLP	Mean	0.809	0.907
	Median	0.816	0.863
	Mode	0.846	0.881
	Dropping Null	0.5	0.895
LightGBM	Mean	0.998	0.998
	Median	0.996	0.998
	Mode	0.996	0.996
	Dropping Null	0.988	0.988
QDA	Mean	0.5	0.948
	Median	0.5	0.954
	Mode	0.5	0.952
	Dropped Null	0.5	0.5



After investigating all the performance metrics for all the algorithms, Best performance metrics among all the algorithms are given below:

TABLE

BEST PERFORMANCE METRICS

Performance Metric	Algorithm [Without Tuning]	Algorithm [With Tuning]
Accuracy	LightGBM 99.75%	Random Forest AdaBoost LightGBM 99.75%
Precision	Decision Tree Random Forest Adaboost XGBoost LightGBM 0.99	Logistic Regression Random Forest Support Vector Classifier Adaboost XGBoost LightGBM 1.0
Recall	Naïve Bayes Random Forest Adaboost	Random Forest Adaboost XGBoost

	XGBoost LightGBM 1.0	LightGBM 1.0
F1-Score	Naïve Bayes Random Forest LightGBM 1.0	Random Forest Adaboost LightGBM 1.0
AUC-ROC	Random Forest LightGBM 0.998	Random Forest AdaBoost LightGBM 0.998

From the table we can see that Random Forest, AdaBoost, XGBoost and LightGBM algorithms performed better in all performance evaluation. In some performance metric Decision Tree and Naïve Bayes algorithm showed better performance.

TABLE

COMPARISON OF ACCURACIES AMONG OTHER RESEARCH WORKS

Research Paper	Algorithm	Best Accuracy	Our Best Accuracy
Avci, E., Karakus, S., Ozmen, O., &Avci, D. (2018, March). Performance comparison of some classifiers on Chronic Kidney Disease data. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-4). IEEE	J48 Classifier	99%	**99.75%**
Gunarathne, W. H. S. D., K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi."Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)." 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017	MDF	99.10%	
Tazin, Nusrat, Shahed AnzarusSabab, and Muhammed Tawfiq Chowdhury. "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique." 2016 International Conference on	Decision Tree	98-99%	

Medical Engineering, Health Informatics and Technology (MediTec). IEEE, 2016			

Chapter-6

Conclusion and Future Works

Kidney is an indispensable organ of human body and chronic kidney diseases has become a prime and trending area in medical research because of it being one of the leading reasons of mortality. Hundreds of thousands of lives can be saved and severe complexity can be avoided if we are able to detect kidney diseases in its premature state. Consequently, patients can be steered clear of procedures like kidney transplant and dialysis which cannot offer concrete safety. Applying machine learning algorithms on computer aided diagnosis system will be of great service in predicting chronic diseases of kidney. We have analyzed twenty-five attributes associated with kidney disease and implemented eleven machine learning algorithms to detect chronic kidney disease. The major upper hand of this study is the consistency of accuracy while working with an extensive number of features.

However, the models are to be inspected in a much larger scale before it is used as a clinical assistant to medical professionals. Further extension of this work can be done by using a larger collection of data which can help in detecting chronic kidney diseases early and more appropriately. In future, we aim to make our model more interactive which will take live input from a device or sensor instead of a dataset and give the corresponding prediction based on the trained model.

References

[1] https://en.wikipedia.org/wiki/Chronic_kidney_disease#:~:text=Chronic%20kidney%20disease%20(CKD)%20is,loss%20of%20appetite%2C%20and%20confusion.

[2] National Chronic Kidney Disease Fact Sheet,2017

URL-https://www.kidneynews.org/careers/leading-edge/cdc-releases-2017-national-ckd-fact-sheet

[3] https://www.nhs.uk/conditions/kidney-disease/diagnosis/#:~:text=The%20main%20test%20for%20kidney,to%20filter%20in%20a%20minute.

[4] Avci, E., Karakus, S., Ozmen, O., &Avci, D. (2018, March). Performance comparison of some classifiers on Chronic Kidney Disease data. In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)* (pp. 1-4). IEEE.

[5] Gunarathne, W. H. S. D., K. D. M. Perera, and K. A. D. C. P. Kahandawaarachchi. "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (CKD)." *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2017.

[6] Arasu SD, Thirumalaiselvi R. A novel imputation method for effective prediction of coronary Kidney disease. In2017 2nd International Conference on Computing and Communications Technologies (ICCCT) 2017 Feb 23 (pp. 127-136). IEEE.

[7] Rubini LJ, Eswaran P. Generating comparative analysis of early stage prediction of Chronic Kidney Disease. International Journal of Modern Engineering Research (IJMER). 2015 Jul;5(7):49-55.

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

[9] Polat, Huseyin, HomayDanaeiMehr, and Aydin Cetin. "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods." *Journal of medical systems* 41, no. 4 (2017): 55.Polat, Huseyin, HomayDanaeiMehr, and Aydin Cetin. "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods." *Journal of medical systems* 41, no. 4 (2017): 55.

[10] Tazin, Nusrat, Shahed AnzarusSabab, and Muhammed Tawfiq Chowdhury. "Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique." *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*. IEEE, 2016.

[11] Yildirim, Pinar. "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction." *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 2. IEEE, 2017.

[12] Devika, R., Sai Vaishnavi Avilala, and V. Subramaniyaswamy. "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest." *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019.

[13] Subasi, Abdulhamit, EminaAlickovic, and Jasmin Kevric. "Diagnosis of chronic kidney disease by using random forest." *CMBEBIH 2017*. Springer, Singapore, 2017. 589-594.

[14] Pasadana, I. A., et al. "Chronic kidney disease prediction by using different decision tree techniques." *Journal of Physics: Conference Series*. Vol. 1255. No. 1. IOP Publishing, 2019.

[15] Başar MD, Sarı P, Kılıç N, Akan A. Detection of chronic kidney disease by using Adaboost ensemble learning approach. In2016 24th Signal Processing and Communication Application Conference (SIU) 2016 May 16 (pp. 773-776). IEEE.

[16] Indriani, A. F., & Muslim, M. A. (2019). SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis. *Lontar Komputer*, *10*(2).

[17] Ogunleye AA, Qing-Guo W. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM transactions on computational biology and bioinformatics. 2019 Apr 17.

[18] http://archive.ics.uci.edu/ml//datasets/Chronic_Kidney_Disease?fbclid=IwAR31W_qpblp76lJqa1ykeYmdEeG_T6AlVHhzNLlNazaMynL2yE6IpEzzdUM

[19] https://en.wikipedia.org/wiki/Supervised_learning

[20] https://en.wikipedia.org/wiki/Unsupervised_learning

[21] https://en.wikipedia.org/wiki/Reinforcement_learning

[22] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[23] https://en.wikipedia.org/wiki/Logistic_regression

[24] Albon, Chris. Machine learning with python cookbook: Practical solutions from preprocessing to deep learning. " O'Reilly Media, Inc.", 2018.

[25] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[26] https://en.wikipedia.org/wiki/Support_vector_machine

[27] https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4#:~:text=Decision%20Tree%20algorithm%20belongs%20to%20the%20family%20of%20supervised%20learning%20algorithms.&text=The%20goal%20of%20using%20a,prior%20data(training%20data).

[28] https://en.wikipedia.org/wiki/Decision_tree_learning

[29] Dangeti, Pratap. Statistics for machine learning. Packt Publishing Ltd, 2017.

[30] https://en.wikipedia.org/wiki/Random_forest#Bagging

[31] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[32] https://towardsdatascience.com/boosting-algorithm-adaboost-b6737a9ee60c

[33] https://en.wikipedia.org/wiki/AdaBoost

[34] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d#:~:text=XGBoost%20is%20a%20decision%2Dtree,all%20other%20algorithms%20or%20frameworks.

[35] https://en.wikipedia.org/wiki/Multilayer_perceptron

[36] https://lightgbm.readthedocs.io/en/latest/Features.html

[37] https://www.quora.com/How-does-the-LightGBM-algorithm-work-conceptually

[38] https://en.wikipedia.org/wiki/Quadratic_classifier#Quadratic_discriminant_analysis

[39] https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10

[40] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[41] https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

[42] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[43] Levey, Andrew S., and Josef Coresh. "Chronic kidney disease." The lancet 379.9811 (2012): 165-180.

[44] El Nahas, A. Meguid, and Aminu K. Bello. "Chronic kidney disease: the global challenge." The lancet 365.9456 (2005): 331-340.

[45] https://www.expert.ai/blog/machine-learning-definition/