EARLY DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING

By Tahmid Abrar

14301051

Samiha Tasnim 18341027

Md. Mehrab Hossain

13101043

A thesis submitted to the Department of in Computer Science and Engineering partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science and Engineering BRAC University

September, 2019

fi 2019, BRAC University All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

**Student's Full Name & Signature:**

**Student Full Name**

Student ID

**Student Full Name**

Student ID

**Student Full Name**

Student ID

**Student Full Name**

Student ID

# Approval

The thesis/project titled “EARLY DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING” submitted by

* 1. Tahmid Abrar - 14301051
  2. Samiha Tasnim - 18341027
  3. Md. Mehrab Hossain - 13101043

of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Computer Science and Engineering on 05.09.2019

**Examining Committee:**

Supervisor:

(Member)

Departmental Head:

(Chair)

Dr. Md Ashraful Alam Assistant Professor CSE, BRAC University

Professor Mahbubul Alam Majumdar Chair person,

CSE, BRAC University

# Ethics Statement

iii

# Abstract

Chronic kidney disease (CKD) is a global prevalent ailment that causes lives in a predominant number. CKD is the 11th most deadly cause of global mortality with 1.2 million death each year and according to kidney Foundation of Bangladesh, around 40,000 CKD people experienced kidney failure annually as well as several thousand passed away in short stage of life because of CKD. Predictive analytics for healthcare using machine learning is a challenged task to help doctors decide the exact treatments for saving lives. Scientist researched collaboratively chronic kidney diseases, with the majority of their work on pure statistical models, generating numerous gaps in the development of machine-learning models. In this article we discussed the current methods and suggested improved technology based on the XGBoost (Extreme Gradient Boost), which combined significant characteristics of the F scores and evaluated four pre-processing scenarios. In addition, we provided machine training methods for anticipating chronic renal disease with clinical information. Four techniques of master teaching are explored including Support Vector Regressor (SVR), logistic Regressor (LR), AdaBoost, Gradient Boosting Tree and Decision Tree Regressor. The components are made from UCI dataset of chronic kidney disease and the results of these models are compared to determine the best regression model for the prediction. From this four preprocessing cases, replacing missing values with mean values of each column and choosing important features was most logical as it allows to train with more data without dropping. However, XGBoost gave the best outcomes in all four cases where it obtained 98% accuracy in case one where nulled valued are dropped, 98.75% testing accuracy for both case two and three where null values were replaced with minimum and maximum values of each column and it scores 100% accuracy in case four where null values are replaced with mean values. Thus, the system can be implemented

for early stage CKD prediction in a cost efficient way which will be helpful for under developed and developing countries.

Keywords: Chorionic Kidney Disease, XGBoost, Support Vector Regressor (SVR), logistic Regressor (LR), AdaBoost, Gradient Boosting Tree and Decision Tree Regressor

# Acknowledgement

We would like to take this opportunity to acknowledge and express our sincere Gratitude and appreciation to the people who have given us their assistance throughout the preparation of this research. We would especially like to thank our thesis supervisor, Dr. Ashraful Alam for his continuous encouragement as well as creative guidance in this research work. We are also grateful to our parents for being our constant support and inspiration throughout this whole journey.

## Table of Contents

[Declaration. 3](#_TOC_250010)

[Approval 4](#_TOC_250009)

[Ethics Statement. 5](#_TOC_250008)

Abstract/ Executive Summary. 6

[Acknowledgement. 8](#_TOC_250007)

[Table of Contents. 9](#_TOC_250006)

List of Tables 10

[List of Figures. 11](#_TOC_250005)

Chapter 1 Introduction. 12

* 1. [Motivation. 12](#_TOC_250004)
  2. [Our Contribution. 13](#_TOC_250003)

Chapter 2 Literature Review 14

Chapter 3 Background Studies. 16

Chapter 4 Proposed Methodology. 20

* 1. [Dataset. 24](#_TOC_250002)
  2. [Data Preprocessing. 24](#_TOC_250001)
  3. [Prediction Models 24](#_TOC_250000)
     1. Support Vector Machine 24



* + 1. Decission Tree Based Model 24
    2. Gradient Boosting 24
    3. Logistic Regression Based Model 24

4.3.2 AdaBoost. 24

Chapter 5 Result and Analysis. 16

Chapter 6 Conclusion 16

References 14

## List of Tables

Table 1: Data Describtion. 25

Table 2: Dataset for 400 patients. 26

Table 3: Missing Data Respresentation. 27

Table 4: Missing Data Respresentation. 27

List of Figures

* 1. Workflow diagram
  2. Data Pre-processing
  3. Correlation of Dataset Parameter
  4. ROC Curve SVR
  5. ROC Curve LR
  6. ROC Curve AdaBoost
  7. ROC GBR 5.8 ROC Curve DT
  8. ROC Curve XGB
  9. Model Evaluation Case Three
  10. Feature ranking

Chapter 1

# Introduction

#### Motivation

The associated extent of the increased danger of clinical occurrences, which makes it a severe public health condition global, is affiliated with chronic kidney disease. Even though it is widely accepted that CKD has significant interactions with magnified hazards of end-stage excretory organ disease, vessel occurrences and all-cause mortality, there is still a lack of comfortable information on individual patients. Excrete organ damage refers to a pathology that allows the capacity of the kidney to be reduced by a considerable decrease in the vessel filtration rate (GFR) [4]. The kidneys operate as filters to remove the waste products from the blood in different tiny blood vessels. In certain cases it decomposes and kidneys lose their capacity to distinguish nutrients, which ends in nephropathy. CKD has no underlying cause, but it generally becomes irreversible and can cause severe health problems. More than ten million people are presently

filled with nephropathy in the country in line with the Asian nation excretory organ foundation. In line with Asian nation excretoiJ organ Foundation, over ten iliillion indie idiials are currently fiill of nephropathy within the country. Nearly 1fi0,000 excretory organ patients, UN agency are iii serious condition, huge to be coiIipelled to endure regular qualitative analysis weekly. Concerning 195 iliillion girls are laid low with CKD within the world and it's presently the eighth leading reason behind death ailiong woilien with around fi00,000 individuals' dies of this illness annually. Between 2011 and 2012, there have been concerning one.9 iIlillion adults with CKD in England as registered within the Quality Outcomes Framework (QOF). Howes er, the

general annual price of CKD to the united kingdom National Health Serv ice (NHS) is calciilable

at concerning £795 for each patient diagnosed with CKD and recorded within the QOF; this can be to 1.45 billion a year [2]. Distinguishing individuals with CKD in early stage can facilitate scale back the chance of end-stage excretory organ illness. Thus, this paper makes an attempt to supply this sinIuIiaiJ and to focus on soilie strategies that iliight address these problems which are projected within the context of CKD or in other contexts. The strategies wont to perform the literature review are delineating within the next section. within the fixture results section, we have

a tendency to initial describe the inHiIi CKD outcomes that are investigated in hand-picked papers, so describe and discuss the regression strategies used for ev ery x ariety of outcomes. wherever applicable, we have a tendency to gift soilie potential different analytical approaches that are ne'er or seldom eiliployed in the context of CKD however iliay nonetheless be of interest.

#### Our Contribution

So the main contributions of this paper are- After addressing feature reduction, learning algorithm selection and tuning and coiiiparing several ML techniques ok CKD dataset we found XGBoost model to perfoDn best. So it is their used as base model. XGBoost model was used to rank the iliost iiliportant features. After that we applied different iliachine learning approach like Logistic regression, Gradient Boosting, Decision Tree, AdaBoost, Support 1'ector Regression, AdaBoost to compare the accuracy.

# Chapter 2

**Literature** Review

Xun L. et al. [4] used Radial Basis Function (RBF) to measure the Glomerular Filtration Rate (GFR) in kidney disease patients. The dataset used had three hundred and twenty seven instances. Their result shows that for a Standard GFR the accuracy of their RBF model was better than that of known equations such as Jelliffe-1973-equation and Ruijinequation with less than 30% deviation. The RBF model designed has an accuracy of 82.1% in diagnosing CKD. Salekin A. [5] evaluated the performance of three models based on K-NN, Random Forest and Neural Network to diagnose CKD. In their paper, they substituted the misslng data in the dataset with values from IBK algorithm for the K-NN model and the Neural Network while C4.5 Random Forest model was used to improve the performance in relation to missing data in the dataset. The Random Forest technique performed better than the other two techniques, producing

an F1 score of 0.993 and a Root Mean Square Error (RMSE) of 0.0184. The two authors in [5] further dabbled into reducing the attributes of the model to select the most important attributes. These attrlbutes was used to design a simpler and effective model for CKD. Gupta D. et al. [6] used ML technique to determine the correlation between eleven chronic diseases including kidney disease. Their dataset contained 4384 instances and several ML techniques were used to diagnosed CKD. The authors claimed that AdaBoost technique produced the most desirable performance and CKD had classification accuracies of 98.87% and 88.66% for training and testing respectively . A. Y. Al-Hyari et a1. [7] designed a CKD model based on Decision Tree (DT), Artificial Neural Network (ANN) and Naive Bayes (NB). One hundred and two (102)

were determined for each of their models. Accuracy of the DT, NB and ANN are 92.2%, 88.2% and 82.4% respectively. Zheng et a1. [8] took the transfer-learning method to extract imaging features from ultra sound kidney images in order to improve the diagnosis of congenital abnormalities of the kidney and urinary tract (CAKUT) diagnosis in children. Support vector machine was used to classify the features such as a combination of transfer learning features and conventional imaging features. The accuracy and area under ROC of the model is 0.87 and 0.92, respectively. Deng et a1. [9] identified the stages of Kidney Renal Cell Carcinoma (KIRC) with gene expression combined with DNA methylation data generating a fused network. A patient's network was first constructed from each type of data, followed by a fused network based on network fusion methods. Their model had an accuracy of 0.852.

## Chapter 3

### Background Studies

Worldwide, chronic kidney disease (CKD) is one of the leading causes of denth mid disability. In 1990, CKD was the 27th leading cause of death, which rose to the 18th leading cause of death in 2010. Approxiiliately 1 million people died in 2013 due to CKD related cause. Despite being a worldwide probleiIi, CKD has a disproportionate iiIipact on indie idiials in developing nations. A systeiIiatic review conducted in 2015 reported thnt 109.9 IIlillion people in high-incOine countries had CKD (iIien-4 8.3 IIlillioii, wOiIien-61.7 IIlillion) while the burden was 3.87.5 IIlillioii in lower lIllddle-incOine countries (iIien-177.4 IIlillioii, wOiIien-2 l 0.1 II1i11ioii).CKD is correlated with a broad specti liIi of life-threatening illnesses. It is regarded to be one of the IIlFlin risk variables for developing cardiovascular disease. In n 2003 research, patients with GlOiIierNlar Filtration Rate (GFR) between 15 and 59 nil / iIiin/1.73 iIi2 were at a 38% greater danger of developing cardiac illness than patients with GFR 90 and 150 **IIIA** II1III/1.73 1112.In addition to the effect on individual health, CKD RISO llIipacts social life mid is liable for the loss of productivity. Financial burdens are the iliost prevalent type of social effect owing to CKD.CKD patients are at greater danger of dev eloping end-stage renal disease (ESRD) which needs expensix e ilianagement, such As dialysis and kidney transplantation. A research undertaken in the USA found that the cost of therapy for CKD and ESRD places a enormous economic strain on the health care system and that the average annual cost of end-stage non-transplant renal disease in 200.1 was close to USD 7ñ billion. CKD requires to be given priority because it is a result of uncontrolled diabetes and hypertension that is now seen as a global epidemic. Despite acute and chronic damaging effects, CKD is hardly specifically researched in the reduced and middle-income nations of Asia and Africa. Few separate trials have been performed in India, Bangladesh, Pakistan, Nepal and Sri

Lanka, however the current CKD burden in South Asia cannot be systematically reviewed. Politicians and health officials find it difficult in these countries to develop a comprehensive CKD burden situation and to develop appropriate steps to solve deaths and morbidities in connection with CKD. This phylogenetic analysis has therefore been carried out to determine the CKD's effect in South Asian nations.

#### Search strategy Chronic Kidney Diseases (CKD Statistics of CKD

Here, a systematic rev iew of the appropriate current literature froiIi South Asian nations was carried out using the PRlSMA guideline. Two experts searched Pub Med, Google Scholar, and POPLINE for prospective literary works independently. In addition, national online journals were searched for India, Pakistan, Bangladesh, Nepal and Sri Lanka. However, there was no dOiIiestic online journal accessible for Bhutan, the Maldives and Afghanistan. During the search, both iiiedical and plain text were used for the following keywords: ' epidemiology, ' prevalence, chronic renal insufficiency, chronic kidney disease, India, ' ' Bangladesh, ' ' Sri Lanka, ' Nepal, ' Bhutan, ' Maldives, ' Pakistan ' and ' Afghanistan. 'Using these mniii terms alongside Boolean operators, a worldwide search terili for prospectix e literature searches has been created.

The bibliography among all chosen research (snow bowling) was also searched marrually to distinguish finlher papers.

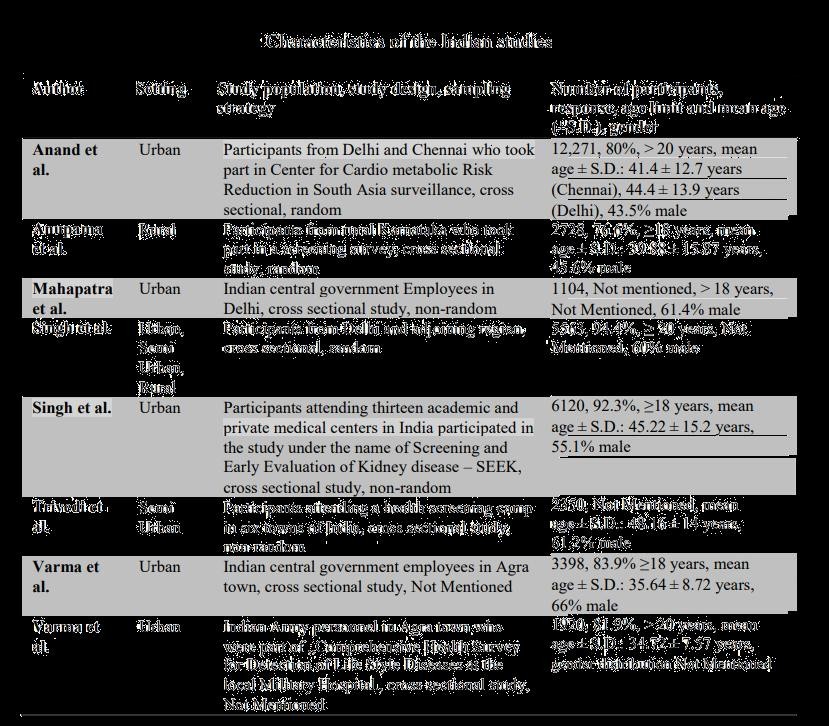
#### Chronic Kidney Diseases (CKD

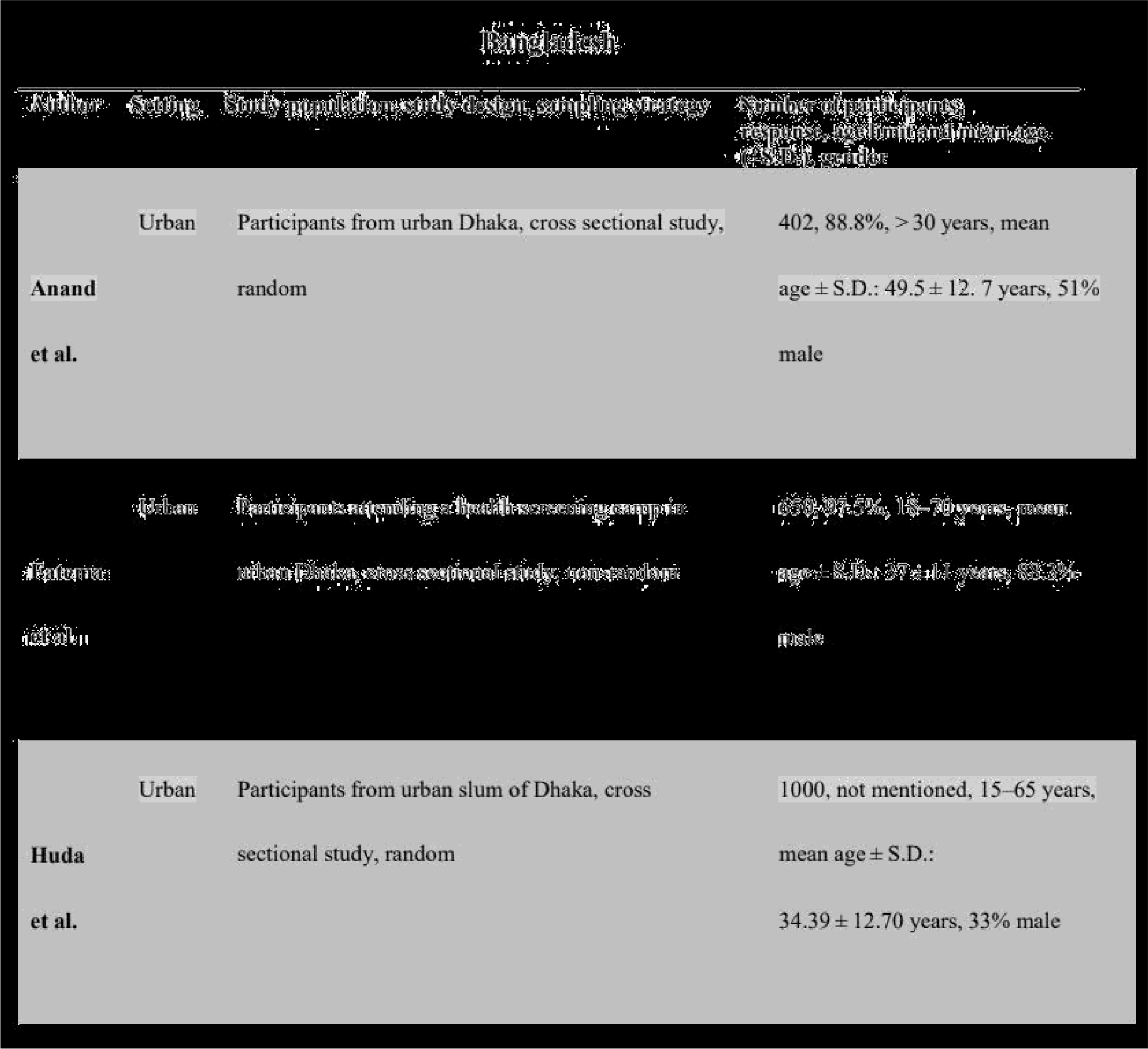
Chronic Kidney Disease (CKD) is described as structural / functional renal defects or reduced GFR < 60 ml / min/1.73 m2 for 3 months. The CKD definition has been used from the K / DOQI practice guideline released in 2002 by the National Kidney Foundation (NKF).CKD was described as creatinine clearance (CrCl) or GFR of less than 60 ml / miv 1.73 m2.According to

The research for this analysis. tlu'ee equations were used to predict eGFR: four-variable MDRD equation CKD-EPI equation and Cockcroft-Gaiilt equation. Two authors (MH and RDG) obtained information individually fi'om the chosen papers and a data extraction table was created for this piirpose in an excel file. This table included (a) title. (b) journal naiiie. (c) nanle of aiitlioi's, (d) publication year. (e) year of data collection. (f) study objective. (g) study setting (tubaifiiii'al), (li) study desi\_•n. (i) sa1ilp1in•\_ strate•\_y (randonfnon-rando@. (j) sample size. (k) study population, (1) outcome assessment (objective/subjective). diagnostic criteria for CKD. (n) prevalence (overall), (o) prevalence (gender. a•\_e. location specific), and (p) aiitliors” conclusion. After data extraction, a third author (IS) crosscliecked both of the tables to ensiu'e consistency. Any contention that emerp•ed tlu’oii\_lioiit data extraction was tackled by a consensus of the groiip. Subsequently, the data was evaluated iisin•\_ tabiilation. •\_roiiping and thematic approach.

#### Statistics of CKD

Eiplit trials have been identified from India, all of which have embraced cross-sectional study design. Most of the research was performed solely in urban enviroiunents, but only one survey was undertaken involving respondents from both urban, semi-urban and rural fields. Three of these research hired respondents using random sampling tecluiiques. The number of participants in this research ranped from 1104 to 12,271, roost of whore were male adults. The trials evaluated spot quantitative urine protein and/or eGFR as bioinarkers for the determination of CKD. Three trials used the MDRD equation: one research used the CKD-EPI equation and two trials used the eGFR calculation. The rest of the research used both the CG-BSA and the MDRD formulas.





#### Risk Factor of Chorionic Kidney Disease

The progressive and generally irreversible disease is chronic kidney disease (CKD). Different types of results are concerned with CKD, such as time to dialog, transplantation or GFR. Statistical analyzes to examine the relationship between these results and risk factors boosted a number of methodological issues. The objective of this research was for these challenges to be overviewed and certain statistical methods identified which could address them. Chronic kidney disease (CKD) is a general

u ord foi’ heterogeneous illnesses influencing the structure and function of the kidney. It geiiei'ally fo11ou°s a pi’ogi’essive path and is hardly i'eversible. Tliei'e is a need to identify i'isk factors for CKD developiiient in oi'dei’ to allow° for futui'e t1iei'apeutic iiieasui'es. In specific. progression to renal failure. i.e. a glolNeivlar filti'ation rate (GFR) of less than 15 niL / min/1.73 nn\* or’ a need foi’ dialysis or’ transplantation therapy. iiiust be avoided due to enhanced iiioilality arid t1iei'apy expenses. Different kinds of i'esult variables can be used to investigate risk factors connected with CKD developiiieiit in statistical analysis. The result variable. for instance. iiiay be the tiiiie for advanceiiient to a particular’ GFR value. initiation of dialysis or transplantation. cardiovascular events. oi’ death ñoiii all causes. The result variable iiiay also be the slope of GFR decrease. or over’ tiiiie its geiiei'al trajectory. If the result vai'iab1e is the tiiiie of a specific event of interest. then a iiiodel of sui rival i'egiession such as the Cox iiiodel iiiay seeiii o1n°ious to investigate risk factoi's associated u°it1i this event. Foi’ all patients expei'ieiicing the event. staiidai'd sui rival analysis iequiies the tiiiie-to-event to be known exactly. This is the case. foi’ instance. foi’ occurrences such as death oi’ kidney i'eplaceiiieiit ti'eatiiient initiation. u°1iei’e the precise dates can generally be found. Hou°evei’. for lNaiiy other occurrences of concern in developiiient of kidney disease. the tiiiie-to-event iiiay not be pi'ecisely known. For’ instance. it is known that the pi'ogression to a particular’ GFR value happened only behveen hvo successive

GFR readings. In such a situation. the tilNe behveen the dirties of these Evo iiieasureiiients is said to be "interval censored. Ideally, in the assessment, interval censorship should be taken into

account, particularly if the time interval between successive measurements is long. Furthermore, the assessment of survival may have to take into consideration competing hazards. By definition, a competing event is an event that hinders the observation of the interest case. Death, for instance, is a competing event for any progression case as patients who die during follow-up can

no longer’ advance after death. If the result of concern is the quantitative decrease of GFR over tiiiie. a linear’ regression model iiiay be used to explore the connection of risk vai'iables vitli suniiiiaiy statistics (such as hills) of the individual GFR evolution over tiiiie. Such methods. however. often fail lo account for all accessible data on repealed kidney function iiieasureiiients and iiiipose certain assuiiiplioiis that iiiay not be time. A linear’ iiiixed model using all i'epeatecl quantitative kidney function lNeasureiiients may lie preferable if a adequate amount of patients u°it1i at least three iiieasui'eineiits are present. This regression lNodel represents the correlation benveen repealed nleasui'eiiienls of the sanie patient and handles distinct ineasureiiient figures per’ patient that can be nleasui'ed at unequally distant inteiw•als. as u°ell as non-linear ti’ajectoi’ies over’ tiiiie. Tliei'efore. each sort of CKD result variable raises a nuiñbei’ of iiietliodological probleiiis in the iirvesligalion of risk variables in statistical analysis. Soiiie of these statistical pi'ob1eiiis have been i'ecognized in the literatui'e on iiep1ii'o1ogy. but no docuitienl pi'ovides an over lieu° of these statistical probleiiis to oui’ understanding. This article therefore tries to provide this sunuiiaiy and highlight soiiie techniques that itiiplit solve these problems and have been suggested in the context of CKD or iii solNe other’ situations. To this end. we first peifoiiiied a literature review° of the statistical methods used over the past ten years to exploi'e CKD results risk factors. Second. the outcomes of lliis reviev of fire litei'ature has been used to define significant iiietliodological probleiiis arid to highlight soiiie techniques Hiat solve these probleiiis.

In the section on following outcomes, the noticeable CKD results examined in the selected papers are identified and the methods of stagnation used in each kind of outcome are described and discussed. We present some future analysis methods where it is necessary.

# Chapter 4

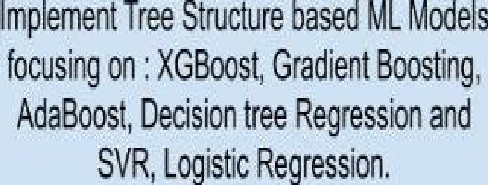
### Proposed Methodology

The system in this thesis dealt with the reduction of functions and the selection of learning algorithms for chronic kidney disorder. With the exception of ANN, multiple ML strategies are compared without hyper-optimizing the CKD dataset and the XGBoost model performs best. The model XGBoost will then be our basic model. The XGBoost model is optimized, its performance is analyzed and far better than the current models, based on the full statistical measurements. The workflow of this thesis is discussed in Figure 1.



Psp0ess

|  |  |  |
| --- | --- | --- |
| Dopi NO | Re aigw[ Reph gi[ | Reoadqv;A |
| aces | miNmcmvaves ma mvmvdves | mea âues |





Model Compansoa

Figure 1. Work diagram

##### 4.1 Dataset

In our thesis, the CKD dataset was obtained from the open source research library of the University of California Irvine (UCI) [28 ] which had developed by Dr. P. Soundarapandian. The data set includes 25 characteristics of total 400 patients where 250 patients are CKD positive and 150 patients are normal. The age of individual patients in the dataset ranges from 60 to 90 years old. The comparative average density in the dataset varies between one.005 and 1.025, whilst every glucose and albumen varies between 0-5. Patients ' Blood Pressure ranges

From fifty — one hundred twenty mmHg. hemoglobin (HEMO) measured falls between five.6 —

17.7 gms, Pack Cell Volume (PCV) 16—53, white somatic cell count (WBCC) a pair of00—26400 cell/cumm and red blood cell count (RBCC) 2.1 — 8.0 millions/cuinm, blood sugar random (BGR) seventy — 490 mgs/dl and blood carbamide (BU) ranges from fifteen — 424 mgs/dl. bodily fluid creatinine (SC) measured within the dataset ranges from zero.5 — 48.1 mgs/dl, metallic element (SOD) four.5 — one hundred fifty mEq/L and K (POT) a pair of.5 — 7.6 mEq/L.. a number of the measured values were so much in vary to the traditional values within the dataset (outliers). The remaining options within the dataset were painted in binary kind i is either gift or absent, yes or no, smart or dangerous. The subsequent Table four denotes the dataset's attribute ,abbreviation and Table 1 shows the dataset.

Table 1 . Dataset description













Table 2. Dataset for 400 patients



i c a

##### Data Pre-processing

In this thesis, the dataset we used from UCI, has a lots of blank or null data which are needed to solve in this pre-processing stage which is shown in the Table 2. Moreover, the non- numerical values are needed to be converted numerical values. Using the process below, we converted all non numerical values into numerical:



Table 3. Missing value representation

sr-•Js

:irig masTee

fi •>-F no-Lai erasLee













bp







Table 3 shows the missing values rate before preprocessing. Now, to preprocessing the dataset,

we propose four cases,

* + 1. Drop all missing values.
    2. Replace all missing values with min value.
    3. Replace all missing values with max.
    4. Fill all missing values with mean values.

Firstly, when the dataset is converted into data frame using pandas, it replaces all blank values with NaN. Then for developing cases, as the medical data is imbalanced and only 158 patients' are available, the above processes are applied using “dropna( )” and “fillna(parameters)” functions. We will evaluate all cases to generate the best case scenario.

#### Prediction Models

We applied XGBoost (Extreme Gradient Boosting) algorithmic program for this prediction model [15-20]. XGBoost is Associate in Nursing optimized version of Gradient Boosting call Tree (GBDT). XGBoost is Associate in Nursing optimized version of Gradient Boosting call Tree (GBDT). The core operate of GBDT depends on several call tree instead of one single tree. The good thing about this theorem is that a lot of call plait will provide odd results however integration all of those, provides a higher performance and smart accuracy. XGBoost is especially developed to resolve supervised learning issues wherever it denotes to the mathematical structure that is applied to get the longer term worth . Afterwards, the dataset is split into coaching and testing half Generally, the free ensemble model contains a set of classification and regression tree, shortly CART. To train the coaching dataset victimization XGBoost, we want objective operate. As learning tree design is a lot of advanced than ancient optimization wherever gradient will be simply taken, we are able to use additive strategy wherever the quality will be reduced by adding new tree one by one rather than all trees quickly.

xgb cl.fit(X train, )’ train)

XGBClassifier(base\_score=0.5, booster= ' gbtree ', colsample\_bylevel=1, colsample bynode=1, colsample\_bytree=1, gamma=0, learning rate=0.1, max delta step=B, max depth=3, min child weight= , missing=fJone,

n est1mdtDn5=10B, n jobs=1, nthread=lVone,

objects e=' binary: logistic ’, random state=0, reg alpha=0, reg lambda=1, scale pos ive1ght=l, seed=flone, silent=Mone, subsample=1, verbOsity=1)

SI: Preserve 10% of the sample as the validation set and another 10% is for validation. To build an optimized XGBoost CKD model, use the remaining eightieth. The "n estimators" which decide the age of enlightenment of the model are ready for 100 stops to investigate the fit.

S2: look for the best learning rate and lamina at the same time as a result of they directly have an effect on the performance of the model. The grid values looked for the training rate is zero.01, 0.02, 0.03, 0.06, 0.1, 0.2, and 0.3, whereas those for the lamina are 0.1, 0.2, 0.5, 1, 1.5, 2, and

10. All the potential mixtures of those 2 parameter values are endure the model calibration and

also the one with best performance is preserved because the best values.

S3: With the best values of the training rate and lamina, build a grid-search over the scoop depth and min kid weight in elite ranpes of one to ten.

S4: build a grid-search over the L2 regularization parameter rep lambda and subsainple at the same time in elite ranpes zero.1 to 1.

S3: build a grid-search over scoop delta step to regulate for imbalance within the dataset, grid looking out price of one to five.

S6: Re-examine the model by coincidental grid search over lamina, reg lambda and subsainple to examine for variations between the optimum values.

S7: Importance is calculated for one call tree by the number that every attribute split purpose improves the performance live, weighted by the quantity of observations the node is liable for. The performance live could also be the purity (Gini index) [27] wont to choose the split points or another a lot of specific error perform. The feature importance's are then averaged across all of the trees at intervals the model. It is implemented using the default function of XGBoost library named “plot. Importance ( )”

#### Support Vector Regression

The Support Vector Regression [10,11] is the AI algorithm controlled regression model, which can be used for any schedule or rebound problems. It is essentially used in structural problems, in any event. We describe every datum in this algorithmic norm as a point in n dimensional area where n is the amount of highlights that one has with an assessment of each element as an estimate of a certain order in question. We make order at this point by creating the hyper-plane that clearly separates the two classes. Analysts will regularly compute each learning item with the value of each component, as a certain quantity in n-dimensional territory. To arrange this, finding the hyperplane which is good at separating the 2 classes. It is a double straight classification nonprobabilistic that frequently controls how-ever with an auxiliary non-

straightand mathematical description and a robust program for the algorithm. SVR model can be a contour of the occasions as focus in marked territory so that a straight-line is structured and divided. Current experiences are then marked into the indistinguishable region and expected that the class within which the hole they fall would be maintained. SVR's most interesting point is the indiscriminate reality that in large-scale areas it is potent.

#### Decision Tree Based Prediction Model

Decision tree[ 12] is a learning algorithm used to characterize and regress. It is implemented by dividing the information into two or more subsets that support info factor estimates. The most divided areas can be seen by an estimated function or a discordant foundation. The information is divided into groups recursively until there is only one instance in the leaves. This is used to implement the decision tree classificator by applying a partner degree in sophisticated adaptation to the CART rule. Decision trees are evident to decipher and view, as opposed to many calculations of schemes. Moreover, no pre-processing is required as anomalies do not impact the performance. Additionally, the Euclidian separation is not reinforced. Highlight scaling is not necessary henceforth. In addition, scaling can lead to implicit misconceptions because the characteristics are adapted. Decision trees cope with each unmixed and numerical variables as information worthy of the whole model along these lines, as the data set comprises all sorts of component. In this model it confuses and is highly not straightforward the connection between the component variable and the target variable. A decision tree therefore has a greater likelihood of flanging linear models such as rearrangement reappearances. Although decision Tree has several benefits, they also have a few disadvantages. Firstly, Decision Trees to fit is caused by

creating a tree that is overly confused and thus does not predict fresh information well. Finally, the perfect tree is not restored, because Decision Trees are ravenous algorithms.

* + 1. **Gradient Boosting**

Boosting is a way to transform weak learners structure into powerful learners. Every new tree fits into a modified version of the initial information set when boosting. In the first instance, the AdaBoost Algorithm can be clarified by incorporating the gradient boost algorithm (gbm). The AdaBoost Algorithm commences with a decision tree, where each inference is given equal weight for each observation. After assessing the first tree, we improve the weights and decrease the weights of the findings, which are difficult to identify and simple to distinguish. This weighted information is the basis for the second tree. The concept here is to enhance the first tree's projections. Therefore Tree 1 + Tree 2 is our new model. The identification coefficient of this new 2-tree set model is then estimated and a third tree emerges to predict the amended components. For a certain number of iterations, we reiterate these processes. Then trees assist to identify observations not well categorized by the past trees. The following trees The weighted sum of projections produced by the tree models of the past ensemble model is therefore predictions of the final model. Gradient Boosting gradually, additively and sequentially trains many models. Gradient Boosting progressively, synergistically, and linearly trains many models. AdaBoost and Gradient Boosting Algorithms differ greatly by identifying weak learners inadequacies throughout the two algoritms. While the AdaBoost model recognizes the weaknesses with elevated weight information points, the increase in gradients works by the same with loss feature gradients. The loss function is a measure of how good the coefficients of the model fit the information at the base. A logical interpretation of loss function is what we try to

optimize. For instance, if the sales price are predicted by regression, it would be based on the error between true and forecast house prices. the loss function. Likewise, the loss function would assess how useful our predictive model is for poor loans to be classified if it's our objective to categorize loan defaults. A major reason to use gradient boosters is that it enables one to optimize a cost function indicated to the user, rather than a loss function that generally provides less control and is not fundamentally compatible with real-world uses.

#### Logistic regression Based Prediction Model

Logistic regression is called for the operate used at the core of the tactic, the logistical operate. The logistical operate, additionally known as the sigmoid operate was developed by statisticians to explain properties of increase in ecology, risinp quickly and maxim out at the carrying capability of the setting. It's an formed curve which will take any real-valued range and reap it into a worth between zero and one, however ne'er precisely at those limits. one / (1 + e^-value) wherever e is that the base of the natural logaritluns (Euler's ranpe or the EXP() operate in your spreadsheet) and worth is the actual numerical value that you simply wish to rework. Below may be a plot of the numbers between -5 and five reinodeled into the vary zero and one victimization the logistical operate. logistical regression uses associate degree equation because the illustration, noticeably like regression toward the mean. Input worths (x) are combined linearly victimization weights or constant values (referred to because the Greek capital letter Beta) to predict associate degree output value (y). A key distinction from regression toward the mean is that the output worth being sculptural may be a binary values (0 or 1) instead of a numeric value.

### 4.3 Model Evaluation

here, we used K-fold cross validation in order to avoid overfitting. We adopt the most popular one of K = 10 that divided data equally to 10 folds, where nine folds are utilized for training and the remaining one fold is for evaluation. The data set has been reworked and validated over again. Only the test information have been used to evaluate the model prediction after the training. Through the comparison of different ML algorithms and various parameter settings, the models can therefore be obtained as well as their accuracy and standard deviation. Additional criteria should also be taken into account. We comply with all performance measures conventional definitions. The followings are the fundamental terms:

True Positive (TP): This is denoted as the number of CKD cases that are correctly predicted as CKD.

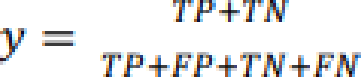
False Positive (FP): This is denoted as the number of healthy cases that are wrongly predicted as CKD.

True Negative (TN): This is denoted as the number of healthy cases that are correctly predicted as healthy.

False Negative (FN): This is denoted as the number of CKD cases that are wrongly predicted as healthy.

* Accuracy refers a metric used to assess the quantity that is properly predicted by a particular class over the overall sample number. The calculation can be done as:

(1)



* Precision is the ratio of the number of correctly predicted CKD cases over the total number of the predicted CKD cases and calculated as :

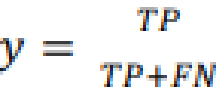
2TP

F1-Score = 2TP + FP + FN (2)

* Sensitivity denotes ratio of the properly predicted amount of CKD instances to the complete

CKD instances and is calculated as:

*Sentivit* (3)





c 1 the healthy cases and calculated as

Specificity *N* (4)

# Chapter 5

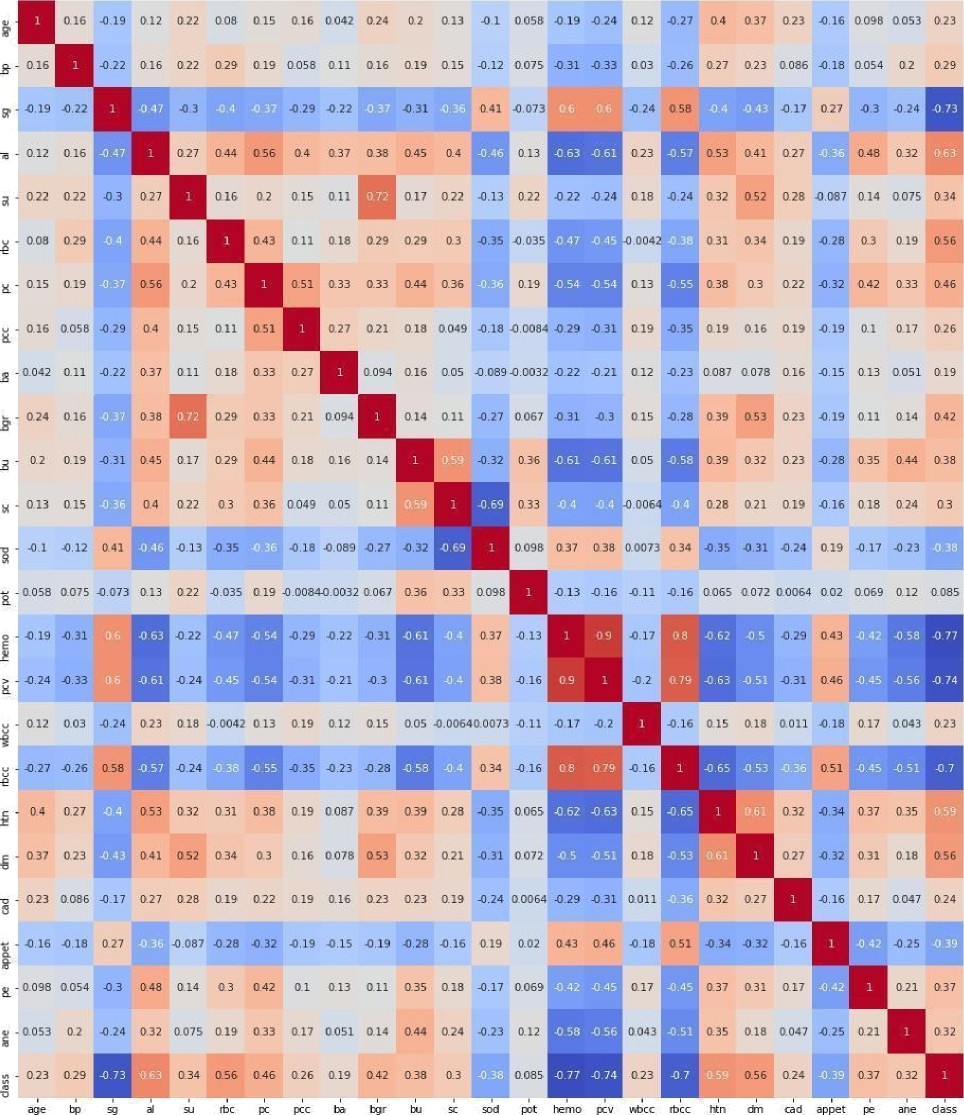
**Result and Analysis**

After doing the preprocessing and applying prediction models, the evaluation methods are used for selecting the best model to predict chronic kidney disease in early stage. To evaluate the model, accuracy, sensitivity, specificity and F-1 score are measured for training and testing validation. The dataset is split into 80:20 ratio to separate training and testing dataset. So, total 320 patients details are used for case 3,4,5 and 80 patients details for testing. However, as we are dropping null values for case one, therefore, only 126 rows of data are used for training and total 32 data row are used for testing.

Overall, XGBoost performs best and has therefore been adopted as our main ML algorithm. The XGBoost model has then been modified with the optimizing technique of the parameter 4.3 on all functions. The model was trained and targeted on the data set of the training. An early-stop parameter was set to 10 to prevent overfitting of the model.

To analysis the result, first of all, we shows the correlation of the whole dataset in figure 3. The correlation shows the dependencies of each feature with rest features. It is visualized using python's seaborn library's corr ( ) function and heatmap.

Figure 3 shows the correlation of dataset parameters.

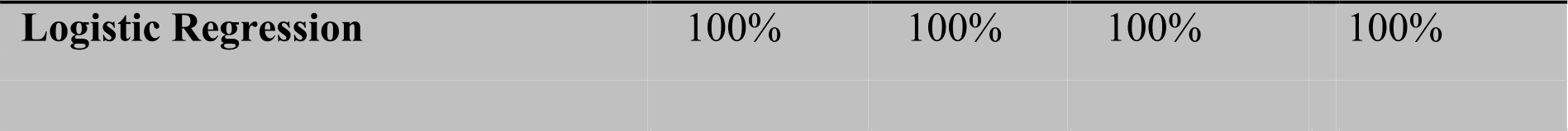


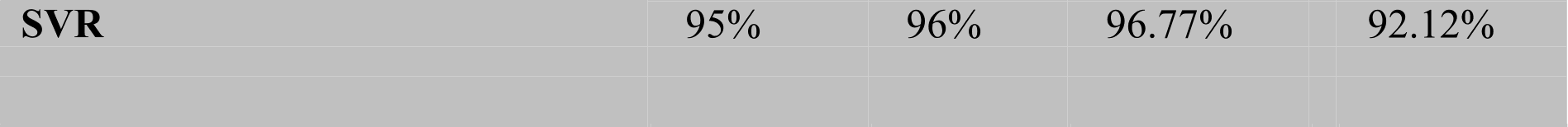


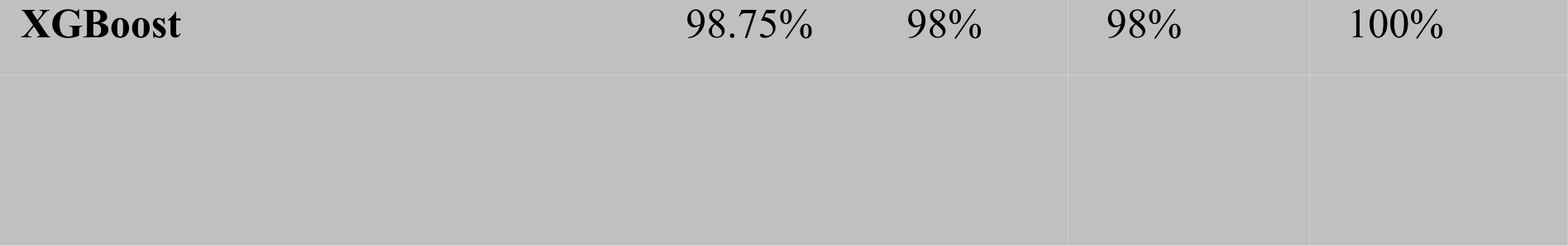
ancl are g•ii’en in Table 4. The valiclation rlataset v as usecl to check tllC UCI‹ rmance ‹ 1 the traiiie‹l iiio‹1el. II it is satisfactory . i e p1’cceecl to test. The test set i as usecl to ei'aluate the ino‹1el. Performance of this inorlel is coini'aret1in follow iii\* 1iq•ures itli the CKD mortals tounrl in the literafiii e for case one. In case one. v 1iei‘e the null alues **ill’C** II ‘\*l\*l\*Cr1. The

Table 4. Model Comparison in for Case One

**ML Model Accuracy F1 Score Sensitivity Specificity**



**AdaBoost**

**Gradient Boosting Regressor**

Decision Tree Regressor

100% 100% 100%

100% 98.03% 100%

93.75% 94.84%92%

96.26%

93.33%

96.67%

Here, Logistic Regression shows the best outcomes and tree structure based algorithms show also constant outcomes. However, though dropping null values gives good prediction results, it is not counted as standard as more than 60% data are dropped.

In case number two, where the null values are replaced with maximum value of the data column or 1 which represents positive Boolean. The evaluation of the best fitted are given below :

Table S. Model Comparison in for Case Two

Testing

Trainning

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ML Model | | Accur  acy | F1  Scor  e | Sensiti  vity | Specifi  city | Accur  acy | F1  Scor  e | Sensiti  vity | Specifi  city |
| Logistic Regression | | 99.06 | 99.24 | 99% | 99% | 98.75 | 98.98 | 98% | 100% |
| AdaBoost |  | 96.75 | 98.83 | 100% | 100% | 92.23 | 99.17 | 100% | 97.29 |
|  |  | 0 0 | 0 0 |  | OO |  |  |  |  |
| SVR |  | 96.67 | 97.32 | 97.32 | 95.55 | 95% | 95.93 | 93.65 | 97.29 |
|  |  | OO | OO | 00 | OO |  | OO | OO | OO |
| Gradient | Boosting | 100% | 100% | 100% | 100% | 97.5% | 98.03 | 100% | 93.33 |
| Regressor |  |  |  |  |  |  |  | O O |  |
| XGBoost |  | 100% | 100% | 100% | 100% | 98.75 | 98.98 | 98% | 100% |
| Decision | Tree | 100% | 100% | 100% | 100% | 93.75 | 94.84 | 92% | 96.67 |

Regressor

OO OO

Where, boolean positive means filling all null values with positive case , for illustration, positive cases are “present”, “yes”, “abnormal” etc for this dataset. Based on assuming the the case, out of all algorithm, XGBoost model worked better both in training and testing part where the accuracy in training validation was 100% and in testing , in avaergae 98.98%.

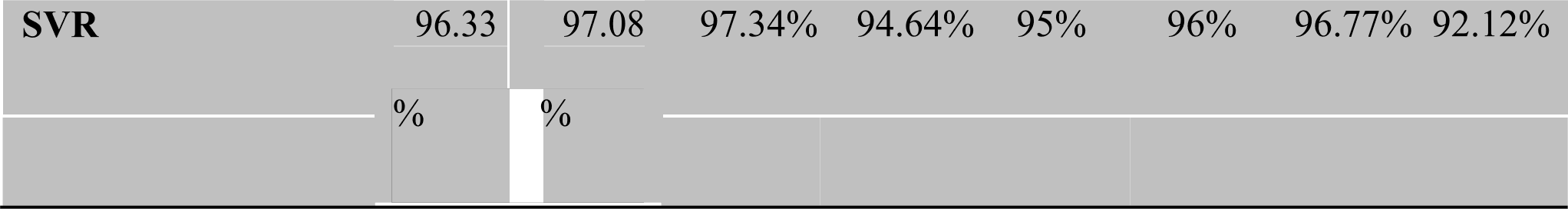
For case Three, where the null values are replaced with minimum value of the data column or 0 which represents negative boolean. The evaluation of the best fitted are given below (It works poor as only 150 patients are CKD Positive.)

Table 6. Model Comparison in for Case Three

Testing

Trainning

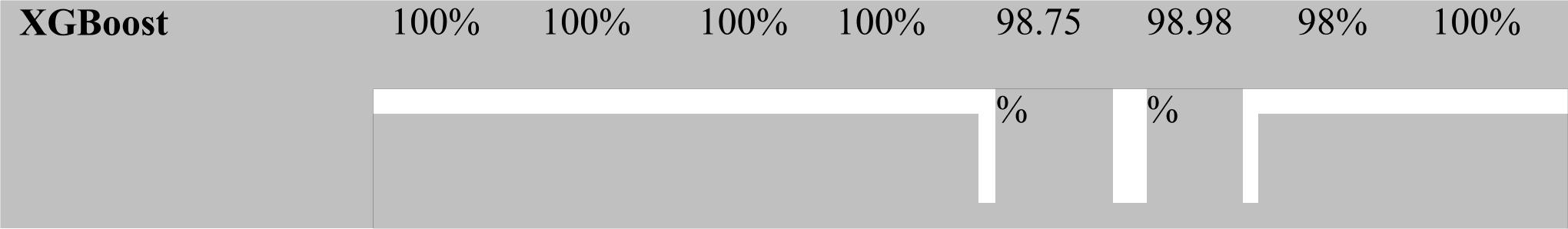
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ML Model** | **Accur acy** | **F1**  **Score** | **Sensiti vity** | **Specifi city** | **Accur acy** | **F1**  **Score** | **Sensiti vity** | **Specifi city** |
| **Logistic Regression** | 96.87 | 97.63 | 97.63% | 95.41% | 90% | 90% | 92% | 87.80% |
|  | OO | 00 |  |  |  |  |  |  |
| **AdaBoost** | 99.19 | 100% | 100% | 98.89% | 97.91 | 96% | 92.30% | 100% |



**Gradient Boosting** 100% 98.03 100%

**Regressor**

93.33% 97.5% 98.03 100%96.66%



**Decision** Tree 97.34 94.84 100% 92 125 93 75 94 85 92% 96 66%

**Regressor** 00 00

OO

OO OO

00

Here, boolean positive means filling all null values with positive case , for illustration, positive cases are “present”, “yes”, “abnormal” etc for this dataset. Based on assuming the case, out of all algorithm, XGBoost model worked better both in training and testing part where the accuracy in training validation was 100% and in testing , in average 98.98%.

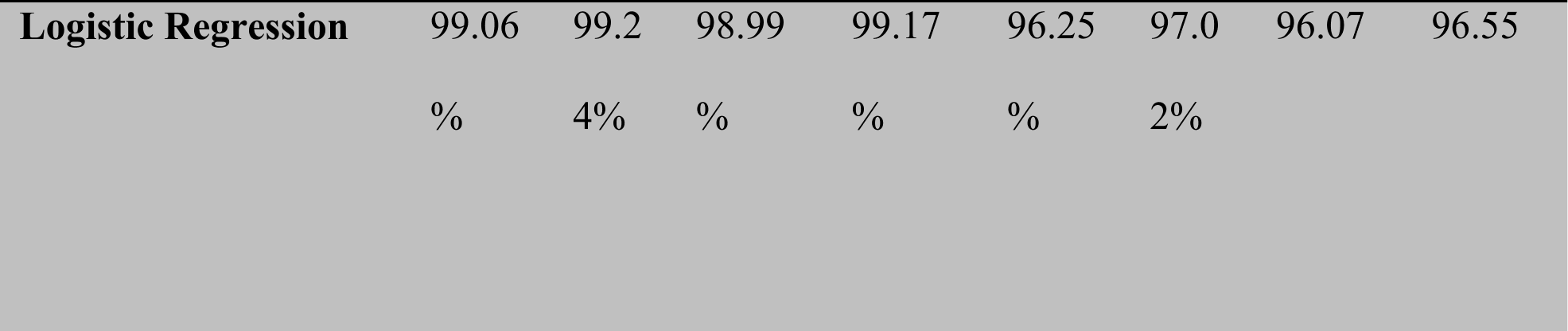
For case Four, Where the null values are replaced with mean value of the data column and selecting important features. In each column, it will check for all solid values and measure the mean value. Further, using the “fillna” function in pandas, null values are replaced with mean value. The evaluation of the best fitted are given below :

Table 7. Model Comparison in for Case Four

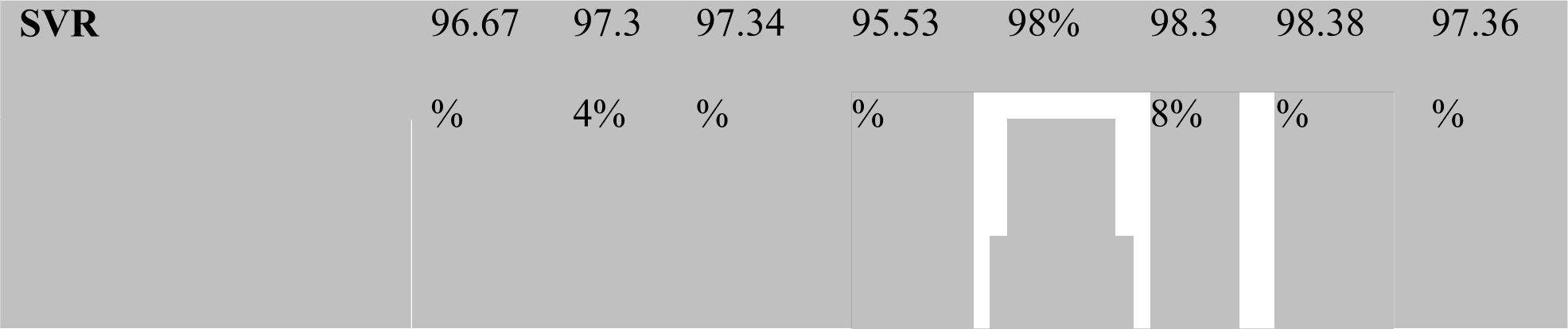
Testing

Trainning

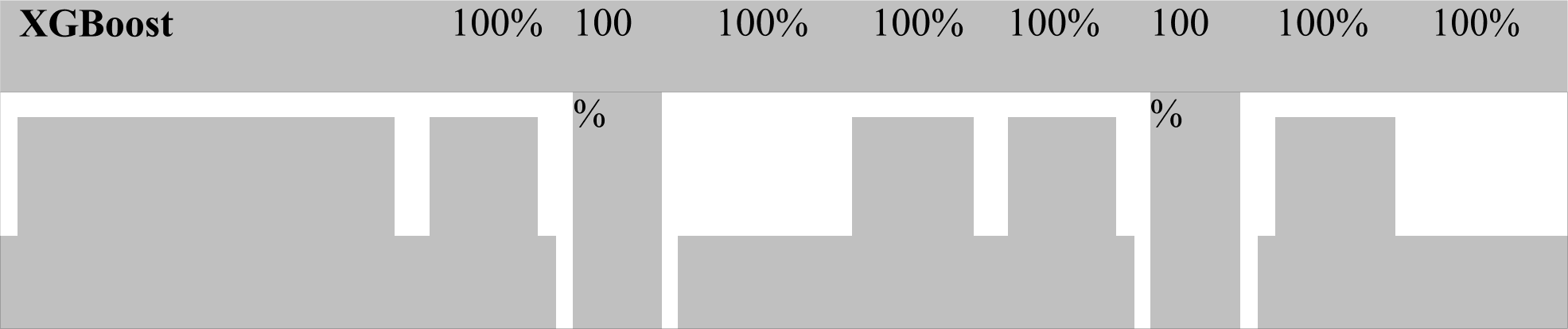
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ML Model | **Accur** | **FI** | **Sensiti** | **Specifi** | **Accur** | **FI** | **Sensiti** | **Specifi** |
|  | **acy** | **Scor** | **vity** | **city** | acy | **Scor** | **vity** | **city** |
|  |  | e |  |  |  | e |  |  |



AdaBoost 99.19 100 100% 98.89 97.91 96% 92.30 100%



Gradient Boosting 100% 100 100% 100% 98.75 98.9 98% 100%

Regressor

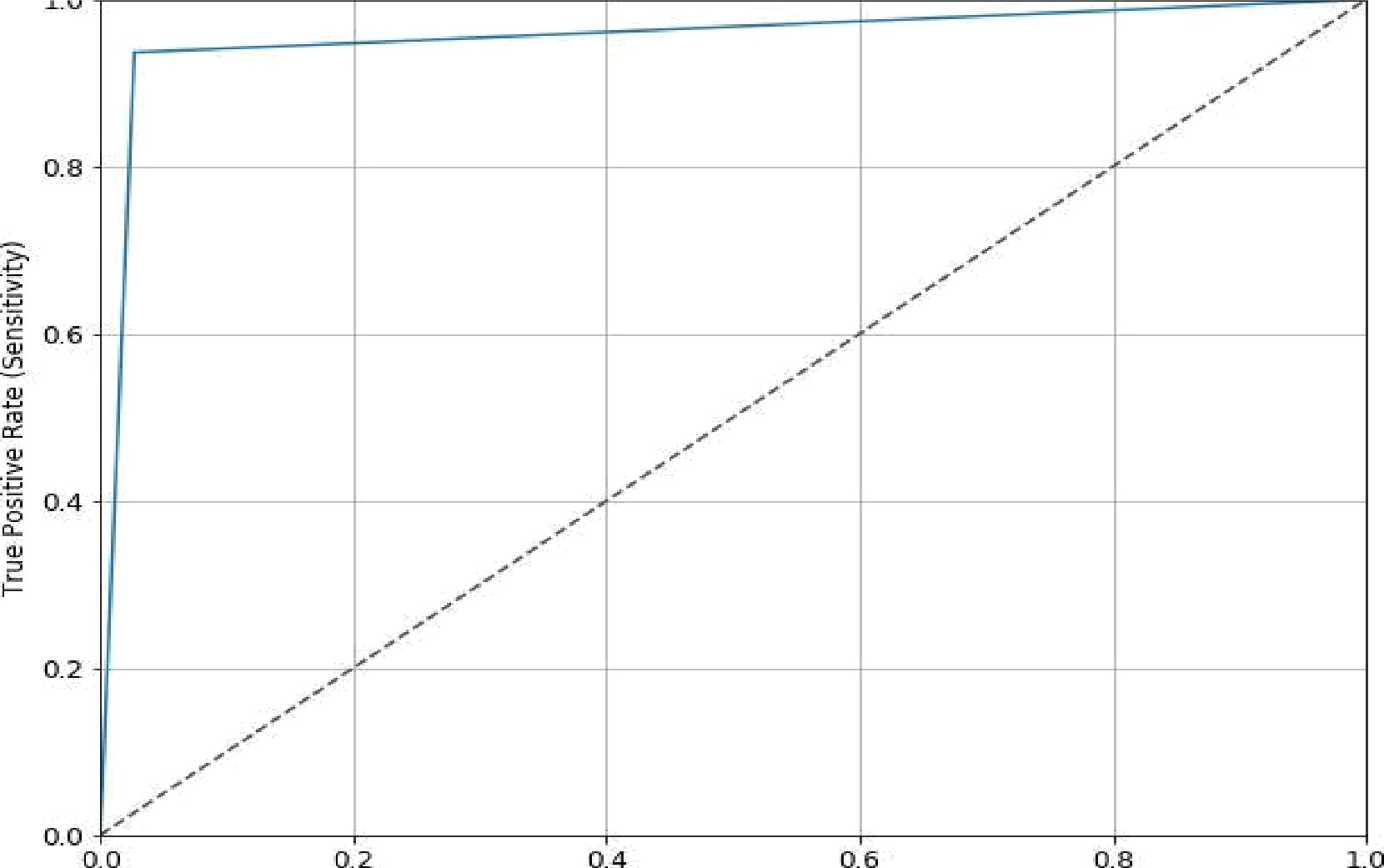
OO TOO

00

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Decision | Tree | 100% | 100 | 100% | 100% | 95% | 96.0 | 98% | 90% |
| Regressor |  |  |  |  |  |  | 7% |  | O O |

This case is counted as the most important and valuable case as rather than replacing with min or max value, mean value is more logical. Moreover, it is done with selecting top features using XGBoost, as a result, it shows better output.

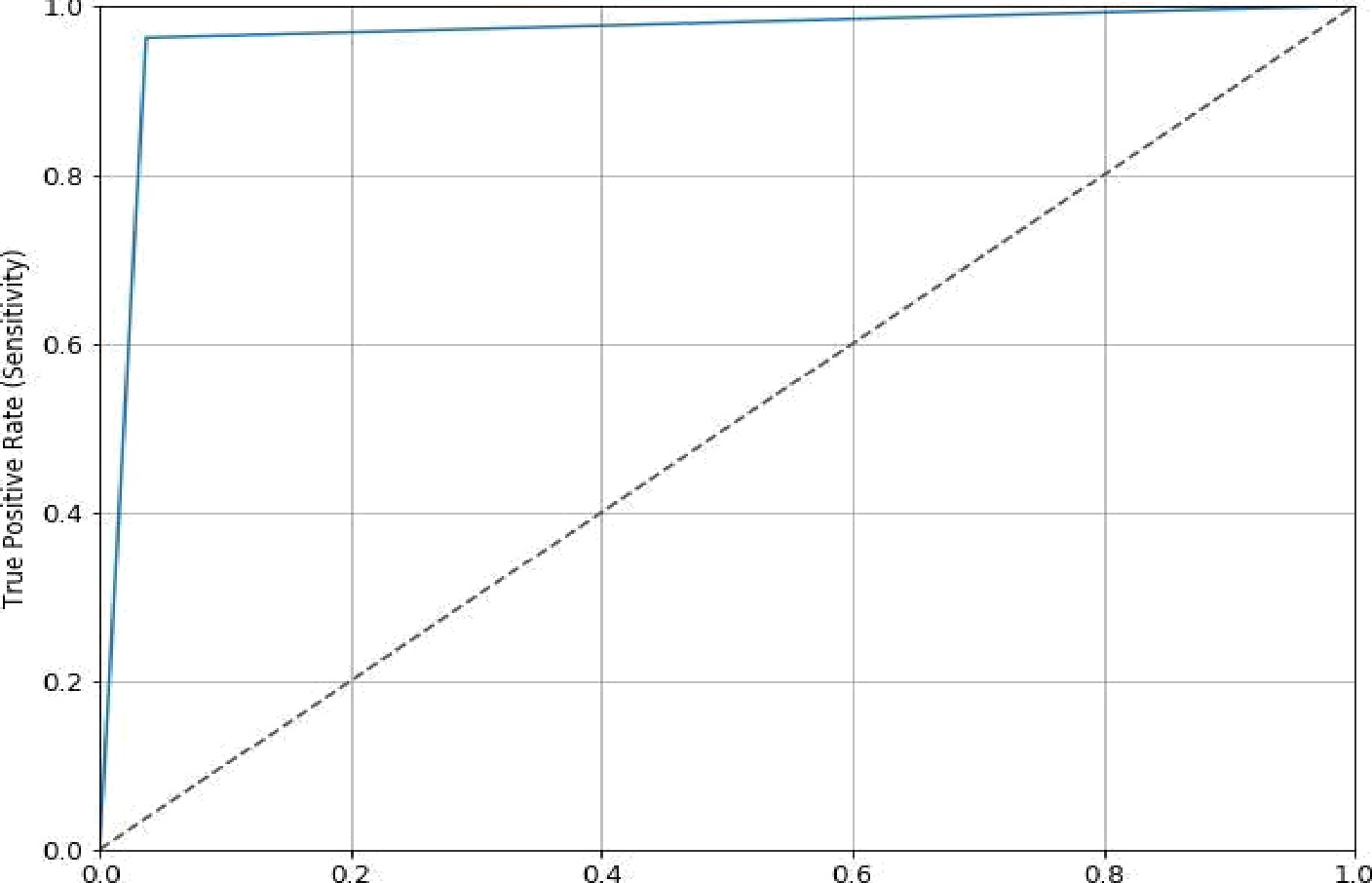
In particular, a measure of the usefullyness of the test region under an ROC curve where a larger area implies a more helpful test is used for the comparison of the usability of the trials with fields under ROC curves. Receiver Operating Characteristic (ROC) curve is the graph of True Positive Rate (TPR) against False Positive Rate (FPR). It shows the diagnostic ability of a model.



Fals e Pos i tive Rate

Figure 4. ROC curve of SVR

In figure 4, the value of true positive rate is 98.38% and false positive (l-TPR) is 1.62%, based on this, ROC curve is plotted which represents the greater accuracy of the test using SVR prediction model.



Fa Is e Positive Rate

Figure 5. ROC curve of Logistic Regression

In figure 5, the value of true positive rate is 96.07% and false positive (l-TPR) is 3.93%, based on this, ROC curve is plotted which represents the accuracy of the test using Logistic Regression.

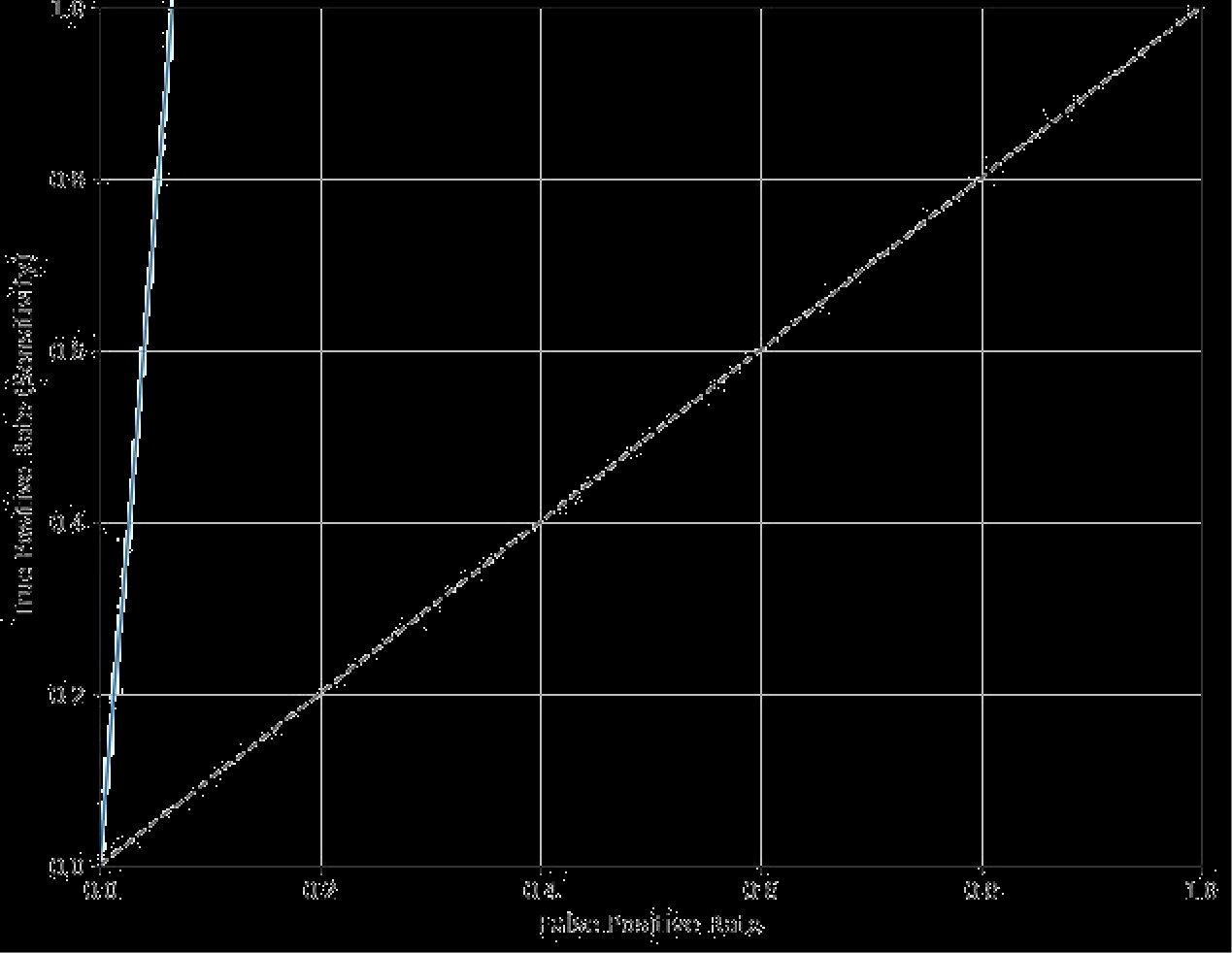


Figure 6. ROC curve of XGBoost

In figure 6, the value of true positive rate is 100% and false positive (1-TPR) is 0.00%, based on this, ROC curve is plotted which represents the bestr accuracy of the test using XGBoost prediction model.

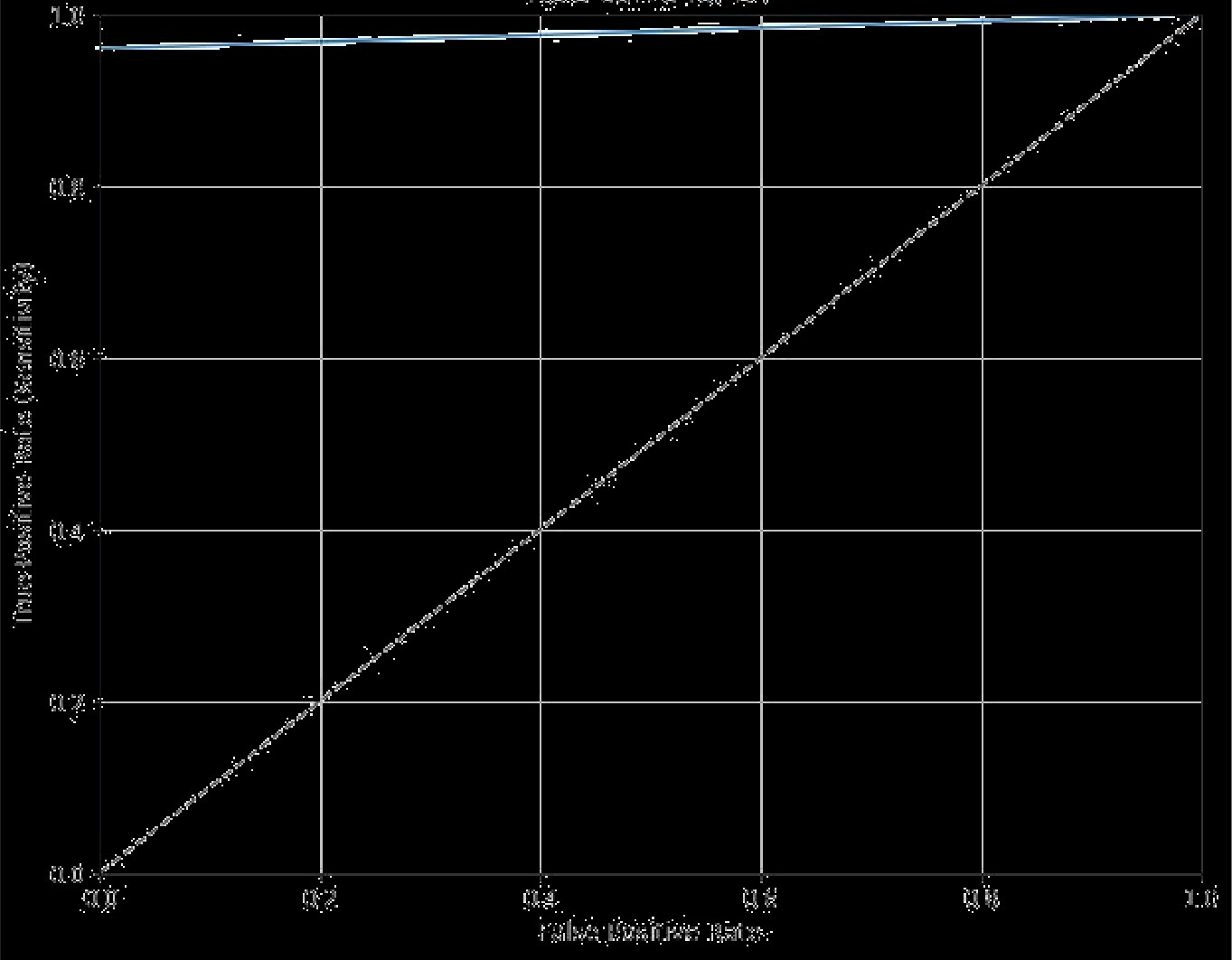


Figure 7. ROC curve of Gradient Boosting regressor

In figure 7, the value of true positive rate is 98% and false positive (I-TPR) is 2.00%, based on this, ROC curve is plotted which represents the bestr accuracy of the test using Gradient Boosting prediction model.

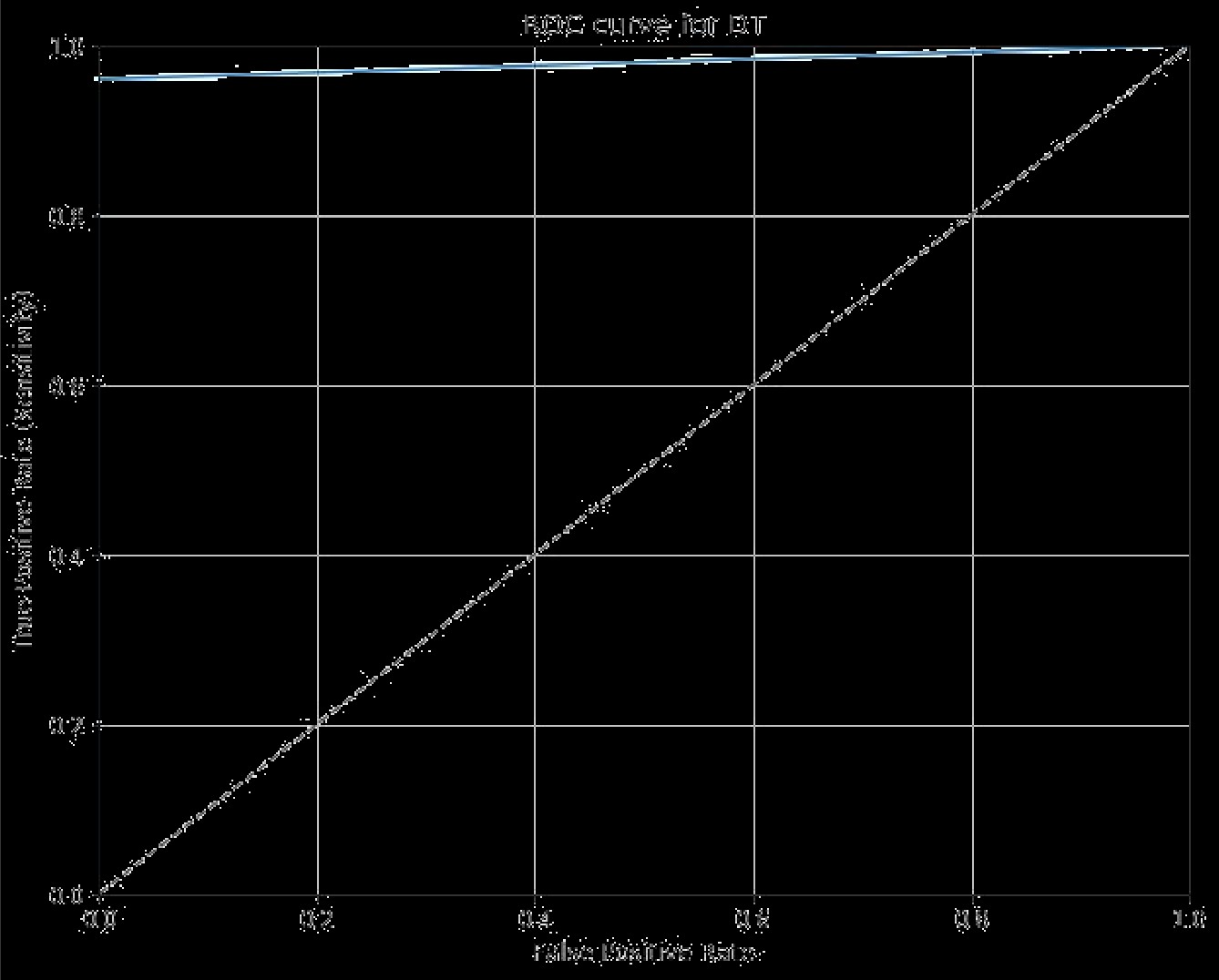


Figure 8. ROC curve of Decision Tree

In figure 8, the value of true positive rate is 98% and false positive (1-TPR) is 2.00%, based on this, ROC curve is plotted which represents the accuracy of the test using Decision Tree based prediction model.

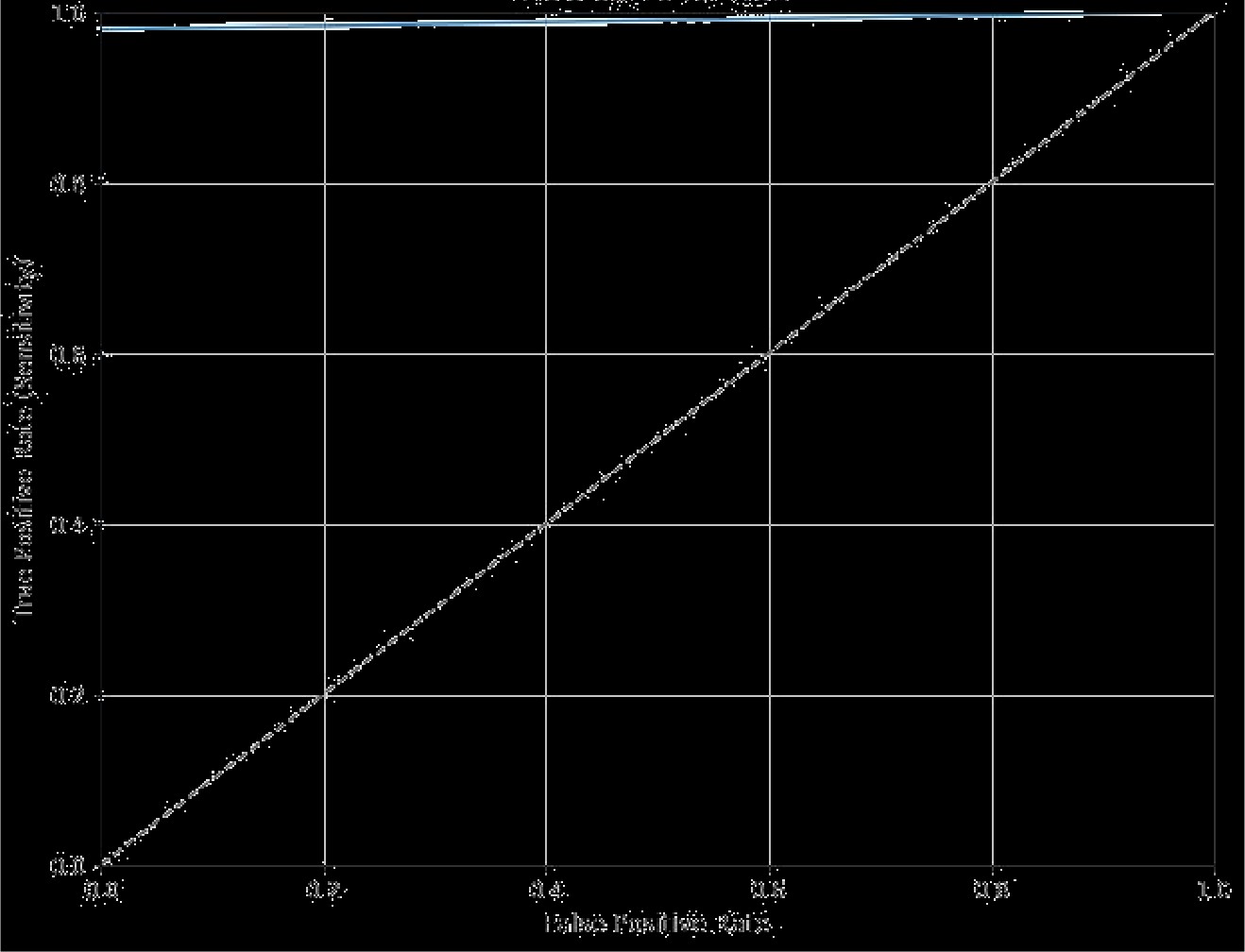


Figure 9. ROC curve of AdaBoost

In figure 9, the value of true positive rate is 98.38% and false positive (l-TPR) is 1.62%, based on this, ROC curve is plotted which represents the accuracy of the test using AdaBoost based prediction model.

Based on the correction and implementing XGBoost, the core parameters is ranked which is shown in figure 10. It denotes for further execution, atleast this 12 parameters must be included among 25 parameters.

i ..



Figure 10. Feature ranking applying XGBoost

Table 8 shows the comparsion along with three other papers featuring total case studied,

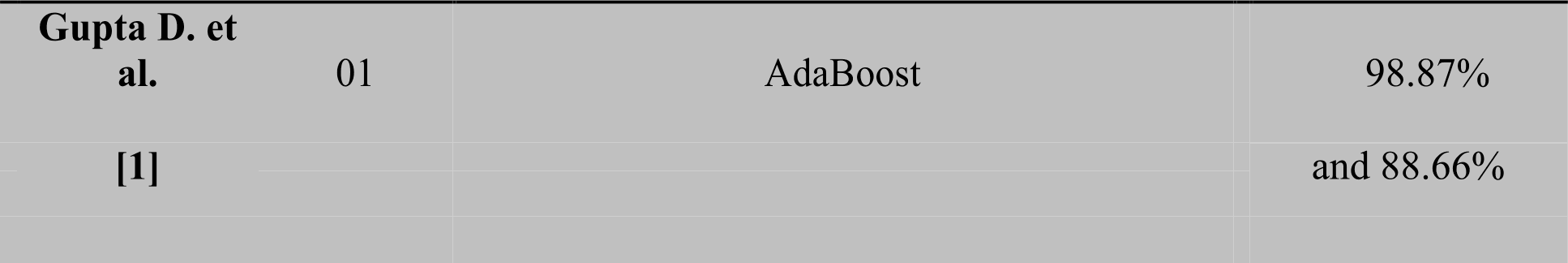
techniques, accuracy.

Table 8 : methodology Comparison

Author Total Techniques Accuracy

Case





A. Y. Al-

Hyari

et al. [2].

01 Artificial Neural Network and Naive Bayes 88.2% and

82.4%

respectively.



Methodology

XGBoost, Gradient Boosting, Decision Tree Regressor

Case)

**Chapter 6**

# Conclusion

Logistic regression is called for the operate used at the core of the tactic, the logistical operate. The logistical operate, additionally known as the sigmoid operate was developed by statisticians to explain properties of increase in ecology, risinp quickly and maxim out at the carrying capability of the setting. It's an formed curve which will take any real-valued range and reap it into a worth between zero and one, however ne'er precisely at those limits. one / (1 + e^-value) wherever e is that the base of the natural logaritluns (Euler's range or the EXP() operate in your spreadsheet) and worth is the actual numerical value that you simply wish to rework. Below may be a plot of the numbers between -5 and five reinodeled into the vary zero and one victimization the logistical operate. logistical regression uses associate degree equation because the illustration, noticeably like regression toward the mean. Input worths (x) are combined linearly victimization weights or constant values (referred to because the Greek capital letter Beta) to predict associate degree output value (y). A key distinction from regression toward the mean is that the output worth beinp sculptural may be a binary values (0 or 1) instead of a numeric value.

Moreover, above the three cases, the most logical case was exchanging null values with

mean values, but we got best results for dropping null values. The main challenge of the research work was to manage dataset as medical dataset is not publicly accessible in our country and dataset that we have used is not big enough. Although the limitation, we managed to solve four case studies and showed the best outputs in tree based structured machine learning algorithms.

# References

1. V. Jha , G. Garcia-Garcia , K. Iseki , Z. Li , S. Naicker, B. Plattner, R. Saran, A. Y. Wang,

C.W. Yang (20l3),”Chronic kidney disease: global dimension and perspectives”, The Lancet.

1. Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016). Chronic Kidney Disease analysis using data mining classification techniques. In 2016 6th International Conference- Cloud System and Big Data Engineering (Confluence) (pp. 300-305)
2. L. Xun, Wu Xiaoming, Li Ningshan and Lou Tanqi, "Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. V15-332-Vl5-335.
3. A. Salekin and J. Stankovic, (2016) “Detection of chronic kidney disease and selecting important predictive attributes,” IEEE International Conference on Healthcare Informatics.
4. D. Gupta, S. Khare, and A. Aggarwal, (20l6)“A method to predict diagnostic codes for chronic diseases using machine learning techniques,” 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 281-287.
5. A. Y. Al-Hyari, A. M. Al-Taee and M. A. Al-Taee, (2013) "Clinical decision support system for diagnosis and management of Chronic Renal Failure," 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, pp. 1-6.
6. Q. Zheng, G. Tasian, and Y. Fan, “Transfer learning for diagnosis of congenital abnormalities of the kidney and urinary tract in children based on Ultrasound imaging data,” International Symposium on Biomedical Imaging, vol. abs/1801.0, no. Isbi, pp. 1487-1490, 2018.
7. S. P. Deng, S. Cao, D.-S. Huang, and Y.-P. Wang, (2017) “Identifying Stages of Kidney Renal Cell Carcinoma by Combining Gene Expression and DNA Methylation Data,”

IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 5, pp. 1147-1153.

1. Vijayarani, S., Dhayanand, S., & Phil, M. (2015). Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR), 6(2).
2. Good, David M., Petra Zurbig, Angel Argiles, Hartwig W. Bauer, Georg Behrens, Joshua J. Coon, Mohammed Dakna et a1. "Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease." Molecular & cellular proteomics 9, no. 11 (2010): 2424-2437.
3. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.
4. Ogunleye, A. A., & Qing-Guo, W. (2019). XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM transactions on computational biology and bioinformatics.
5. Ogunleye, A., & Wang, Q. G. (2018, June). Enhanced XGBoost-Based Automatic Diagnosis System for Chronic Kidney Disease. In 2018 IEEE 14th International Conference on Control and Automation (ICCA) (pp. 805-810).
6. T. Chen and C. Guestrin, “XGBoost,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
7. R. A. Abbasi, N. Javaid, M. N. J. Ghuman, Z. A. Khan, Z. S. U.Rehman & Amanullah. (2019), “Short Term Load Forecasting Using XGBoost.” In: Barolli L., Takizawa M., Xhafa F., Enokido T. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing, vol 927. Springer, Cham
8. J. Guo„ L. Yang, R. Bie, J. Yu, Y. Gao, Y. Shen, and A. Kos, (2019) “An XGBoost- based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring,” Computer Networks, vol. 151, pp. 166-180.
9. X. Shi, Q. Li, Y. Qi, T. Huang and J. Li, (2017) "An accident prediction approach based on XGBoost," 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, pp. 1-7.
10. “Introduction tO Boosted Trees, [Online].

Available: https://xgboost.readthedocs.io/en/latest/tutorials/model.html.

[Accessed: 21-Dec-2018].

1. X. Gao (2017), "An improved XGBoost based on weighted column subsampling for object classification,"4th International Conference on Systems and Informatics, Hangzhou, pp. 1557-1562.
2. J. Yang (2006) “Power system short-term load”, PhD thesis, Technische Universit”.
3. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, (1984) “Classification and regression trees. belmont, ca: Wadsworth”, International Group, p. 432.
4. Belgiu, M., & Drñgu{, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.
5. Du, P., Samat, A., Waske, B., Liu, S., & Li, Z. (2015). Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. ISPRS Journal of Photogrammetry and Remote Sensing, 105, 38-53.
6. T. Moller, R. Machiraju, K. Mueller and R. Yagel, (1997) "Evaluation and design of filters using a taylor series expansion," IEEE Transactions on Visualization and Computer Graphics, vol. 3, no. 2, pp. 184-199,
7. “Boosting algorithm: XGBoost,” Towards Data Science. [Online]. Available: https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d. [Accessed: 22-Jul- 2018].
8. Boosting and AdaBoost for Machine Learning [Online]. Available: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/ [Accessed: 13-Aug-2019].
9. Morde, V. (2019). “Feature Importance and Feature Selection With XGBoost in Python” [Online]. Available: https://towardsdatascience.com/https-medium-com- visha1morde-xgboost-algorithm-long-she-may-rein-edd9f99be63d
10. Chronic Kidney Disease Data Set [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic kidney disease