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

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)

_____

Dr. Md Ashraful Alam
Assistant Professor
CSE, BRAC University

Departmental Head:
(Chair)

_____

Professor Mahbubul Alam Majumdar
Chair person,
CSE, BRAC University

Ethics Statement

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

List of Tables

List of Figures

Chapter 1

Introduction

1.1 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 excretory organ Foundation, over ten million individuals are currently full of nephropathy within the country. Nearly 160,000 excretory organ patients, UN agency are in serious condition, have to be compelled to endure regular qualitative analysis weekly. Concerning 195 million girls are laid low with CKD within the world and it's presently the eighth leading reason behind death among women with around 600,000 individuals' dies of this illness annually. Between 2011 and 2012, there have been concerning one.9 million adults with CKD in England as registered within the Quality Outcomes Framework (QOF). However, the general annual price of CKD to the united kingdom National Health Service (NHS) is calculable

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 summary and to focus on some strategies that might 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 future results section, we have a tendency to initial describe the main CKD outcomes that are investigated in hand-picked papers, so describe and discuss the regression strategies used for every variety of outcomes. wherever applicable, we have a tendency to gift some potential different analytical approaches that are ne'er or seldom employed in the context of CKD however may nonetheless be of interest.

1.2 Our Contribution

So the main contributions of this paper are- After addressing feature reduction, learning algorithm selection and tuning and comparing several ML techniques ok CKD dataset we found XGBoost model to perform best. So it is then used as base model. XGBoost model was used to rank the most important features. After that we applied different machine learning approach like Logistic regression, Gradient Boosting, Decision Tree, AdaBoost, Support Vector 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 missing 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 attributes 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 al. [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 al. [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 al. [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 death and disability. In 1990, CKD was the 27th leading cause of death, which rose to the 18th leading cause of death in 2010. Approximately 1 million people died in 2013 due to CKD related cause. Despite being a worldwide problem, CKD has a disproportionate impact on individuals in developing nations. A systematic review conducted in 2015 reported that 109.9 million people in high-income countries had CKD (men-48.3 million, women-61.7 million) while the burden was 387.5 million in lower middle-income countries (men-177.4 million, women-210.1 million).CKD is correlated with a broad spectrum of life-threatening illnesses. It is regarded to be one of the main risk variables for developing cardiovascular disease. In a 2003 research, patients with Glomerular Filtration Rate (GFR) between 15 and 59 ml / min/1.73 m2 were at a 38% greater danger of developing cardiac illness than patients with GFR 90 and 150 ml / min/1.73 m2.In addition to the effect on individual health, CKD also impacts social life and is liable for the loss of productivity. Financial burdens are the most prevalent type of social effect owing to CKD.CKD patients are at greater danger of developing end-stage renal disease (ESRD) which needs expensive management, 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 2001 was close to USD 75 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.

3.1 Search strategy Chronic Kidney Diseases (CKD Statistics of CKD

Here, a systematic review of the appropriate current literature from South Asian nations was carried out using the PRISMA 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 domestic online journal accessible for Bhutan, the Maldives and Afghanistan. During the search, both medical 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 main terms alongside Boolean operators, a worldwide search term for prospective literature searches has been created. The bibliography among all chosen research (snow bowling) was also searched manually to distinguish further papers.

3.2 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 / min/1.73 m2.According to

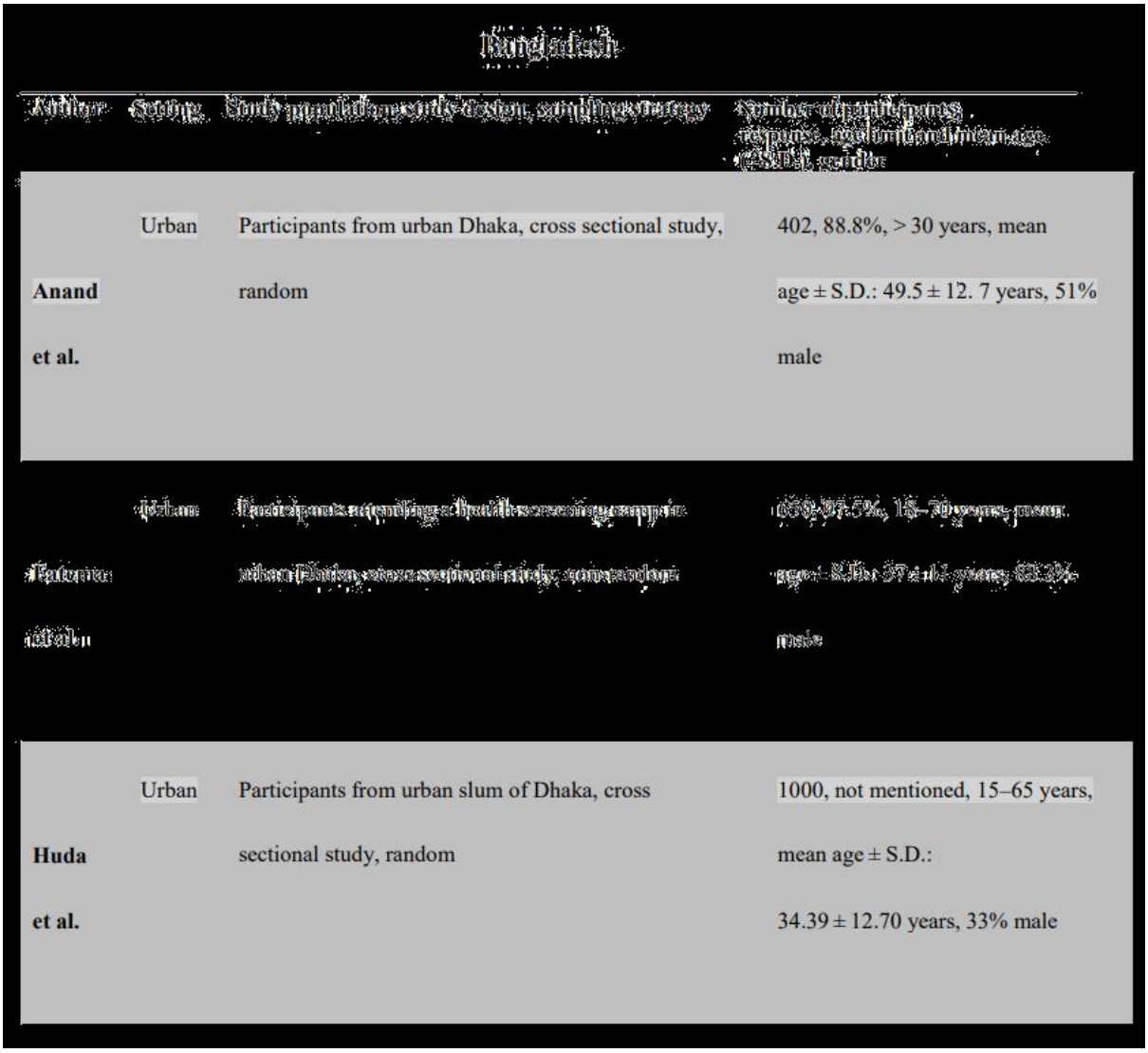
The research for this analysis, three equations were used to predict eGFR: four-variable MDRD equation CKD-EPI equation and Cockcroft-Gault equation. Two authors (MH and RDG) obtained information individually from the chosen papers and a data extraction table was created for this purpose in an excel file. This table included (a) title, (b) journal name, (c) name of authors, (d) publication year, (e) year of data collection, (f) study objective, (g) study setting (urban/rural), (h) study design, (i) sampling strategy (random/non-random), (j) sample size, (k) study population, (l) outcome assessment (objective/subjective), diagnostic criteria for CKD, (n) prevalence (overall), (o) prevalence (gender, age, location specific), and (p) authors" conclusion. After data extraction, a third author (IS) crosschecked both of the tables to ensure consistency. Any contention that emerged throughout data extraction was tackled by a consensus of the group. Subsequently, the data was evaluated using tabulation, grouping and thematic approach.

3.3 Statistics of CKD

Eight trials have been identified from India, all of which have embraced cross-sectional study design. Most of the research was performed solely in urban environments, 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 techniques. The number of participants in this research ranged from 1104 to 12,271, most of whom were male adults. The trials evaluated spot quantitative urine protein and/or eGFR as biomarkers 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.

Characteristics of the Indian studies

Author	Setting	Study population, study design, sampling strategy	Number of participants, response, age limit and mean age (±SD), gender
Anand et al.	Urban	Participants from Delhi and Chennai who took part in Center for Cardio metabolic Risk Reduction in South Asia surveillance, cross sectional, random	12,271, 80%, > 20 years, mean age ± S.D.: 41.4 ± 12.7 years (Chennai), 44.4 ± 13.9 years (Delhi), 43.5% male
Ajmera et al.	Rural	Participants from rural Karnataka who took part in a screening survey, cross sectional study, random	2723, 70.6%, ≥18 years, mean age ± S.D.: 39.83 ± 15.87 years, 45.6% male
Mahapatra et al.	Urban	Indian central government Employees in Delhi, cross sectional study, non-random	1104, Not mentioned, > 18 years, Not Mentioned, 61.4% male
Singhal et al.	Urban, Semi Urban, Rural	Participants from Delhi and adjoining region, cross sectional, random	5583, 94.4%, ≥ 20 years, Not Mentioned, 61% male
Singh et al.	Urban	Participants attending thirteen academic and private medical centers in India participated in the study under the name of Screening and Early Evaluation of Kidney disease – SEEK, cross sectional study, non-random	6120, 92.3%, ≥18 years, mean age ± S.D.: 45.22 ± 15.2 years, 55.1% male
Talwalkar et al.	Semi Urban	Participants attending a health screening camp in environs of Delhi, cross sectional study, non-random	2350, Not Mentioned, mean age ± S.D.: 45.16 ± 15 years, 61.2% male
Varma et al.	Urban	Indian central government employees in Agra town, cross sectional study, Not Mentioned	3398, 83.9% ≥18 years, mean age ± S.D.: 35.64 ± 8.72 years, 66% male
Verma et al.	Urban	Indian Army personnel in Agra town who were part of a Comprehensive Health Survey & Evaluation of Life Style Diseases at the level of Army Hospital, cross sectional study, Not Mentioned	1520, 91.9%, > 20 years, mean age ± SD: 34.82 ± 7.57 years, gender distribution Not Mentioned



Bangladesh

Author	Setting	Study population, study design, sampling strategy	Number of participants, response, age, gender
Anand et al.	Urban	Participants from urban Dhaka, cross sectional study, random	402, 88.8%, > 30 years, mean age ± S.D.: 49.5 ± 12. 7 years, 51% male
Fatema Huda	Urban	Participants attending a health screening camp in urban Dhaka, cross sectional study, convenient	650, 97.5%, 18–70 years, mean age ± S.D: 37 ± 13 years, 62.2% male
Huda et al.	Urban	Participants from urban slum of Dhaka, cross sectional study, random	1000, not mentioned, 15–65 years, mean age ± S.D.: 34.39 ± 12.70 years, 33% male

3.4 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

word for heterogeneous illnesses influencing the structure and function of the kidney. It generally follows a progressive path and is hardly reversible. There is a need to identify risk factors for CKD development in order to allow for future therapeutic measures. In specific, progression to renal failure, i.e. a glomerular filtration rate (GFR) of less than 15 mL / min/1.73 m2 or a need for dialysis or transplantation therapy, must be avoided due to enhanced mortality and therapy expenses. Different kinds of result variables can be used to investigate risk factors connected with CKD development in statistical analysis. The result variable, for instance, may be the time for advancement to a particular GFR value, initiation of dialysis or transplantation, cardiovascular events, or death from all causes. The result variable may also be the slope of GFR decrease, or over time its general trajectory. If the result variable is the time of a specific event of interest, then a model of survival regression such as the Cox model may seem obvious to investigate risk factors associated with this event. For all patients experiencing the event, standard survival analysis requires the time-to-event to be known exactly. This is the case, for instance, for occurrences such as death or kidney replacement treatment initiation, where the precise dates can generally be found. However, for many other occurrences of concern in development of kidney disease, the time-to-event may not be precisely known. For instance, it is known that the progression to a particular GFR value happened only between two successive GFR readings. In such a situation, the time between the times of these two measurements 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 time, a linear regression model may be used to explore the connection of risk variables with summary statistics (such as hills) of the individual GFR evolution over time. Such methods, however, often fail to account for all accessible data on repeated kidney function measurements and impose certain assumptions that may not be true. A linear mixed model using all repeated quantitative kidney function measurements may be preferable if a adequate amount of patients with at least three measurements are present. This regression model represents the correlation between repeated measurements of the same patient and handles distinct measurement figures per patient that can be measured at unequally distant intervals, as well as non-linear trajectories over time. Therefore, each sort of CKD result variable raises a number of methodological problems in the investigation of risk variables in statistical analysis. Some of these statistical problems have been recognized in the literature on nephrology, but no document provides an overview of these statistical problems to our understanding. This article therefore tries to provide this summary and highlight some techniques that might solve these problems and have been suggested in the context of CKD or in some other situations. To this end, we first performed a literature review of the statistical methods used over the past ten years to explore CKD results risk factors. Second, the outcomes of this review of the literature has been used to define significant methodological problems and to highlight some techniques that solve these problems.

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.

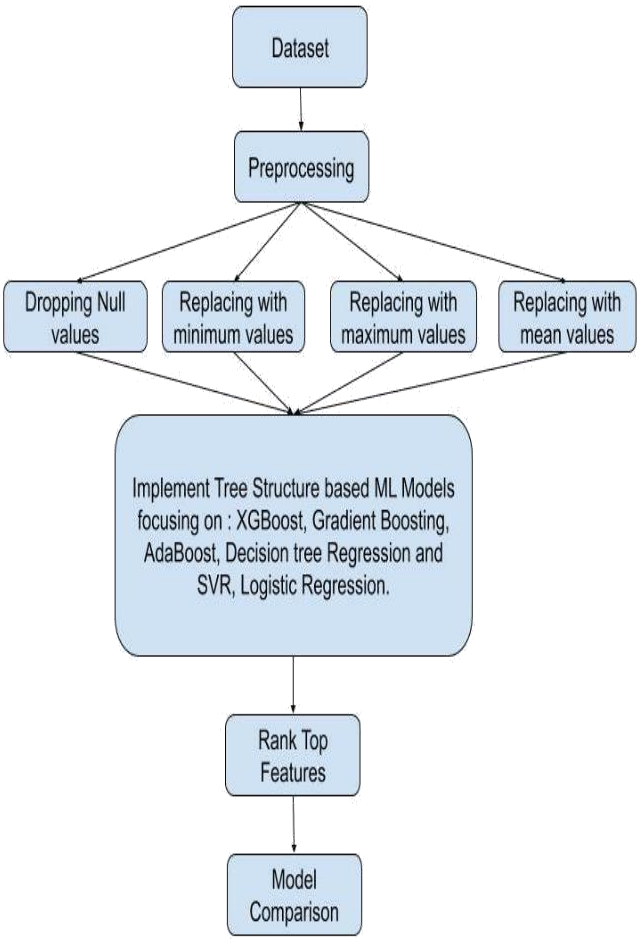


Figure 1. Work diagram

4.1 Dataset

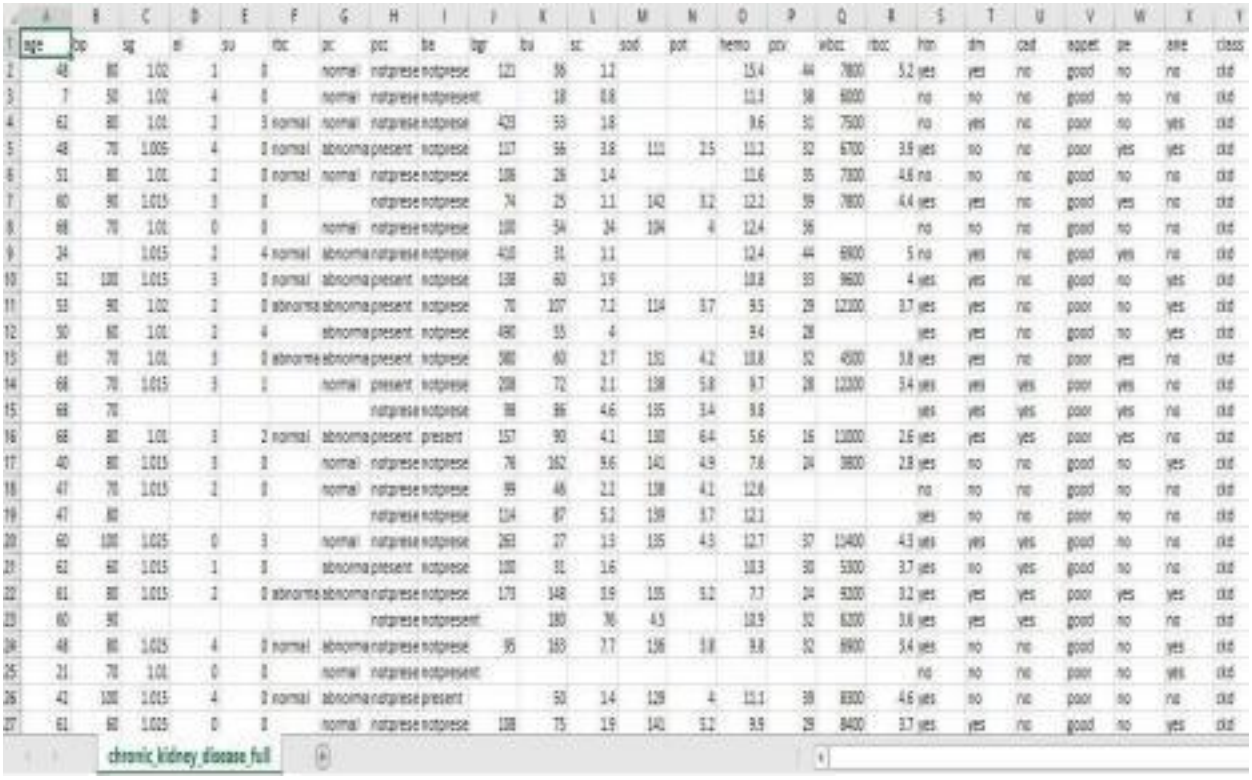
In our thesis, the CKD dataset was obtained from the open source research library of the University of California Irvine (UCI) [28] which had developed by Dr. P. Soundarapandian. The data set includes 25 characteristics of total 400 patients where 250 patients are CKD positive and 150 patients are normal. 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/cumm, 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

Number	Attribute	Abbreviation
1	Age	Age
2	Appetite	APPET
3	Anemia	ANE
4	Albumin	AL
5	Bacteria	BA
6	Blood Pressure	BP
7	Blood Glucose Random	BGR
8	Blood Urea	BU
9	Coronary Artery Disease	CAD
10	Diabetes Mellitus	DM
11	Pus Cell Clump	PCC
12	Serum Creatinine	SC
13	Sodium	SOD
14	Potassium	POT
15	Hemoglobin	HEMO
16	Pack Cell Volume	PCV
17	White Blood Cell Count	WC
18	Red Blood Cell Count	RBCC
19	Hypertension	HTN
20	Pus Cell	PC
21	Specific Gravity	SG
22	Red Blood Cell	RBC
23	Petal Edema	PE
24	Sugar	SU
25	Class	CLASS

Table 2. Dataset for 400 patients



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wbcc	rbcc	htn	dm	cad	appet	pe	ane	class
2	48	80	1.02	1	0		normal	notpresent	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
3	7	50	1.02	4	0		normal	notpresent	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
4	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
5	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
6	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
7	60	90	1.015	3	0			notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
8	68	70	1.01	0	0		normal	notpresent	notpresent	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
9	24		1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
10	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
11	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12000	3.7	yes	yes	no	poor	no	yes	ckd
12	50	60	1.01	2	4		abnormal	present	notpresent	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
13	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
14	68	70	1.015	3	1		normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
15	68	70						notpresent	notpresent	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
16	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
17	40	80	1.015	3	0		normal	notpresent	notpresent	76	162	9.6	141	4.9	7.8	24	9800	2.8	yes	no	no	good	no	yes	ckd
18	47	70	1.015	2	0		normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
19	47	80						notpresent	notpresent	114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
20	60	100	1.025	0	3		normal	notpresent	notpresent	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
21	62	60	1.015	1	0		abnormal	present	notpresent	100	31	1.6			10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
22	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
23	60	90						notpresent	notpresent		180	76	4.5		10.9	32	6200	3.8	yes	yes	yes	good	no	no	ckd
24	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd
25	21	70	1.01	0	0		normal	notpresent	notpresent										no	no	no	poor	no	yes	ckd
26	42	100	1.015	4	0	normal	abnormal	notpresent	present		50	1.4	129	4	11.1	38	8300	4.6	yes	no	no	poor	no	no	ckd
27	61	60	1.025	0	0		normal	notpresent	notpresent	108	75	1.9	141	3.2	9.5	29	8400	3.7	yes	yes	no	good	no	yes	ckd

chronic_kidney_disease_full

4.2 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:

```
data['class'] = data['class'].map({'ckd':1,'notckd':0})
data['htn'] = data['htn'].map({'yes':1,'no':0})
data['dm'] = data['dm'].map({'yes':1,'no':0})
data['cad'] = data['cad'].map({'yes':1,'no':0})
data['appet'] = data['appet'].map({'good':1,'poor':0})
data['ane'] = data['ane'].map({'yes':1,'no':0})
data['pe'] = data['pe'].map({'yes':1,'no':0})
data['ba'] = data['ba'].map({'present':1,'notpresent':0})
data['pcc'] = data['pcc'].map({'present':1,'notpresent':0})
data['pc'] = data['pc'].map({'abnormal':1,'normal':0})
data['rbc'] = data['rbc'].map({'abnormal':1,'normal':0})
```

Figure 2. Numerical value conversation process

Table 3. Missing value representation

	Missing Values	% of Total Values
rbc	152	38.00
rc	131	32.75
wc	106	26.50
pot	88	22.00
sod	87	21.75
pcv	71	17.75
pc	65	16.25
hemo	52	13.00
su	49	12.25
sg	47	11.75
al	46	11.50
bgr	44	11.00
bu	19	4.75
sc	17	4.25
bp	12	3.00
age	9	2.25
ba	4	1.00
pcc	4	1.00
htn	2	0.50
dm	2	0.50
cad	2	0.50
appet	1	0.25
pe	1	0.25
ane	1	0.25

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.

4.3 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, y_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=0, max_depth=3, min_child_weight=1, missing=None,
       n_estimators=100, n_jobs=1, nthread=None,
       objective='binary:logistic', random_state=0, reg_alpha=0,
       reg_lambda=1, scale_pos_weight=1, seed=None, silent=None,
       subsample=1, verbosity=1)
```

S1: 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 gamma 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 gamma 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 gamma, build a grid-search over the scoop depth and min kid weight in elite ranges of one to ten.

S4: build a grid-search over the L2 regularization parameter reg lambda and subsample at the same time in elite ranges zero.1 to 1.

S5: 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 gamma, reg lambda and subsample 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 ( )".

4.3.1 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.

4.3.2 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.

4.3.3 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.

4.3.4 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, rising quickly and maxing out at the carrying capability of the setting. It's an formed curve which will take any real-valued range and map 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 logarithms (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 remodeled 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:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

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

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (2)$$

• Sensitivity denotes ratio of the properly predicted amount of CKD instances to the complete CKD instances and is calculated as:

$$Sentivity = \frac{TP}{TP+FN} \quad (3)$$

• Specificity is the ratio of the number of correctly predicted healthy cases over the total number of the healthy cases and calculated as :

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

Chapter 5

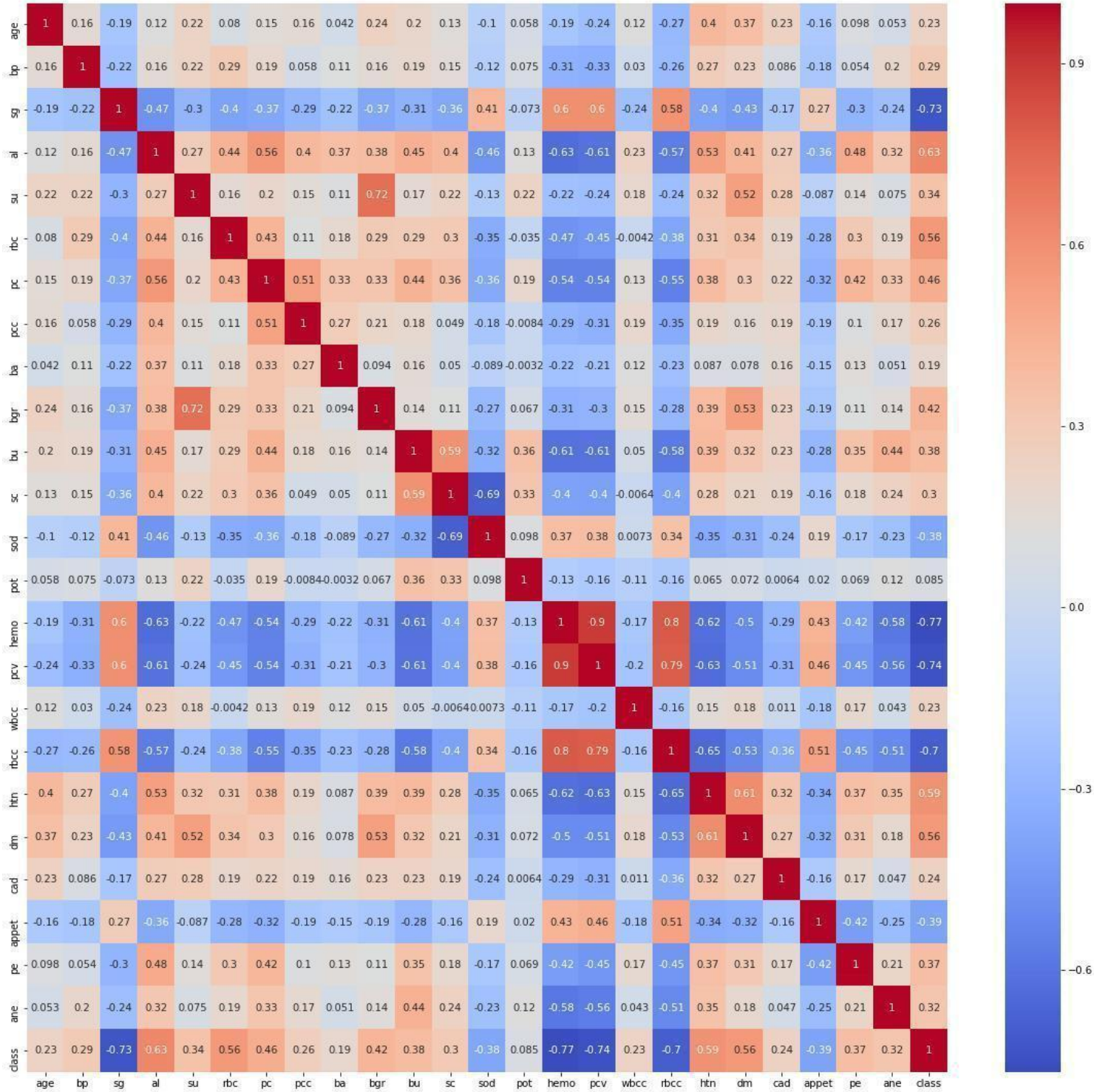
Result and Analysis

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.



process is shown in Figures 2 and 3. The optimal values of hyper parameters were obtained and are given in Table 4. The validation dataset was used to check the performance of the trained model. If it is satisfactory, we proceed to test. The test set was used to evaluate the model. Performance of this model is compared in following figures with the CKD models found in the literature for case one. In case one, where the null values are dropped. The evaluation of the best fitted are given below:

Table 4. Model Comparison in for Case One

ML Model	Accuracy	F1 Score	Sensitivity	Specificity
Logistic Regression	100%	100%	100%	100%
AdaBoost	100%	100%	100%	96.26%
SVR	95%	96%	96.77%	92.12%
Gradient Boosting Regressor	100%	98.03%	100%	93.33%
XGBoost	98.75%	98%	98%	100%
Decision Tree Regressor	93.75%	94.84%92%		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 5. Model Comparison in for Case Two

	Trainning				Testing			
ML Model	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**
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 %
SVR	96.67 %	97.32 %	97.32 %	95.55 %	95%	95.93 %	93.65 %	97.29 %
Gradient Boosting Regressor	100%	100%	100%	100%	97.5%	98.03 %	100%	93.33 %
XGBoost	100%	100%	100%	100%	98.75 %	98.98 %	98%	100%
Decision Tree Regressor	100%	100%	100%	100%	93.75 %	94.84 %	92%	96.67 %

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

	Trainning				Testing			
ML Model	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**
Logistic Regression	96.87 %	97.63 %	97.63%	95.41%	90%	90%	92%	87.80%
AdaBoost	99.19 %	100%	100%	98.89%	97.91 %	96%	92.30%	100%
SVR	96.33 %	97.08 %	97.34%	94.64%	95%	96%	96.77%	92.12%

Gradient Boosting Regressor	100%	98.03 %	100%		93.33%	97.5%	98.03 %	100%96.66%
XGBoost	100%	100%	100%	100%	98.75 %	98.98 %	98%	100%
Decision Tree Regressor	97.34 %	94.84 %	100%	92.125 %	93.75 %	94.85 %	92%	96.66%

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

	Trainning				Testing			
ML Model	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**	**Accuracy**	**F1 Score**	**Sensitivity**	**Specificity**

Logistic Regression	99.06 %	99.24%	98.99 %	99.17 %	96.25 %	97.02%	96.07	96.55
AdaBoost	99.19 %	100 %	100%	98.89 %	97.91 %	96%	92.30 %	100%
SVR	96.67 %	97.34%	97.34 %	95.53 %	98%	98.38%	98.38 %	97.36 %
Gradient Boosting Regressor	100%	100 %	100%	100%	98.75 %	98.98%	98%	100%
XGBoost	100%	100 %	100%	100%	100%	100 %	100%	100%
Decision Tree Regressor	100%	100 %	100%	100%	95%	96.07%	98%	90%

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.

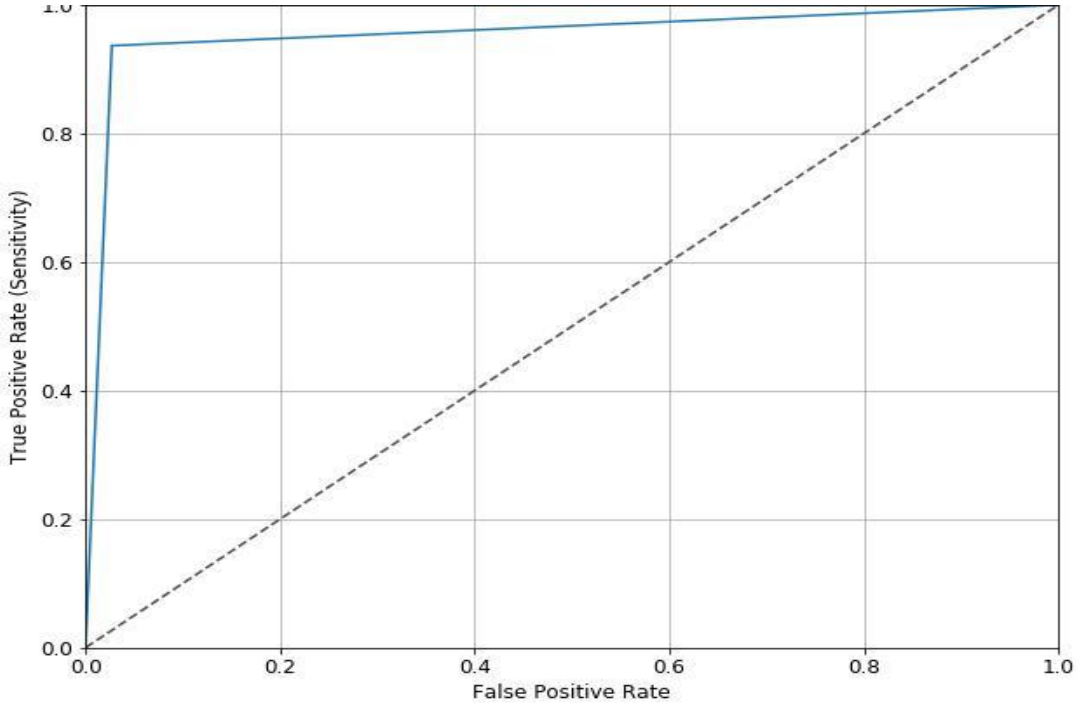


Figure 4. ROC curve of SVR

In figure 4, the value of true positive rate is 98.38% and false positive (1-TPR) is 1.62%, based on this, ROC curve is plotted which represents the greater accuracy of the test using SVR prediction model.

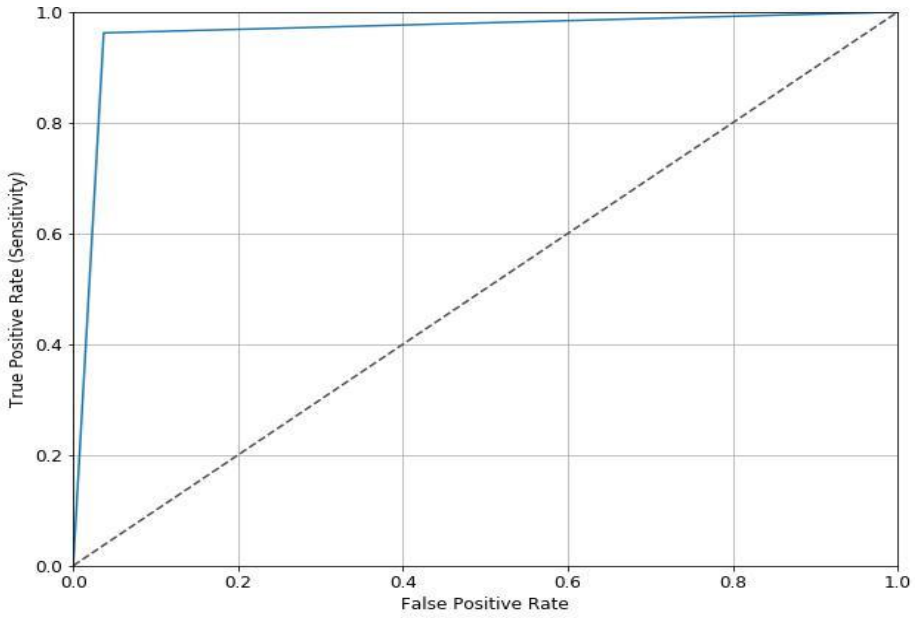


Figure 5. ROC curve of Logistic Regression

In figure 5, the value of true positive rate is 96.07% and false positive (1-TPR) is 3.93%, based on this, ROC curve is plotted which represents the accuracy of the test using Logistic Regression.

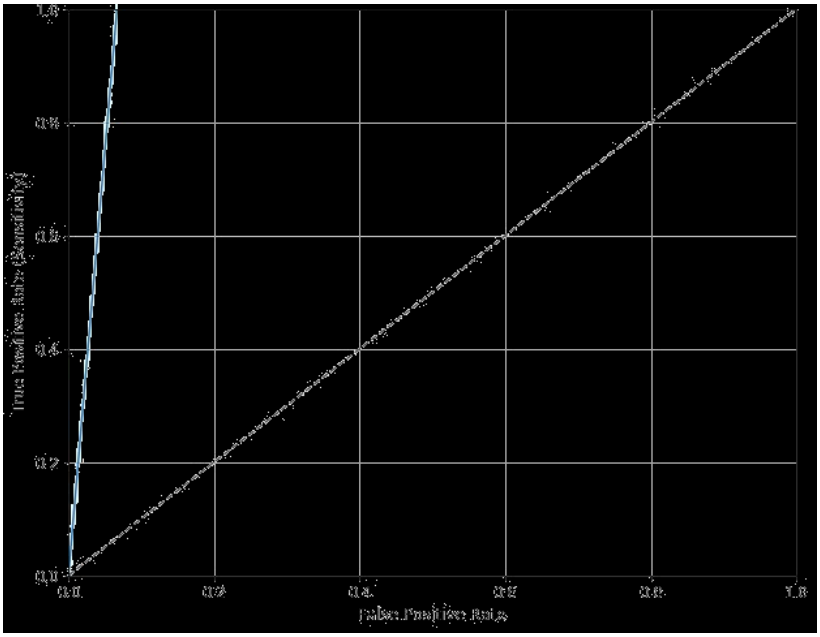


Figure 6. ROC curve of XGBoost

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.

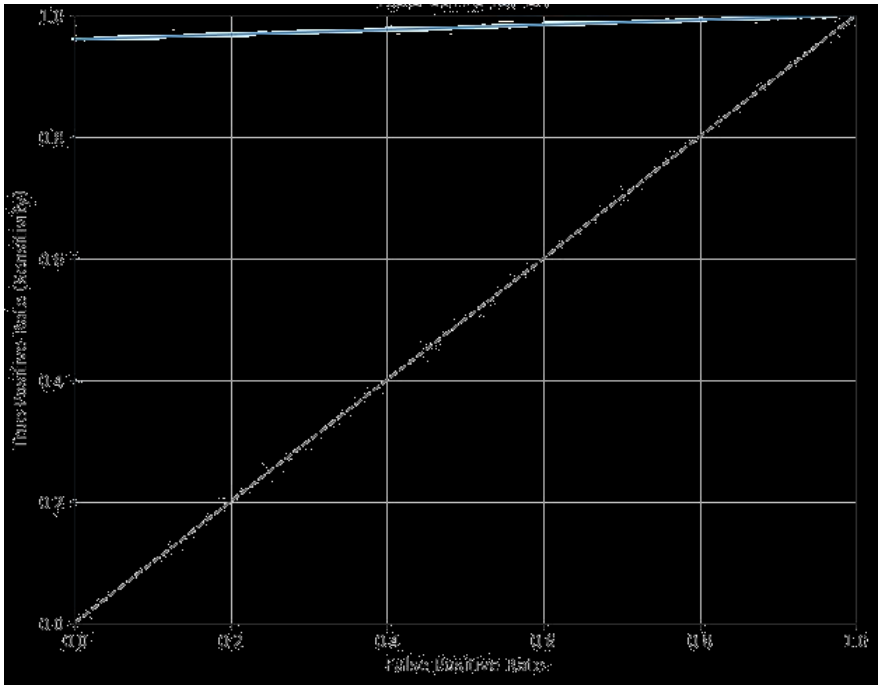


Figure 7. ROC curve of Gradient Boosting regressor

In figure 7, the value of true positive rate is 98% and false positive (1-TPR) is 2.00%, based on this, ROC curve is plotted which represents the bestr accuracy of the test using Gradient Boosting prediction model.

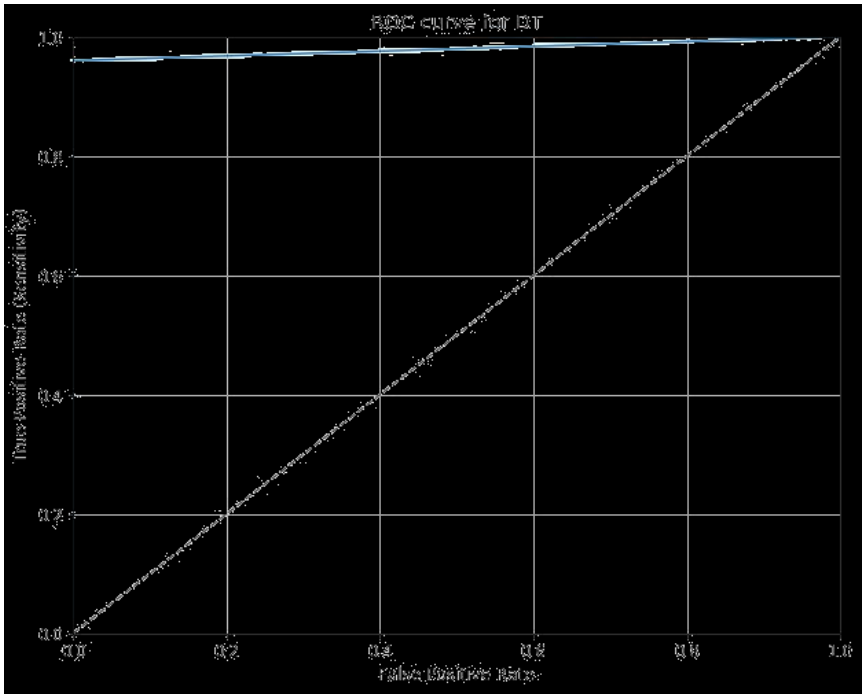


Figure 8. ROC curve of Decision Tree

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.

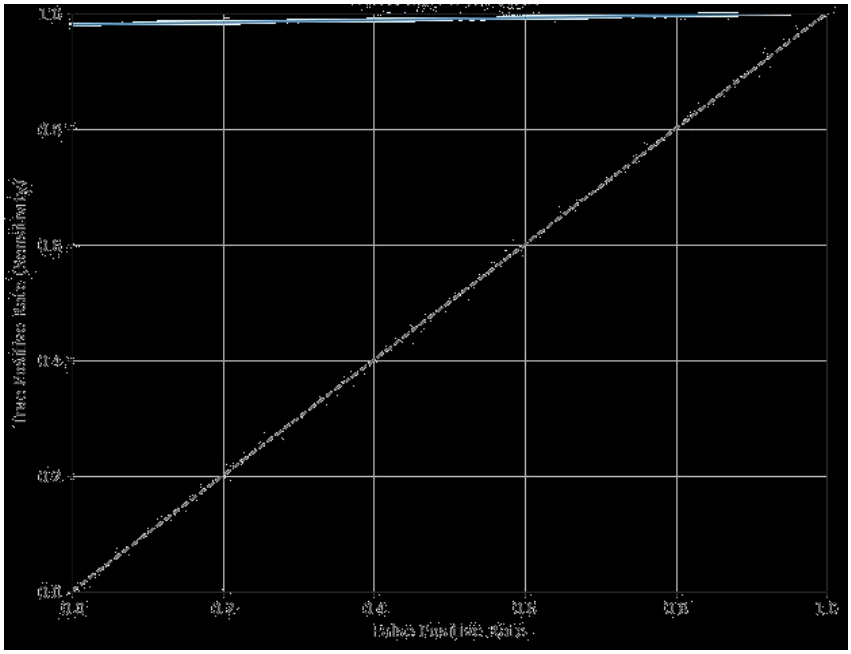


Figure 9. ROC curve of AdaBoost

In figure 9, the value of true positive rate is 98.38% and false positive (1-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.

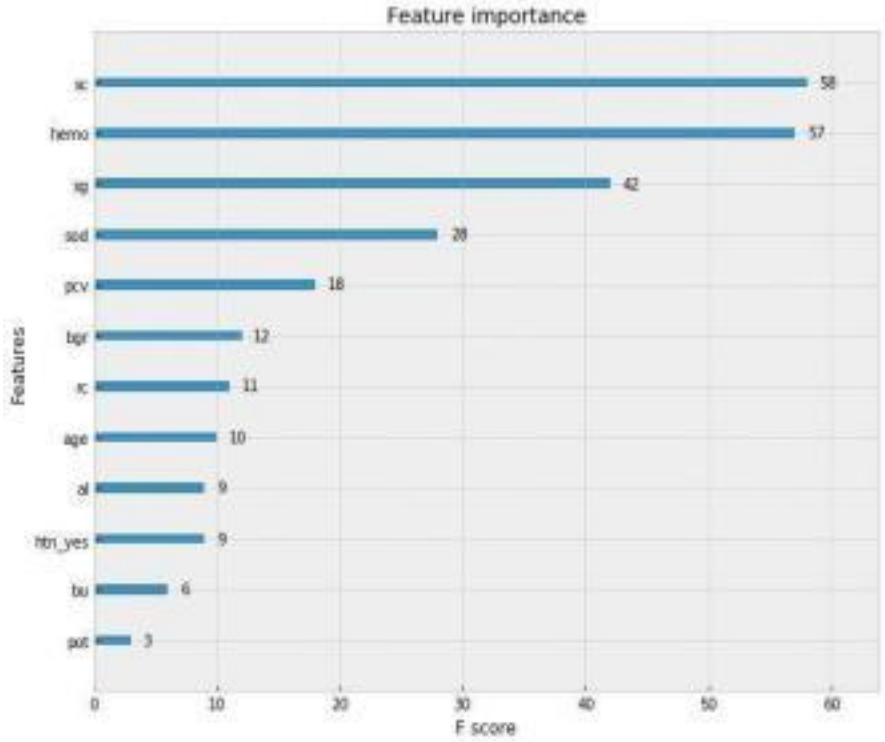


Figure 10. Feature ranking applying XGBoost

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

Table 8 : methodology Comparison

Author	Total Cases	Techniques	Accuracy
Gupta D. et al. [1]	01	AdaBoost	98.87% and 88.66%
A. Y. Al-Hyari et al. [2].	01	Artificial Neural Network and Naive Bayes	88.2% and 82.4% respectively.
Salekin A. [3]	01	K-NN, Random Forest and Neural Network.	99.3% (F1 Score in Best Case)
Our Proposed Methodology	04	Logistic Regression, AdaBoost, SVR, XGBoost, Gradient Boosting, Decision Tree Regressor	100% (in Best 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, rising quickly and maxing out at the carrying capability of the setting. It's an formed curve which will take any real-valued range and map 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 logarithms (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 remodeled 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.

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 (2013),"Chronic kidney disease: global dimension and perspectives", The Lancet.

[2] 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)

[3] 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-V15-335.

[4] A. Salekin and J. Stankovic, (2016) "Detection of chronic kidney disease and selecting important predictive attributes," IEEE International Conference on Healthcare Informatics.

[5] D. Gupta, S. Khare, and A. Aggarwal, (2016)"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.

[6] 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.

[7] 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.

[8] 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.

[9] 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).

[10] Good, David M., Petra Zürbig, Angel Argiles, Hartwig W. Bauer, Georg Behrens, Joshua J. Coon, Mohammed Dakna et al. "Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease." Molecular & cellular proteomics 9, no. 11 (2010): 2424-2437.

[11] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[12] Ogunleye, A. A., & Qing-Guo, W. (2019). XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM transactions on computational biology and bioinformatics.

[13] 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).

[14] T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016.

[15] 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

[16] 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.

[17] 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.

[18] “Introduction to Boosted Trees, [Online]. Available: https://xgboost.readthedocs.io/en/latest/tutorials/model.html. [Accessed: 21-Dec-2018].

[19] 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.

[20] J. Yang (2006) “Power system short-term load”, PhD thesis, Technische Universit¨.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, (1984) “Classification and regression trees. belmont, ca: Wadsworth”, International Group, p. 432.

[22] Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.

[23] 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.

[24] 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,

[25] “Boosting algorithm: XGBoost,” Towards Data Science. [Online]. Available: https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d. [Accessed: 22-Jul-2018].

[26] Boosting and AdaBoost for Machine Learning [Online]. Available: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/ [Accessed: 13-Aug-2019].

[27] Morde, V. (2019). “Feature Importance and Feature Selection With XGBoost in Python” [Online]. Available: https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[28] Chronic_Kidney_Disease Data Set [Online]. Available: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease