KIDNEY DISEASE PREDICTION USING MACHINE LEARNING

BY
MD SAJIB HOSSAIN
ID: 171-15-9122

AND
MD MUSHFIQUR RAHMAN
ID: 171-15-8730

AND
ABDULLAH AL ROMAN
ID: 171-15-8616

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

Supervised By
Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-supervised by:
Majidur Rahman
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2021

APPROVAL

This Project titled **“Kidney disease prediction using machine learning”**, submitted by MD Sajib Hossain, MD Mushfiqur Rahman and Abdullah Al Roman to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 27th January, 2021.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Zahir

Gazi Zahirul Islam **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Raja Tariqul Hasan Tusher **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Farid

Dr. Dewan Md. Farid **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

DECLARATION

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

Supervised by:

Sadik

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-supervised by:

for Sadik

Majidur Rahman
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md Sajib Hossain
ID: 171-15-9122
Department of CSE
Daffodil International University

মুশফিকুর রহমান

Md Mushfiqur Rahman
ID: -171-15-8730
Department of CSE
Daffodil International University

Roman

Abdullah Al Roman
ID: 171-15-8616
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We are very grateful and thankful to **Almighty Allah** for giving us His blessings and showed us the right way to successfully complete this final year thesis timely.

We are really thankful and wish our significant obligation to **Supervisor Md. Sadekur Rahman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Profound Knowledge and unmistakable fascination of our supervisor in the field of "Artificial Intelligence" to do this project. His unending tolerance, insightful direction, nonstop support, steady and enthusiastic management, productive analysis, significant counsel, perusing numerous sub-par drafts and revising them at all stage have made it conceivable to finish this project.

We are also really thankful to **Holy Care Hospital, Purbo Bazar, Feni Road, Chowmuhani, Noakhali** for providing us the medical report of various patients which helped us a lot in making this paper successful one.

We would also like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan, Head, Dept. of Computer Science & Engineering** and other faculty members and the staffs of CSE department of Daffodil International University.

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

ABSTRACT

In present, Machine learning is one of the most used technique to predict various disease in health sector. Not only in health sector, but also in every sector, predicting something using machine learning has been a popular one. In Bangladesh, most people are unaware of Kidney related problems. When they suffer from serious kidney problems, they start to take treatment. But if they know about the features which are important to find kidney related problems, then the total number of affected people will be decreased. We have applied various feature selection method in our collected dataset to find the most suitable attributes which are responsible for various kidney disease such as: CKD (Chronic Kidney Disease), ESRD (end stage Renal Disease), UTI (Urinary Tract Infection). We have applied 11 machine learning model to predict the various kidney diseases. We have split our dataset into train dataset and test dataset to predict the kidney disease. Highest performance was achieved by Random Forest Classifier and Decision Tree Classifier. In future, we will extend our dataset to find even more better performance and we will implement it using an android application and web application.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

TABLES

LIST OF ABBREVIATION

CKD – Chronic Kidney Disease

ESRD – end stage Renal Disease

UTI – Urinary Tract Infection

RBC – Red Blood Cell

BP – Blood Pressure

Hb – Hemoglobin

PCV – Packed Cell Volume (Hematocrit)

SC – Serum Creatinine

KNN – K-Nearest Neighbors.

SVM – Support Vector Machine.

LDA – Linear Discriminant Analysis.

ANN – Artificial Neural Network.

CNN – Convolutional Neural Network

MLP – Multilayer Perceptron

RBFN – Radial basis function network

DT – Decision Tree

RF – Random Forest

LR – Logistic Regression

CART – Classification and Regression Trees

LVQ – Learning vector quantization

RBF – Radial Basis Function

NB – Naïve Bayes

EM – Expectation–Maximization

CM – Confusion Matrix

ROC – Receiver Operating Characteristic Curve

AB – ADA Boost

GB – Gradient Boost

XGB – Extra Gradient Boost

AUC – Area Under The Curve

CHAPTER 1

INTRODUCTION

1.1 Introduction

In this century, people are suffering from serious diseases related to heart, kidney, liver, lung etc. One of the common but serious one is kidney disease. Kidney is the most important and vital part of human body. The main function of kidney is to remove waste materials and toxic fluids from human body. Kidney is also responsible for the regulation of body's acid content, potassium, and salt. It also produces hormones which controls blood pressure and also stimulates red blood cell production and also helps to control calcium metabolism. Kidney can filter and return 200 quarts of fluid in the bloodstream. Almost 2 quarts of fluid is removed with the urine and 198 quarts of fluid is recovered [18]. Whether a kidney has a problem or not can be found out by blood test and urine test. In blood test, there are some attributes that are related to kidney problems. Red blood Cell in blood has a specific value which is 4.5-6.5 for male and 3.8-5.8 for female. Higher than normal will cause Renal Cell Carcinoma (Kidney Cancer) and lower than normal will cause anemia with chronic kidney disease. If Blood Pressure has higher than 80 diastolic pressure, that will cause kidney problem. Another one Hemoglobin has also a reference value which is 13-18 for male and 11-16.5 for female. Lower than that causes kidney disease. Packed Cell Volume is also important feature to detect kidney disease. The reference value of male is 40%-54% and for female is 37%-47%. Lower than that causes kidney disease. The most important feature is Serum Creatinine. The reference value of SC is 0.6-1.4. Higher than that, causes kidney diseases. Initial stage of kidney disease can be experienced by some symptoms or signs but most of the time it remains hidden until a major test is performed. In the last stage, kidney failure can happen that will result into death. The treatment to recover that is transplanting a kidney.

In this research, we have applied a total of 11 classification method which are Logistic Regression Classifier, Random Forest Classifier, Decision Tree Classifier, Nave Bayes Gaussian Classifier, K-Nearest Neighbors Classifier (KNN), Support Vector Machine Classifier (SVM), ADA Boost Classifier, Gradient Classifier, XG Boost Classifier, Fisher's Linear Discriminant Analysis Classifier (LDA), Artificial Neural Network

Classifier (ANN).

The objective of our thesis:

- To observe the most important features which are helpful to predict kidney disease.
- To observe the performance from the dataset that we have collected by applying different classification techniques.
- To observe the most suitable classification technique.

1.2 Motivation

In Bangladesh, most of the people are not conscious about kidney related problems. They don’t really know whether they are suffering from kidney disease or not.

Kidney related problem causes estimated 1.2 million deaths yearly in the world. About 35,000-40,000 CKD patients, out of about 18 million people, develop kidney failure every year in Bangladesh. According to a study, higher than 40 years old people are likely to suffer from kidney disease. There is not much of study related to kidney prediction which has better performance.

Other than this, we see that today’s world is so much focusing on recommendation system. Users expect all that the better things will be prescribed to them by the system.

To make a system to be recommendation should be able to take choice without anyone else. To take decision without anyone else should have to have classified data.

All these reasons made us to do this kind of research where we will classify data and predict kidney disease.

1.3 Rationale of the Study

There are many works on kidney disease prediction. However, these much classification techniques dealing with kidney disease data is rare. Right now, we see that there are so many researches on this kidney disease yet every one of those are not that much unmistakable from one another while we are utilizing 11 different classification techniques to discover the best accuracy results which methods provide for make a further application in future.

1.4 Research Question

- Can we collect authentic data from any hospital?
- Can we pre-process data and find the best features?
- Can our dataset give us a better accuracy than other related works?

1.5 Expected Outcome

Expected result of this research-based thesis is to construct a calculation or making a total proficient system that will find Kidney Disease dataset concerning the assembled model of prepared dataset.

Our applied algorithms will find the best features which are responsible for various kidney related diseases and can be use on new data to predict kidney diseases. Also we found best accuracy on different machine learning algorithm and that can be use in future applications. The model's performance will be attempted by applying different kinds of machine learning algorithms on our dataset. The work would test how exact estimations react to our dataset.

1.6 Report Layout

Chapter 1 Gives a brief Introduction related to kidney, kidney function, important features, name of applied algorithm, objectives of the research. Also discusses about our thesis motivation, Rationale of the Study, Research Question and Expected Outcome.

Chapter 2 Gives a short introduction related to our paper, a brief description about previous study related to our work, a research summary, scope of the problems and challenges that we had to overcome.

Chapter 3 Gives a short introduction, Research Subject and Instrumentation, data collection procedure and implementation requirement. Also, a brief description about data collection, pre-processing, merging and cleaning data and finally feature selection.

Chapter 4 Gives a short introduction on model training, a brief experimental result on eleven machine learning model, descriptive analysis and summary of the results.

Chapter 5 Discusses summary of study, conclusion, recommendation and future work.

CHAPTER 2

BACKGROUND

2.1 Introduction

This section mirrors the connected works that all around done by certain researchers in the past time in this field. Plus, giving an away from of this, this part will show what the restrictions of these works were and ultimately, this section portrays extent of our examination just as the difficulties of it.

2.2 Related Works

Predicting any disease using machine learning has become a popular way nowadays. Lots of researchers have implemented various algorithm, various data mining techniques to classify various diseases. Some related study has been observed by us to develop this paper.

Kerina Blessmore Chimwayi, Noorie Haris, Ronnie D. Caytiles, N. Ch. S. N. Iyenger has given a statement about the risk level prediction of CKD using Neuro-Fuzzy classifier. They also used Hierarchical Clustering and found three form of clusters and a strong relationship between CKD and diabetes [1]. Another work on prevalence of Kidney disease in women with the help of their blood test report was done by Dr. S. Gomathi alias Rohini, C. Karpagam in 2020. They implemented two classification techniques and compared them. They also collected data from 32 districts in India [2]. The Diagnosis and Estimate of Chronic Kidney Disease using the Machine Learning Methods was done by Enes Celik, Muhammet Atalay, Adil Kondiloglu in 2016. They used a dataset from UCI Repository and used Decision Tree Classifier and Support Vector Machine Classifier [3]. K. R. Lakshmi, Y. Nagesh, M. VeeraKrishna has worked on kidney disease in 2014 by using ANN, Decision Tree, Logical Regression [4]. J Van Eyck, J Ramon, F Guiza, G Meyfroidt, M Bruynooghe & G Van den Berghe has worked on data mining technique on acute kidney injury on 2012. Their aim was to develop a statistical model capable of predicting the occurrence of AKI in patients after elective cardiac surgery using Gaussian process [5]. In 2012, Morteza Khavanin Zadeh, and Mohammad Mehdi Sepehri discussed about Early AVF Failure using W-Simple Cart and WJ48[6]. In 2012, Abeer Y. Al-Hyari, Ahmad M. Al-Taee, Majid A. Al-Taee worked on 102 patients and applied ANN, Naïve Bayes and Decision Tree [7]. Another work on Renal Failure Hemodialysis was done by Xudong Song, Qiu Zhanzhi, Jianwei Mu. They introduce briefly data mining technology, focuses on data mining decision tree classification method, and proposes a new variable precision rough set decision tree classification algorithm based on weight limit number explicit region [8]. In 2009, Kaushal Kumar & Abhishek worked on ANN for Diagnosis of Kidney Stones Disease. They introduce briefly diagnose kidney stone disease by using various algorithm which have different characteristics and architecture [9]. In 2020, Shawni Dutta and Prof. Samir Kumar Bandyopadhyay proposed and implemented Neural network model and 10-fold cross validation methodology under a single platform and classified patients with CKD [10]. Also Dr. S. Vijayarani, Mr. S. Dhayanand research work was related to predict kidney diseases using classification algorithms such as Naïve Bayes and Support Vector Machine in 2015. Their work focused on the classification accuracy and execution time performance factors. From the results it was observed that the SVM is better than the Naive Bayes classifier [11]. There was another excellent work done by Sunil Belur Nagaraj, Michelle J. Pena, Wenjun Ju, Hiddo L. Heerspink in 2020. They predicted ESRD with machine learning models with multiple baseline demographic and clinical characteristics [12]. In 2019, S. Mahalakshmi, P. Menaka, R.S. Rajkumar used some classification techniques such as CNN, MLP, RBFN to predict kidney disease in diabetic patients. Their main work was to find the best classification technique based on excellent accuracy [13]. In 2017, Tabassum S, Mamatha Bai B G, Jharna Majumdar used data mining technique to predict CKD [14]. Another was in 2017, by Maryam Soltanpour Gharibdousti, Kamran Azimi, Saraswathi Hathikal, Dae H Won. They applied Decision Tree, Linear Regressing, Super Vector Machine, Naive Bayesian and Neural Network model to predict kidney disease. The dataset was also collected from UCI repository [15]. The Hybrid Neural model and some machine learning algorithm such as: LR, NB, SVM, GBDT was proposed by Yafeng Ren, Hao Fei, Xiaohui Liang, Donghong Ji and Ming Cheng in 2018. The proposed model outperformed traditional statistical models with discrete features and neural baseline systems [16]. In 2019, Jing Xiao, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, and Zhibin

Ye developed a machine learning tool in the prediction of chronic kidney disease progression [17].

2.3 Research Summary

After discussing the related works, we found out that, the dataset used in the research is from UCI repository, whereas we are using a dataset provided by Holy Care Hospital. Pre-processing data, dealing with missing value, feature selection and model training on that dataset provide us a better accuracy.

Table 2.1- Research Summary

No	Author	Year	Domain	Algorithm	Accuracy
1	Kerina Blessmore Chimwayi, Noorie Haris, Ronnie D. Caytiles, N. Ch. S. N. Iyenger[1]	2017	CKD	Neuro-Fuzzy	97%
2	Dr. S. Gomathi alias Rohini, C. Karpagam[2]	2020	Kidney	DT RF	98% 96%
3	Enes Celik, Muhammet Atalay, Adil Kondiloglu[3]	2016	CKD	DT SVM	96% 91%
4	K. R. Lakshmi, Y. Nagesh, M. VeeraKrishna[4]	2014	Kidney	ANN DT LR	93% 78% 74%
5	J Van Eyck, J Ramon, F Guiza, G Meyfroidt, M Bruynooghe, G Van den Berghe[5]	2012	AKI	Gaussian	75%
6	Mohammad Rezapour, Morteza Khavanin Zadeh, and Mohammad Mehdi Sepehri[6]	2013	Early AVF Failure	CART	85%
7	Abeer Y. Al-Hyari, Ahmad M. Al-Taee, Majid A. Al-Taee[7]	2013	CKD	DT	--
8	Xudong Song, Qiu Zhanzhi, Jianwei Mu[8]	2012	RFH	DT	60%
9	Kaushal Kumar, Abhishek[9]	2009	Kidney Stone	MLP LVQ RBF	92% 84% 87%
10	Shawni Dutta and Prof. Samir Kumar Bandyopadhyay[10]	2020	CKD	KNN DT	91% 94%
11	Dr. S. Vijayarani, Mr. S. Dhayanand[11]	2015	CKD	NB SVM	70% 76%
12	Sunil Belur Nagaraj, Michelle J. Pena, Wenjun Ju, Hiddo L. Heerspink[12]	2020	CKD	LR SVM RF	77% 78% 80%

No	Author	Year	Domain	Algorithm	Accuracy
13	S. Mahalakshmi, P. Menaka, R.S. Rajkumar[13]	2019	CKD	RBFN MLP CNN	85% 91% 94%
14	Tabassum S, Mamatha Bai B G, Jharna Majumdar[14]	2017	CKD	EM ANN	70% 75%
15	Maryam Soltanpour Gharibdousti, Kamran Azimi, Saraswathi Hathikal, Dae H Won[15]	2017	CKD	ANN	63%
16	Yafeng Ren, Hao Fei, Xiaohui Liang, Donghong Ji, Ming Cheng[16]	2018	CKD	LR NB SVM	64% 67% 42%
17	Jing Xiao, Ruifeng Ding, Xiulin Xu, Haochen Guan, Xinhui Feng, Tao Sun, Sibo Zhu, Zhibin Ye[17]	2019	CKD	LR SVM KNN XG-Boost	82% 81% 74% 83%

2.4 Scope of the Problem

In Bangladesh, many people suffer from kidney disease and dies because of not taking proper medication. Also, they are not aware of this disease. They don’t know or don’t take proper prevention to deal with the disease. Also, there are not much work or study related to kidney disease that is use by any organization. Some study was made but they used the dataset from internet which might not be valid one. In our research we have used an authentic dataset given by a hospital and predicted kidney disease with a better performance. Many applications can be built with our proposed model and dataset.

2.5 Challenges

Several challenges were overcome including collecting data from patient’s report was a tough one. Firstly, we have sort out all the reports that have been provided by hospital. Then we collected the data from the report and insert it into excel file. Dealing with the missing value, pre-processing data was harder one. Also feature selection, algorithm selection which is really tough for the initial stage. Despite these challenges it was possible to extract visual information contained in the high-risk data for patients within the study utilizing machine learning.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter mainly deals with the data collection, pre-processing, and feature selection. The brief description has been given here. How we have collected our data, then merge it with another dataset and after this we have clean the dataset and performed some algorithms to find out the strongly correlated features for further processing.

3.2 Research Subject and Instrumentation

We mean by research subject is that exploration region that is being read and investigated for clear understandings. For clear arrangement, yet in addition research subject is liable for giving the correct information on different examination boundaries. Then again, Instrumentation alludes to the necessary instruments or apparatuses that are utilized by the analysts.

3.3 Data Collection Procedure

To research on distinct field, the rapid and head thing is the Data. Data is, genuinely, thought of as the focal point of the machine learning process. Furthermore, for our research, we must choose between limited options of data. Thusly, it has become our most testing task for our research. We have collected 208 data from a hospital and 400 data from UCI repository.

3.4 Implementation Requirement

For completing our research work we have kept up several steps. Those all are connected with each other. As we worked with data so we need to clean up our data prior to fitting them in our supervised algorithms. There were similarly some more significant steps in our exploration work. Figure 3.4.1 exhibiting the entirety of the methods for our research methodology.

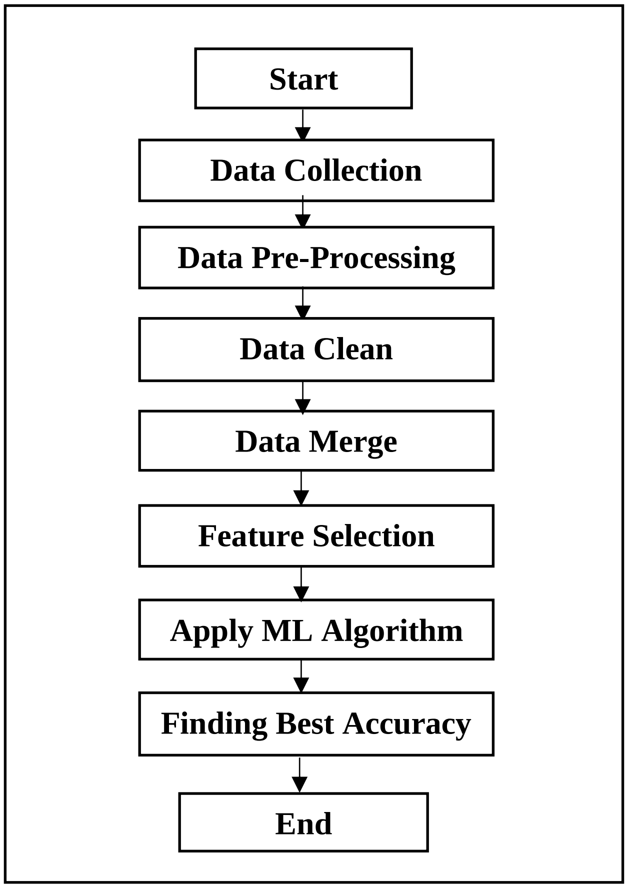


Figure 3.1- Research Methodology

3.4.1 Data Collection

We have collected 208 data from a hospital. Hospital management happily gave us the report of the patients. Then we talked with some doctor to find the features which are relatable to kidney. After that we have collected those data and insert it into excel file. As we have collected only 208 data, we needed more data to apply machine learning algorithms. That's why we have collected 400 data from UCI repository and after some process we merge those data.

3.4.2 Data Preprocessing

In our dataset there are 21 attributes.

Table 3.1- Dataset Attributes

No	Attribute (Short Form)	Attribute (Full Form)	Unit
1	age	Age	Years
2	sex	Sex	N/A
3	bp	Blood Pressure	mmHg
4	ht	Hypertension	N/A
5	rbs	Red Blood Sugar	Mg/du
6	an	Anemia	N/A
7	rbc	Red Blood Cells	m/ul
8	hb	Hemoglobin	g/dl
9	esr	Westergren	mm/1st hour
10	wbc	White Blood Cell	/cumm
11	pcv	.Packed Cell Volume	N/A
12	plt	Platelet	/cumm
13	ec	Epithelial Cells	/HPF
14	pc	Pus cell	/HPF
15	pcc	Pus cell clumps	N/A
16	sc	Serum creatinine	mg/dl
17	ap	Appearance	N/A
18	sg	Specific gravity	N/A
19	al	Albumin	N/A
20	bu	Blood urea	/HPF
21	class	Class	N/A

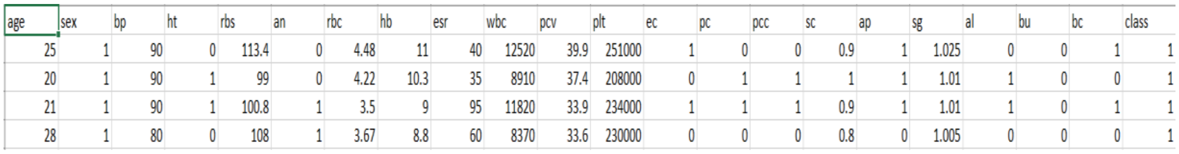
3.4.3 Data Clean

There are two types of data in our dataset. One is categorial and another is numerical. For simplicity, we converted all the categorial values into numerical value by one hot encoding procedure.

Table 3.2- Clean Dataset

Name	Encoding
sex	Male=0, Female=1
ht	Yes=1, No=0
an	Yes=1, No=0
pcc	Present=1, Not Present=0
ap	Turbid=1, Clear=0
al	Trace=1, Nil=0
bu	Trace=1, Nil=0
bc	Present=1, Not Present=0
class	Present=1, Not Present=0

After encoding, we have to deal with two attributes which are ec and pc, because these attributes have value like (5-6), (10-20). For this reason, we have followed a technique. For ec: The minimum value was 0, maximum value was 30. The median was 15. So greater than 15 was categorized into 1 and less than or equal to 15 was represented by 0. For pc: The minimum value was 0, maximum value was 25. The median was 12.5. So greater than 12.5 was categorized into 1 and less than or equal to 12.5 was represented by 0. After that, we deal with the missing values. For this, we find out the mean value of entire column and replaced it by mean value.



age	sex	bp	ht	rbs	an	rbc	hb	esr	wbc	pcv	plt	ec	pc	pcc	sc	ap	sg	al	bu	bc	class
25	1	90	0	113.4	0	4.48	11	40	12520	39.9	251000	1	0	0	0.9	1	1.025	0	0	1	1
20	1	90	1	99	0	4.22	10.3	35	8910	37.4	208000	0	1	1	1	1	1.01	1	0	0	1
21	1	90	1	100.8	1	3.5	9	95	11820	33.9	234000	1	1	1	0.9	1	1.01	1	0	1	1
28	1	80	0	108	1	3.67	8.8	60	8370	33.6	230000	0	0	0	0.8	0	1.005	0	0	0	1

Figure 3.2- Clean Dataset

3.4.4 Data Merge

After pre-processing data, we merge our dataset with the UCI repository dataset. In our dataset we have taken 22 attributes which are most important for kidney. Based on this we have taken only those attributes which are matched with our dataset.

3.4.5 Feature Selection

First of all, we will give a brief description using a table and figure of all the attributes that we have used.

Table 3.3- Description of attributes

Name	Mean	Standard Deviation	Minimum	Maximum
Age	44.808388	19.545154	0.500000	94.000000
BP	77.935855	12.273013	30.000000	180.000000
RBS	118.216118	28.406000	73.800000	444.600000
RBC	4.364786	1.096669	0.100000	6.500000
HB	12.050987	2.477184	3.100000	17.800000
ESR	45.458224	12.631848	7.000000	120.000000
WBC	9130.279605	2736.904172	2200.000000	26400.000000
PCV	38.701974	8.090350	1.000000	67.000000
PLT	303841.983553	56815.691489	21400.000000	692000.000000
SC	1.647023	1.647416	0.400000	13.000000
SG	1.014298	0.011177	0.800000	1.025000

We have found out all the histogram for all attributes and here we have showed only for four attributes.

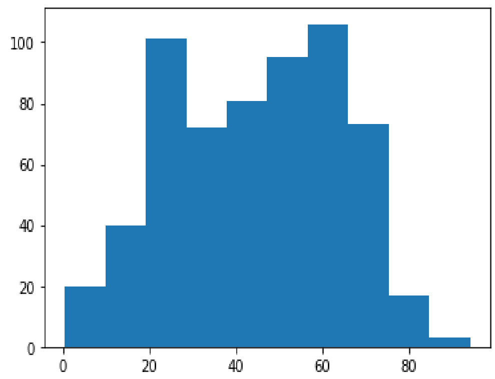


Figure 3.3- Age Histogram

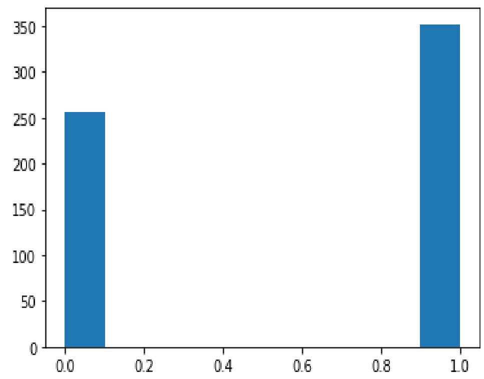


Figure 3.4- Sex Histogram

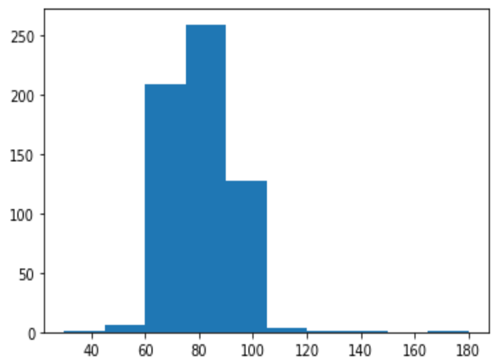


Figure 3.5- BP Histogram

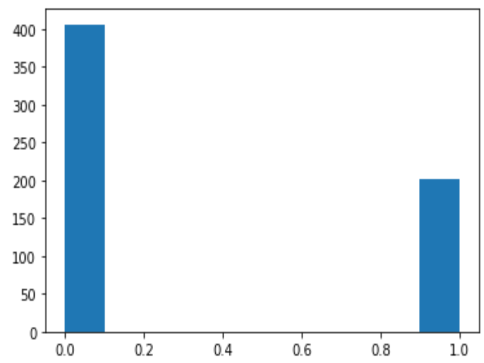


Figure 3.6- Ht Histogram

In our dataset, “class” attribute is the dependent attribute and all other attributes are independent attribute. We find out the correlation between “class” and 21 other attribute using simple corr() method. There are 3 types of method to find strong correlated features. They are:

1. Filter Method
2. Wrapper Method
3. Embedded Method

In filter method, we have used Pearson Correlation technique, Chi Square technique. For wrapper method, Recursive Feature Elimination technique has been used. In embedded method, two types of techniques have been used. One is Lasso L1, and another one is Tree-based technique. Also, we implemented Univariate Selection to find best features. Figure 3.25 shows us the heatmap of correlation between “class” and 21 other attributes.

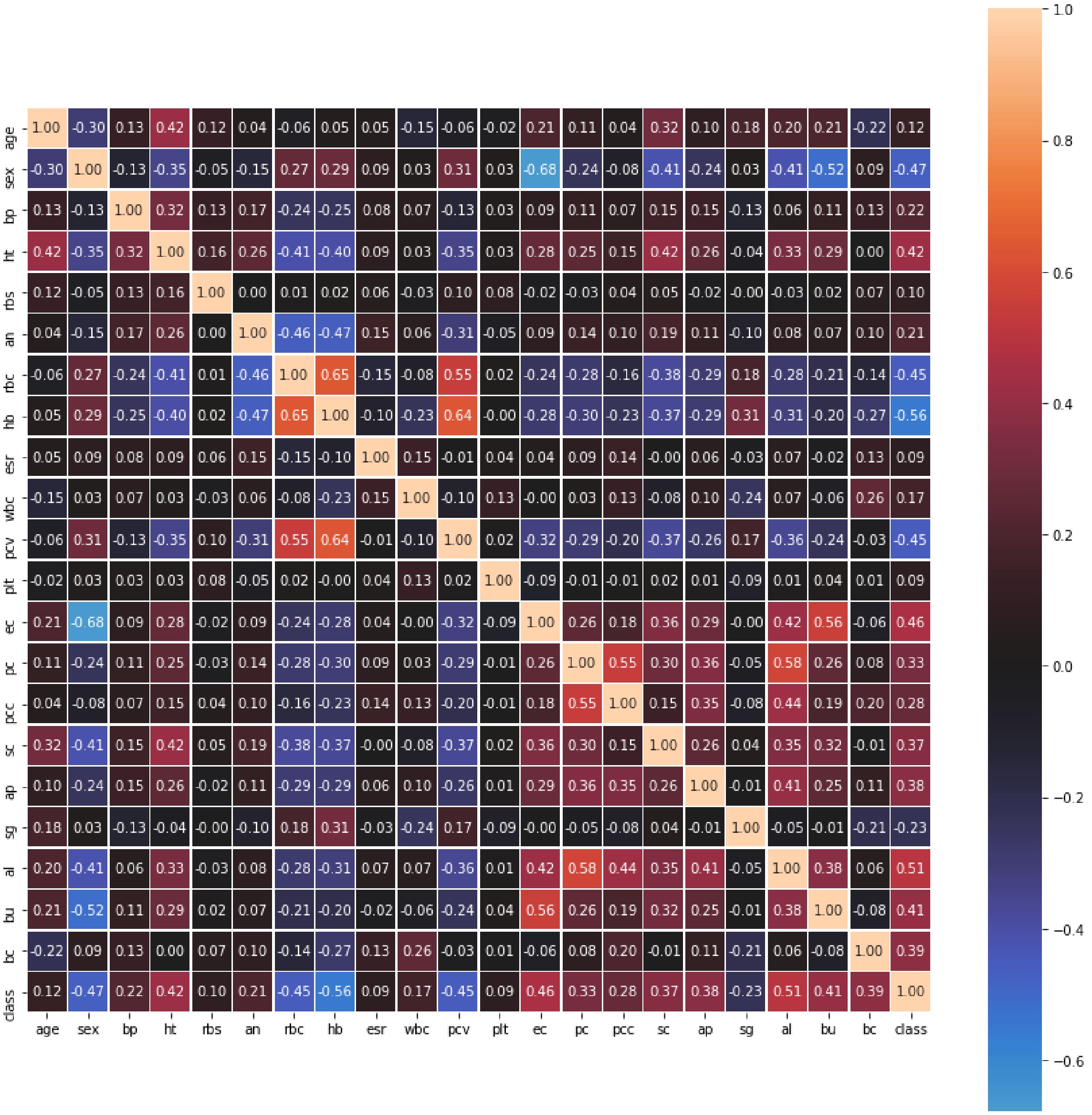
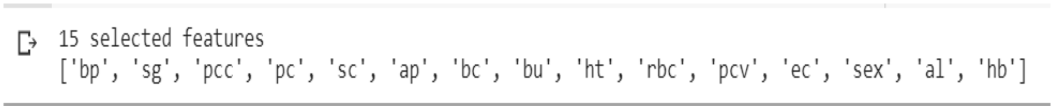


Figure 3.7- Correlation Heatmap

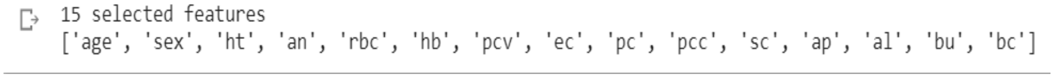
- **Pearson Correlation:** It can be found by calculating covariance of two variables divided by the product of their standard deviations. Figure 3.26 shows the top 15 features selected by Pearson algorithm.



```
15 selected features
['bp', 'sg', 'pcc', 'pc', 'sc', 'ap', 'bc', 'bu', 'ht', 'rbc', 'pcv', 'ec', 'sex', 'al', 'hb']
```

Figure 3.8- Pearson Correlation

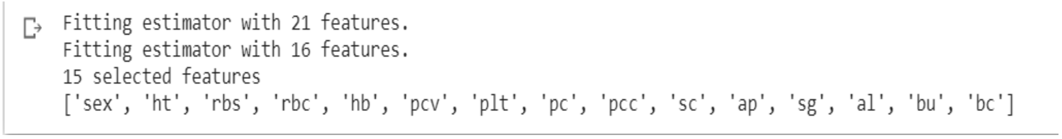
- **<u>Chi Square Features</u>:** We calculate X^2 between the target and every feature, and choose the desired output with the best X^2 scores. Figure 3.27 shows the top 15 features selected by Chi Square Features algorithm.



```
15 selected features
['age', 'sex', 'ht', 'an', 'rbc', 'hb', 'pcv', 'ec', 'pc', 'pcc', 'sc', 'ap', 'al', 'bu', 'bc']
```

Figure 3.9- Chi Square Features

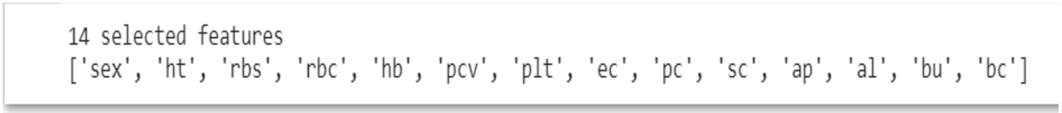
- **<u>Recursive Feature Elimination:</u>** This algorithm ranks features by importance. It discards the lowest importance feature and re-fit the model. Figure 3.28 shows the top 15 features selected by Recursive Feature Elimination algorithm.



```
Fitting estimator with 21 features.
Fitting estimator with 16 features.
15 selected features
['sex', 'ht', 'rbs', 'rbc', 'hb', 'pcv', 'plt', 'pc', 'pcc', 'sc', 'ap', 'sg', 'al', 'bu', 'bc']
```

Figure 3.10- Recursive Feature Elimination

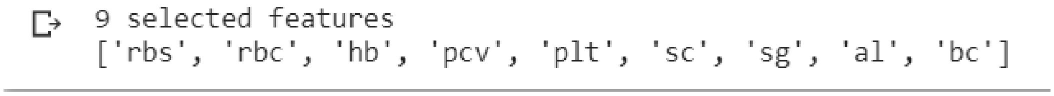
- **<u>Lasso L1- SelectFromModel</u>:** Lasso regression uses L1 norm as regularizer. Figure 3.29 shows the top 14 features selected by Lasso L1 algorithm.



```
14 selected features
['sex', 'ht', 'rbs', 'rbc', 'hb', 'pcv', 'plt', 'ec', 'pc', 'sc', 'ap', 'al', 'bu', 'bc']
```

Figure 3.11- Lasso L1

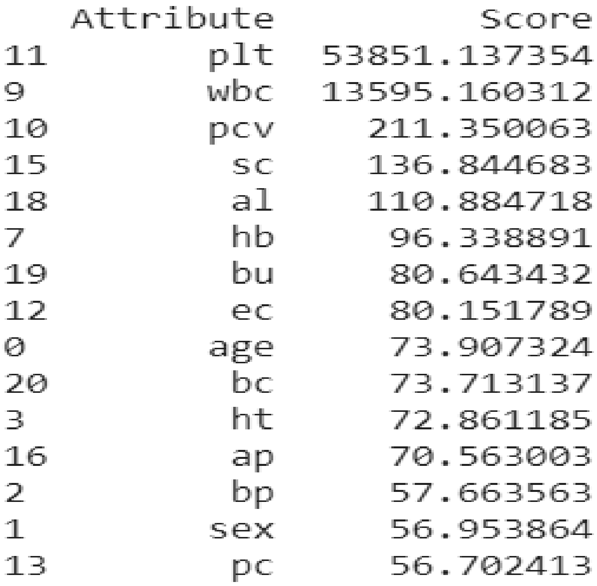
- **<u>Tree based-SelectFromModel:</u>** Random forest is the base interpreter of this feature selection technique. In random forest, the last feature is the mean of all decision tree feature importance. Figure 3.30 shows the top 9 features selected by tree-based algorithm.



```
9 selected features
['rbs', 'rbc', 'hb', 'pcv', 'plt', 'sc', 'sg', 'al', 'bc']
```

Figure 3.12- Tree-based

- **Univariate Selection:** Figure 3.31 shows the top 15 features selected by univariate selection.



```
     Attribute             Score
11         plt  53851.137354
9          wbc  13595.160312
10         pcv    211.350063
15          sc    136.844683
18          al    110.884718
7           hb     96.338891
19          bu     80.643432
12          ec     80.151789
0          age     73.907324
20          bc     73.713137
3           ht     72.861185
16          ap     70.563003
2           bp     57.663563
1          sex     56.953864
13          pc     56.702413
```

Figure 3.13- Univariate Selection

- **Information Gain:** Figure 3.32 shows the top features selected using information gain method.

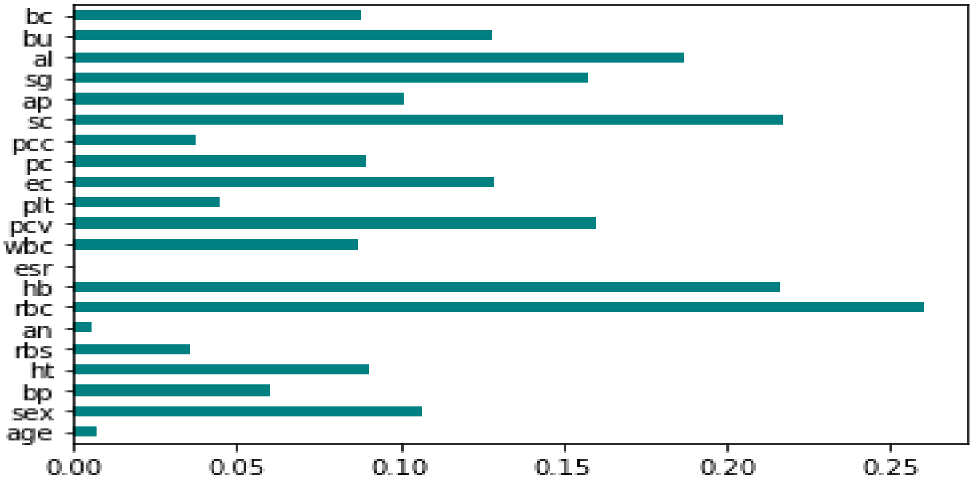
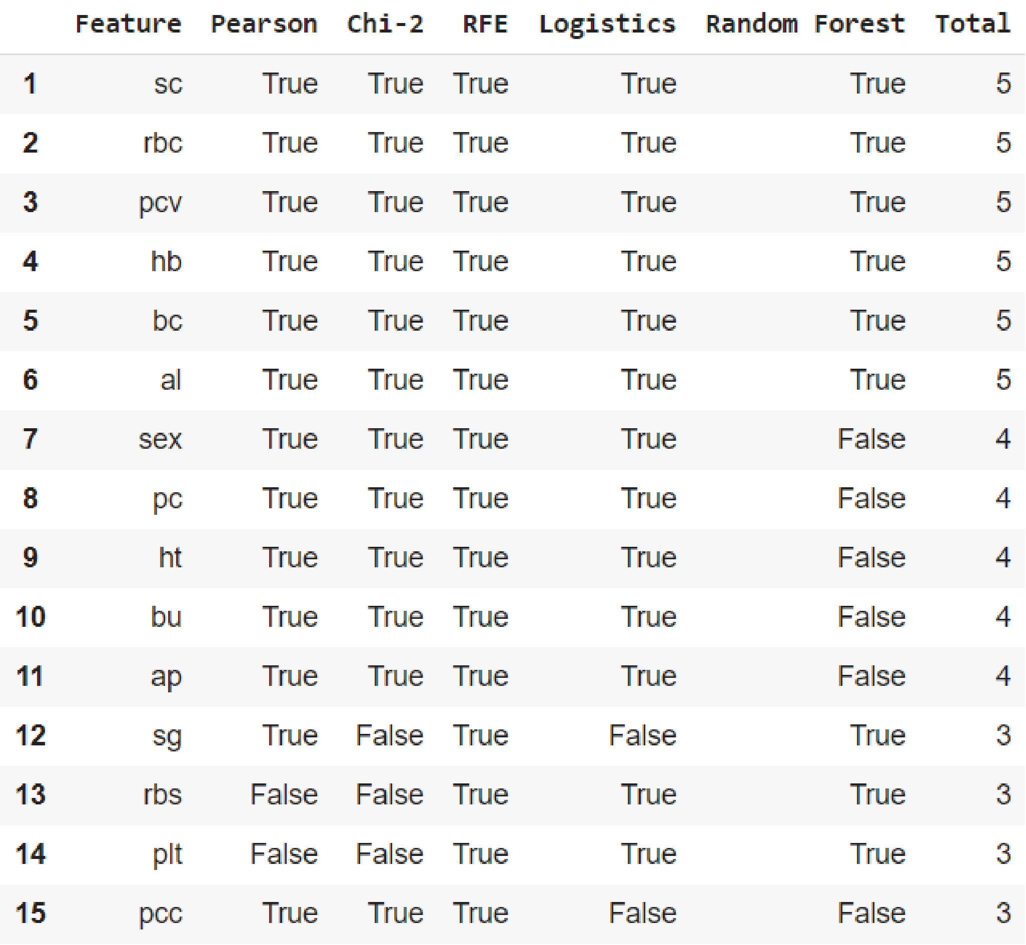


Figure 3.14- Information Gain

If we combine all of this the final features top 5 features will be 'sc', 'rbc', 'pcv', 'hb', 'bc', 'al'. Figure 3.33 shows the combined feature selection.



	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	Total
1	sc	True	True	True	True	True	5
2	rbc	True	True	True	True	True	5
3	pcv	True	True	True	True	True	5
4	hb	True	True	True	True	True	5
5	bc	True	True	True	True	True	5
6	al	True	True	True	True	True	5
7	sex	True	True	True	True	False	4
8	pc	True	True	True	True	False	4
9	ht	True	True	True	True	False	4
10	bu	True	True	True	True	False	4
11	ap	True	True	True	True	False	4
12	sg	True	False	True	False	True	3
13	rbs	False	False	True	True	True	3
14	plt	False	False	True	True	True	3
15	pcc	True	True	True	False	False	3

Figure 3.15- Feature Selection

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

For getting better accuracy we have implemented some feature selection methods, and applied machine learning models. For this we divided our data into two parts. Train data and test data. After dividing the dataset, model training algorithm will be implemented.

4.2 Experimental Results

In our dataset, there were 608 data. We are predicting “class” attribute which value are: Present, Not Present. Total number of Present was 364 and total number of Not Present was 234. We partitioned this dataset into 3:1. We will keep 456 data into train dataset where number of present is 281, and number of not present is 175, which is 75% of our main dataset. Remaining 152 data will be kept in test dataset where number of present is 93, and number of not present is 59, which is 25% of our main dataset. We applied model on our train dataset, and predicted with test dataset. Total 11 machine learning models were implemented and various performance like confusion matrix, accuracy, sensitivity, specificity, F1-score and ROC curve were found.

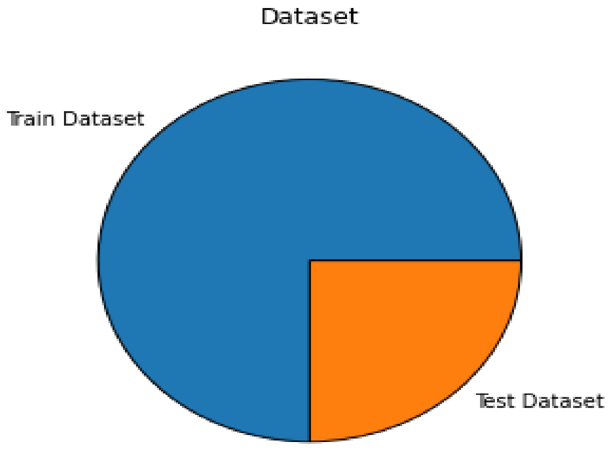
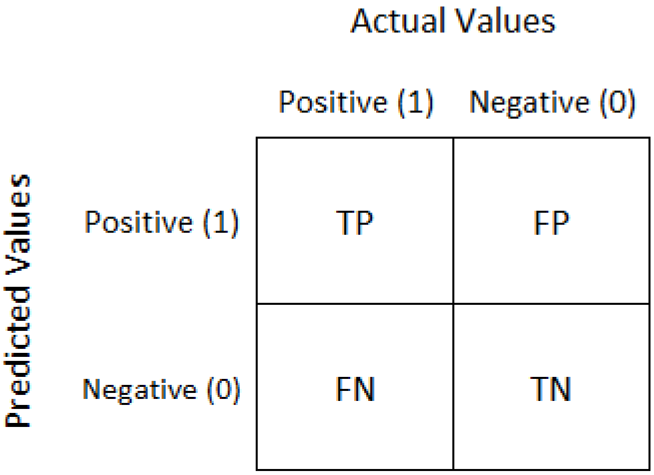


Figure 4.1- Dataset Pie Chart



Predicted Values		Actual Values	
		Positive (1)	Negative (0)
	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4.2- Confusion Matrix

Here we have discussed about the accuracy, F1-score, precision confusion matrix and ROC curve of different classification techniques.

Table 4.1- Performance of 11 models

Name	Accuracy	Precision	F1-Score	Recall	Sensitivity	Specificity
LR	90%	0.96	0.91	0.87	0.87	0.96
RF	96%	0.98	0.95	0.94	0.94	0.98
DT	96%	1	0.95	0.96	0.96	1
NB	85%	1	0.85	0.74	0.74	1
RF	90%	0.92	0.91	0.90	0.90	0.92
SVM	90%	0.99	0.91	0.84	0.84	0.99
AB	91%	0.91	0.93	1	1	0.91
GB	93%	0.94	0.93	0.93	0.93	0.94
XGB	93%	0.94	0.94	0.93	0.93	0.94
LDA	88%	0.86	0.90	0.94	0.94	0.86
ANN	93%	1	0.94	0.90	0.90	1

For Logistic Regression Classifier

1) **Confusion Matrix:**

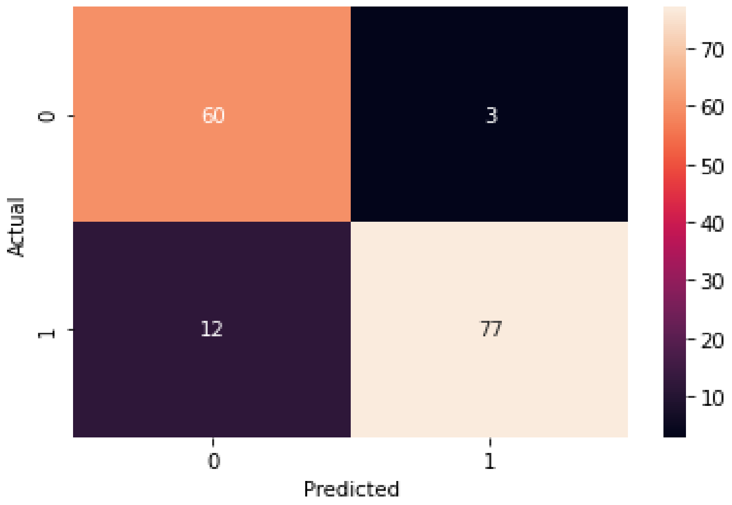


Figure 4.3- CM of LR

2) **ROC Curve:**

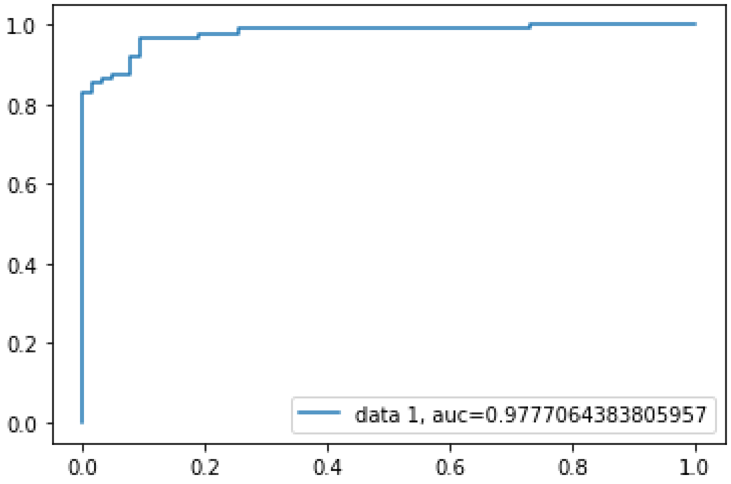


Figure 4.4- ROC of LR

For Random Forest Classifier

1) **Confusion Matrix:**

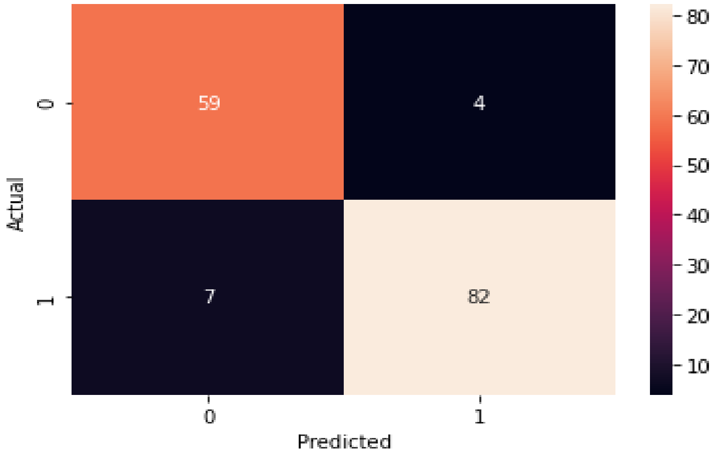


Figure 4.5- CM of RF

2) **ROC Curve:**

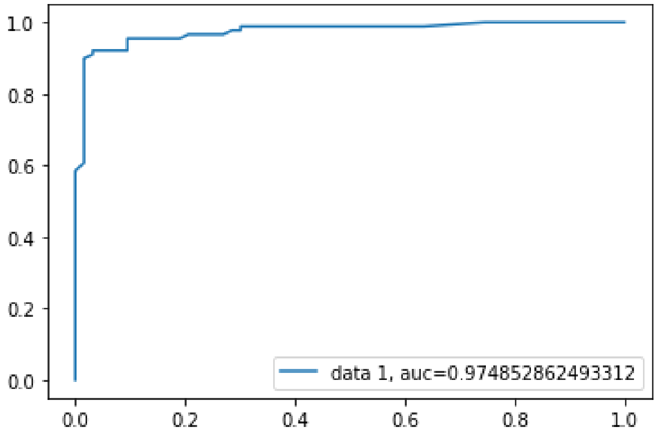


Figure 4.6- ROC of RF

For Decision Tree Classifier

1) Confusion Matrix:

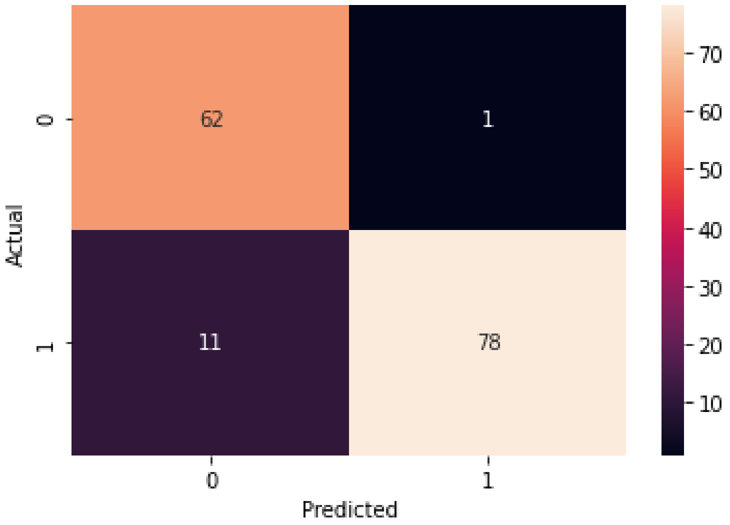


Figure 4.7- CM of DT

2) ROC Curve:

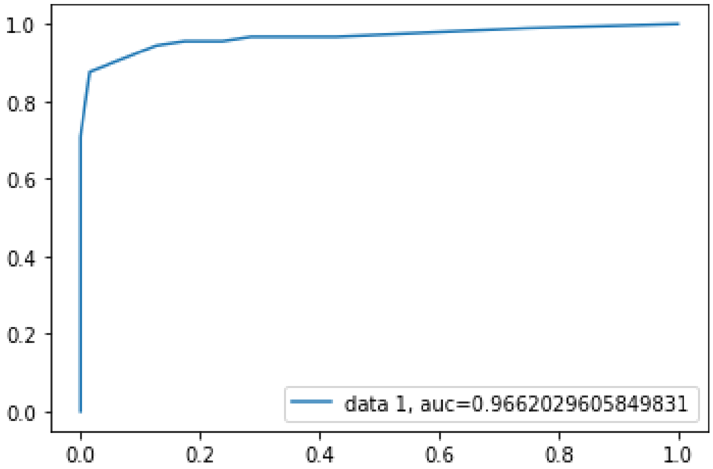


Figure 4.8- ROC of DT

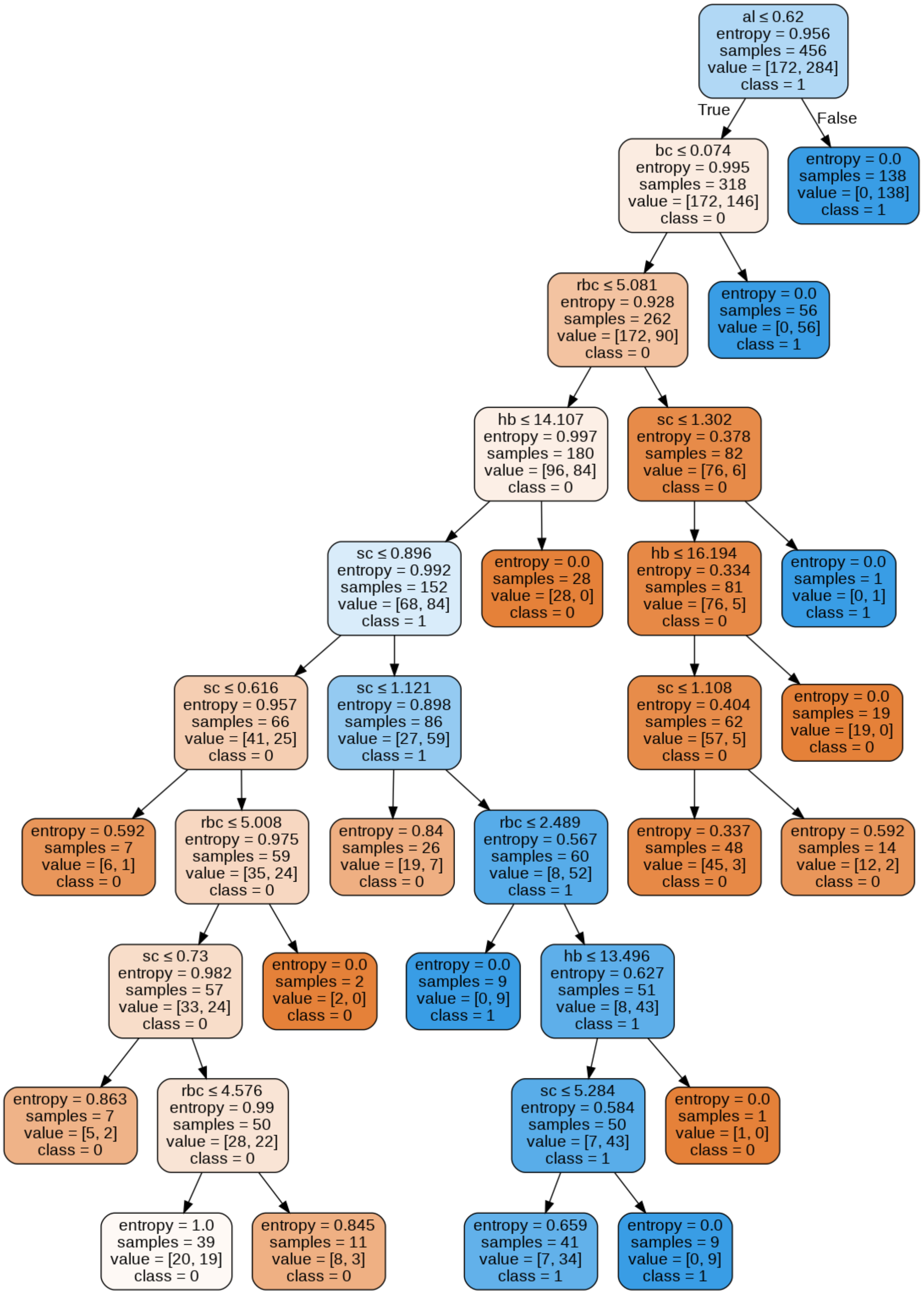


Figure 4.9- Decision Tree

For Naïve Bayes Gaussian Classifier

1) **Confusion Matrix:**

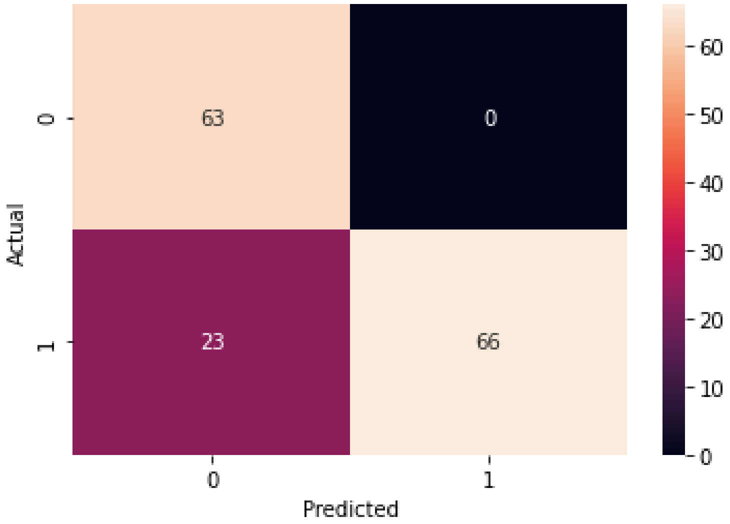


Figure 4.10- CM of NB

2) **ROC Curve:**

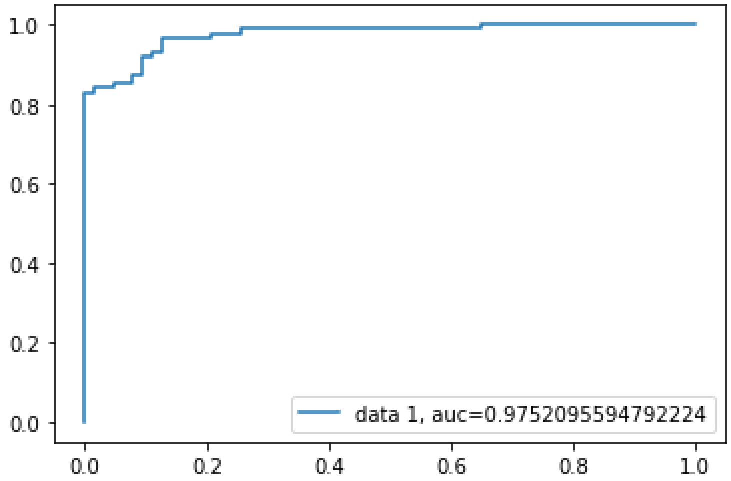


Figure 4.11- ROC of NB

For KNN Classifier

1) **Confusion Matrix:**

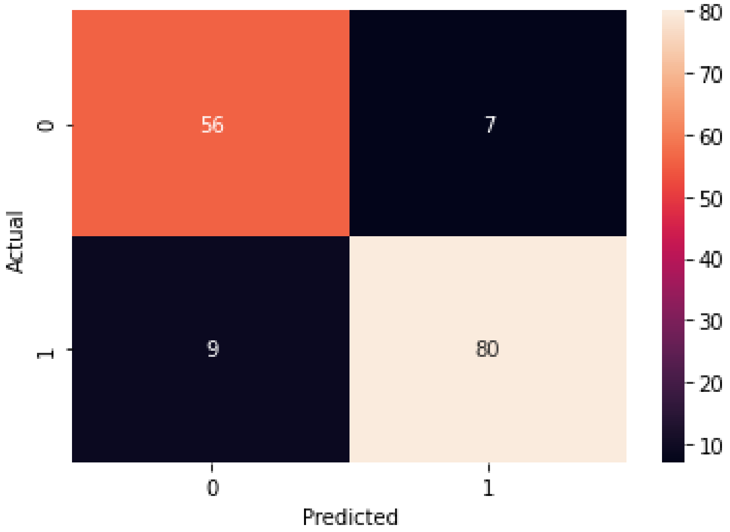


Figure 4.12- CM of KNN

2) **ROC Curve:**

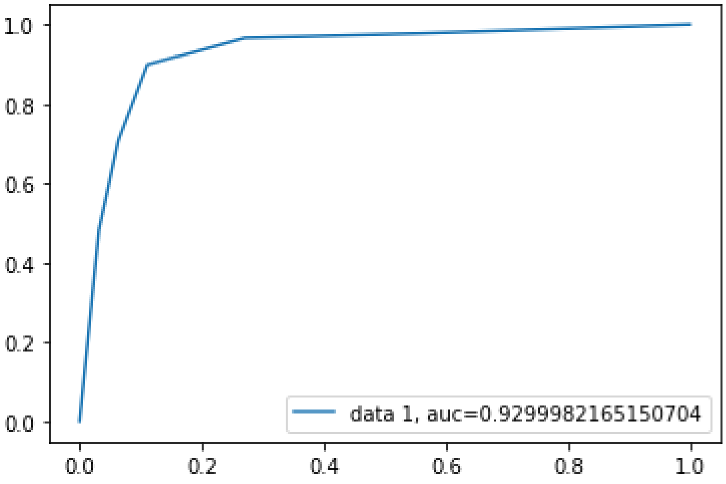


Figure 4.13- ROC of KNN

For SVM Classifier

1) **Confusion Matrix:**

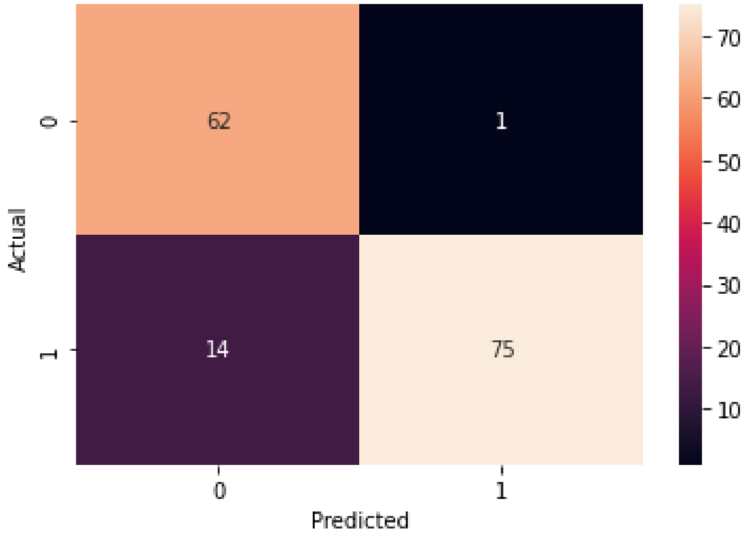


Figure 4.14- CM of SVM

2) **ROC Curve:**

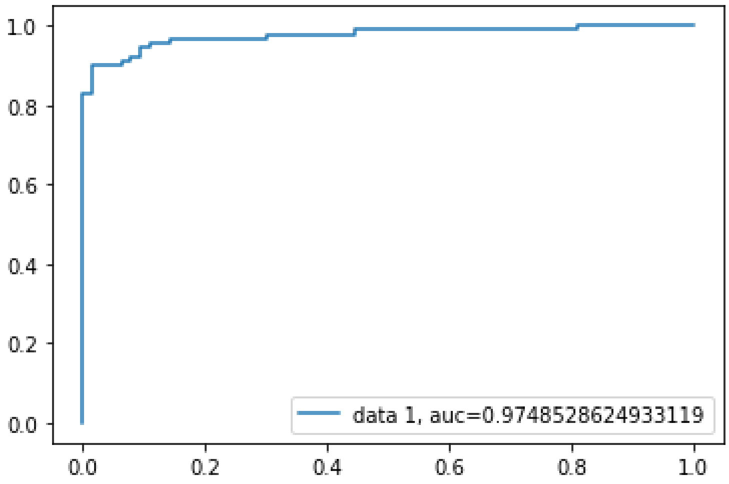


Figure 4.15- ROC of SVM

For ADA Boost Classifier

1) **Confusion Matrix:**

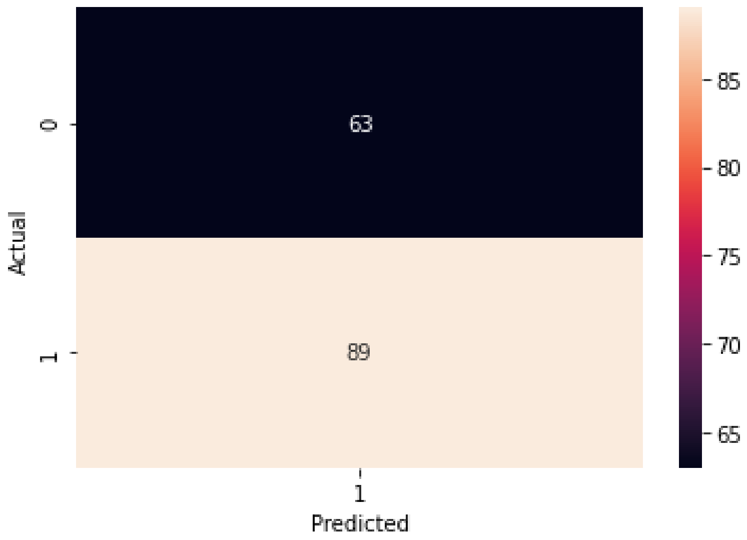


Figure 4.16- CM of AB

2) **ROC Curve:**

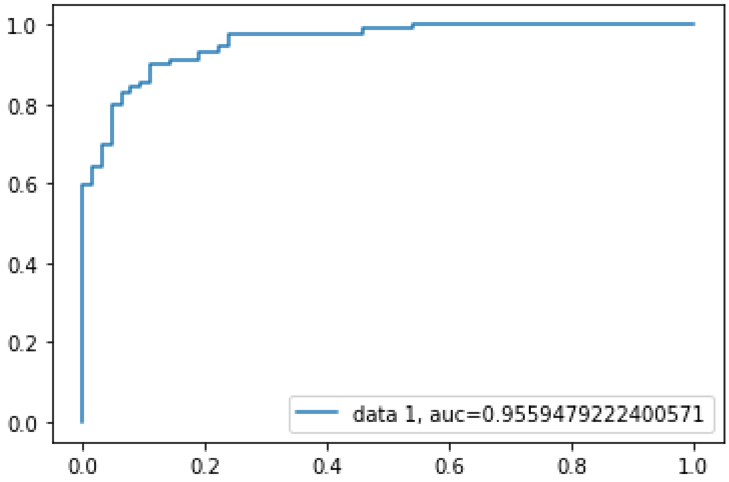


Figure 4.17- ROC of AB

For Gradient Boost Classifier

1) **Confusion Matrix:**

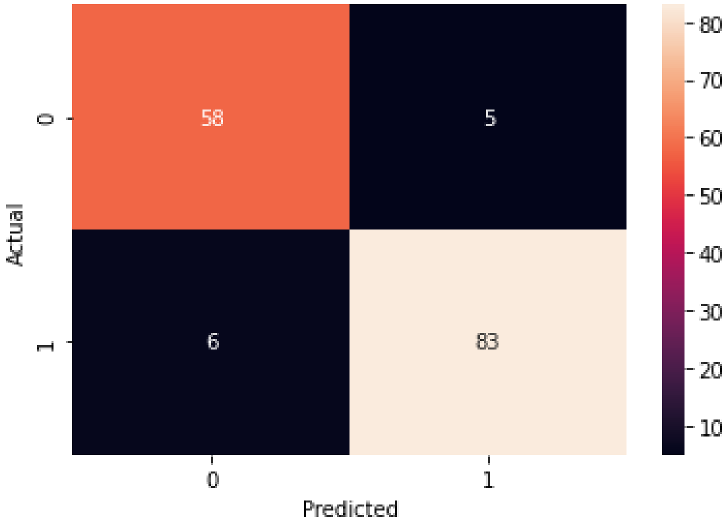


Figure 4.18- CM of GB

2) **ROC Curve:**

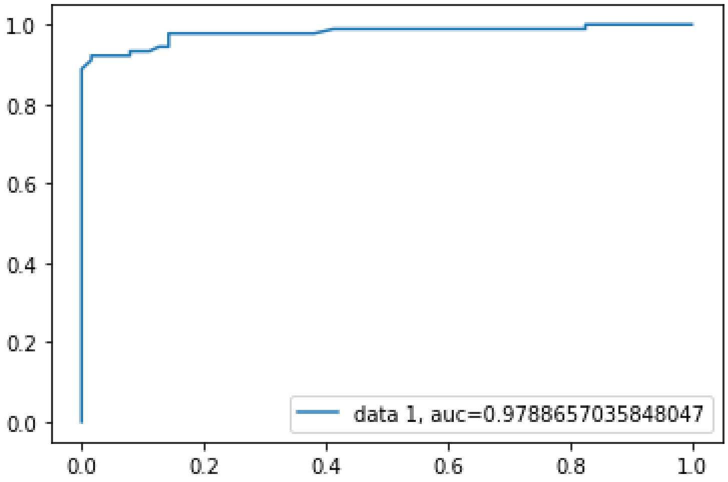


Figure 4.19- ROC of GB

For XG Boost Classifier

1) **Confusion Matrix:**

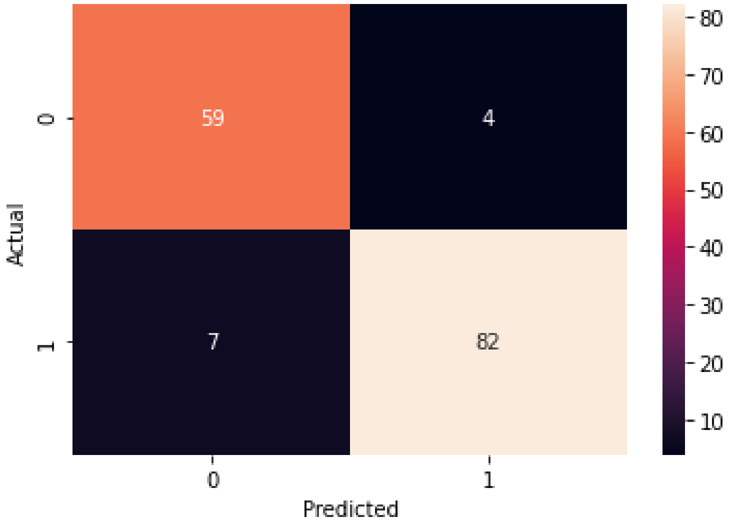


Figure 4.20- CM of XGB

2) **ROC Curve:**

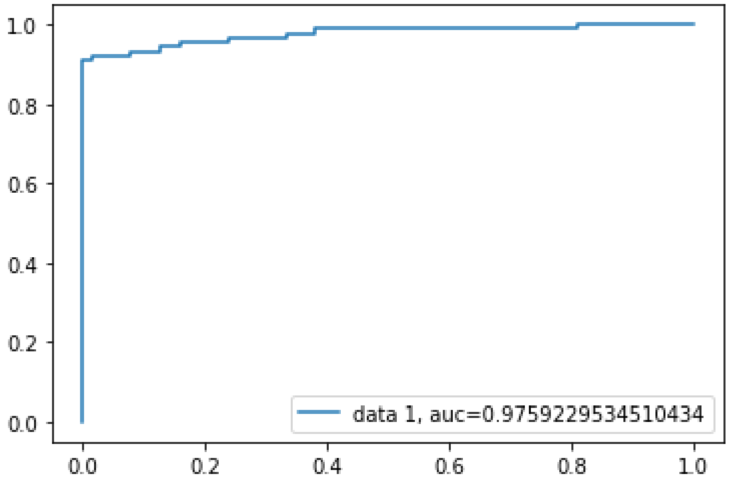


Figure 4.21- ROC of XGB

For Fisher’s LDA Classifier

1) **Confusion Matrix:**

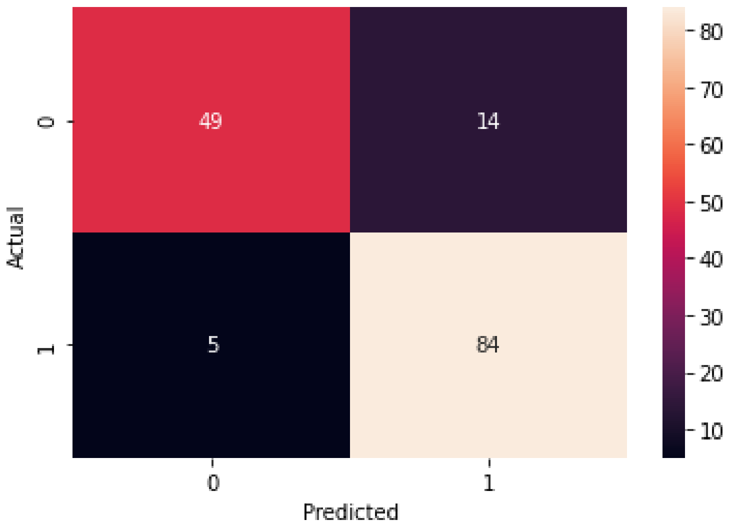


Figure 4.22- CM of LDA

2) **ROC Curve:**

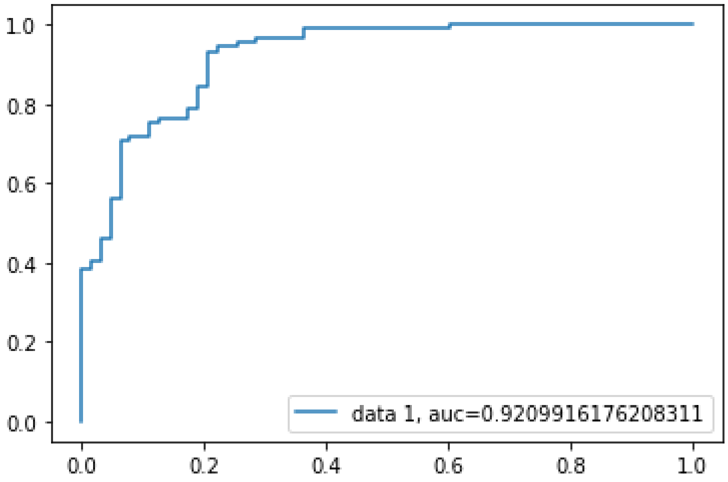


Figure 4.23- ROC of LDA

For ANN Classifier

1) **Confusion Matrix:**

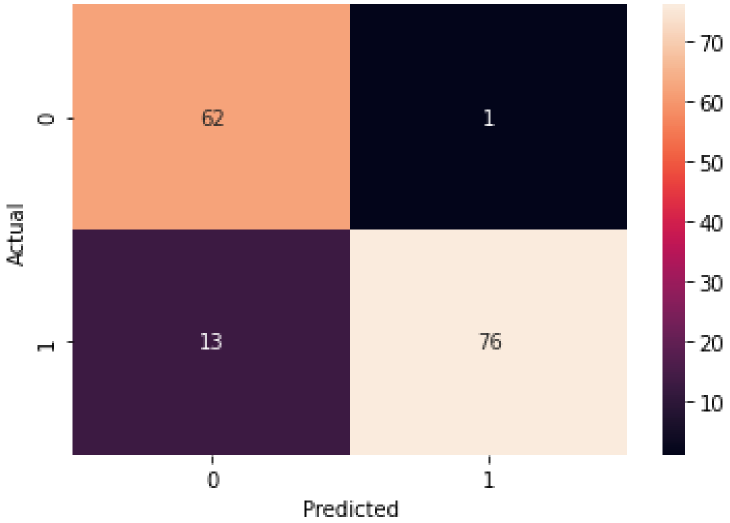


Figure 4.24- CM of ANN

2) **ROC Curve:**

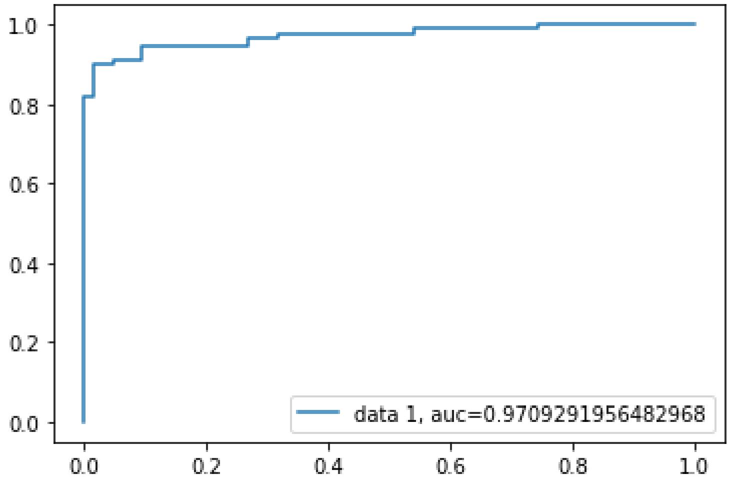


Figure 4.25- ROC of ANN

4.3 Descriptive Analysis

Taking a look at all eleven methodology we finally got the ideal techniques which satisfied the best and those are Decision Tree and Random Forest classifier having 96% accuracy. A ROC curve is a graph showing the performance of a classification model at all classification thresholds. AUC in ROC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0[19]. The data used in this research fuses only two classes for the yield variable, present and not present.

4.4 Summary

Resulting to getting this accuracy, most raised result came from Decision Tree and Random Forest that are the explanation, we are satisfied; in case we endeavor to extend accuracy level, must to set up the dataset properly. The all supreme signs should be also numbered. At that, to extend the precision level, data cleaning has no other choice. The more data are preprocessed, the more precise assumption will be showed up by this classifier.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

This project will significantly influence society. People from rural area can use the application executed through this research to check if they have any urinary tract infection. This research is a present for the typical poor individuals for whom it isn't generally easy to take consultation from a kidney specialist in view regarding the nonappearance of cash and the measure of specialists open. Moreover, people can check the kidney problem in a few seconds which normally sets aside a reasonable effort to check truly. So, this research has a helpful outcome on society.

5.2 Impact on Environment

Any work has a pretty much environmental effect. This project positively affects the environment. There is no opportunity of sending any harmful substances or anything that harms the atmosphere. Also, this project will help people to predict their sickness without a zero loss to the environment.

5.3 Ethical Aspects

Using a machine learning algorithm for clinical purpose will arise some moral issues. At first the authority of using the dataset. A dataset using from internet obviously raises ethical issues. We have merged a dataset with our main dataset and for this we have used a dataset from UCI repository system. Datasets of UCI repository are free to access to any researchers. There is another issue. A specialist can clarify why and how illness happens and the conceivable result of that sickness. Yet, how a machine distinguishes that sickness is totally obscure to the users. Thusly, it might raise an issue about the precision of the end. A probable reaction for the issue is conveying the program under open-source licenses.

5.4 Sustainability Plan

This research means to decrease the sufferings of the average citizens. To guarantee its maintainability greater progress will be brought to the research. Execution of this research at this point will not be restricted to web applications just, rather an android form, as well as an iOS rendition programming, will be executed. The current website page will be refreshed to draw in more individuals and there will be an appropriate rule to the clients so that they realize how to utilize it. More calculations will be applied later on trying to increment the performance of the model.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary

Our aim was to develop a model which will predict kidney disease. We have collected data from hospital and some data from UCI repository system. We pre-processed data, cleaned data, removed missing value and found out the features that are mostly responsible for kidney disease. We have applied eleven machine learning algorithm and among these, Decision Tree and Random Forest classifier gave us 96% accuracy.

6.2 Conclusion

From collecting data to model training, we have learned a lot of things on Supervised learning. Also, working with machine learning algorithms will help us to develop many more predictions in future. Study on supervised learning also help us to know about unsupervised learning and we are all interested to work with different techniques and model in future.

6.3 Recommendation

One notable recommendation is:

- To produce better output of this research, we can prepare the data set more efficiently.

6.4 Future Work

- We will add more data or more attributes to make the model more precise.
- We will implement our work by creating android application and web application to predict new data.

REFERENCES

[1] Chimwayi, K. B., Haris, N., Caytiles, R. D., & Iyengar, N. C. S. (2017). Risk level prediction of chronic kidney disease using Neuro-fuzzy and hierarchical clustering algorithm (s), *International Journal of Multimedia and Ubiquitous Engineering*, vol.12, pp.23-36, 2017

[2] alias Rohini, S. G., & Karpagam, C. A PRIMARY PREDICTIVE STUDY ON PREVALENCE OF KIDNEY DISEASES IN WOMEN USING MACHINE LEARNING TECHNIQUES, vol.16, pp. 143-148, June 2020

[3] Celik, E., Atalay, M., & Kondiloglu, A. (2016). The diagnosis and estimate of chronic kidney disease using the machine learning methods, *International Journal of Intelligent Systems and Applications in Engineering*, vol.4, pp. 27-31, September 2016

[4] Lakshmi, K. R., Nagesh, Y., & Krishna, M. V. (2014). Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology*, vol. 7, pp. 242-254, March 2014

[5] Van Eyck, J., Ramon, J., Guiza, F., Meyfroidt, G., Bruynooghe, M., & Van den Berghe, G. (2012). Data mining techniques for predicting acute kidney injury after elective cardiac surgery. *Critical Care*, *International Journal of Intelligent Systems and Applications in Engineering*, March 2012

[6] Rezapour, M., Khavanin Zadeh, M., & Sepehri, M. M. (2013). Implementation of predictive data mining techniques for identifying risk factors of early AVF failure in hemodialysis patients. *Computational and mathematical methods in medicine,* March 2013

[7] Al-Hyari, A. Y., Al-Taee, A. M., & Al-Taee, M. A. (2013, December). Clinical decision support system for diagnosis and management of chronic renal failure. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),* Jordan, vol.9, 2013

[8] Bala, S., & Kumar, K. (2014). A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing*, vol. *3*(7), pp. 960-967, February 2012

[9] Kumar, K., & Abhishek, B. (2012). Artificial neural networks for diagnosis of kidney stones disease, Germany: GRIN Verlag, I.J. *Information Technology and Computer Science*, vol.1, pp.41-48, 2009

[10] Bandyopadhyay, S. K., & DUTTA, S. (2020). Chronic Kidney Disease Prediction Using Neural Approach. *medRxiv, International Journal of Multimedia and Ubiquitous Engineering*, 2020

[11] Vijayarani, S., & Dhayanand, S. (2015). Data mining classification algorithms for kidney disease prediction. *International Journal on Cybernetics & Informatics (IJCI),* vol. 4, pp. 13-25, August 2015

[12] Belur Nagaraj, S., Pena, M. J., Ju, W., Heerspink, H. L., & BEAt-DKD Consortium. (2020). Machine-learning–based early prediction of end-stage renal disease in patients with diabetic kidney disease using clinical trials data. *Diabetes, Obesity and Metabolism*, *22*(12), 2479-2486.

[13] Mahalakshmi, S., Menaka, P., & Rajkumar, R. S. (2019). Classification of chronic kidney disease stages

in diabetic patients. *International Journal of Research and Analytical Reviews*, vol.6, March 2019.

[14] Tabassum, S., MBB, G., & Majumdar, J. (2017). Analysis and prediction of chronic kidney disease using data mining techniques. *Int. J. Eng. Res. Comput. Sci. Eng*, *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE),* vol. 4, pp. 25-32, September 2017

[15] Gharibdousti, M. S., Azimi, K., Hathikal, S., & Won, D. H. (2017). Prediction of chronic kidney disease using data mining techniques. In *IIE Annual Conference.* Institute of Industrial and Systems Engineers (IISE), 2017

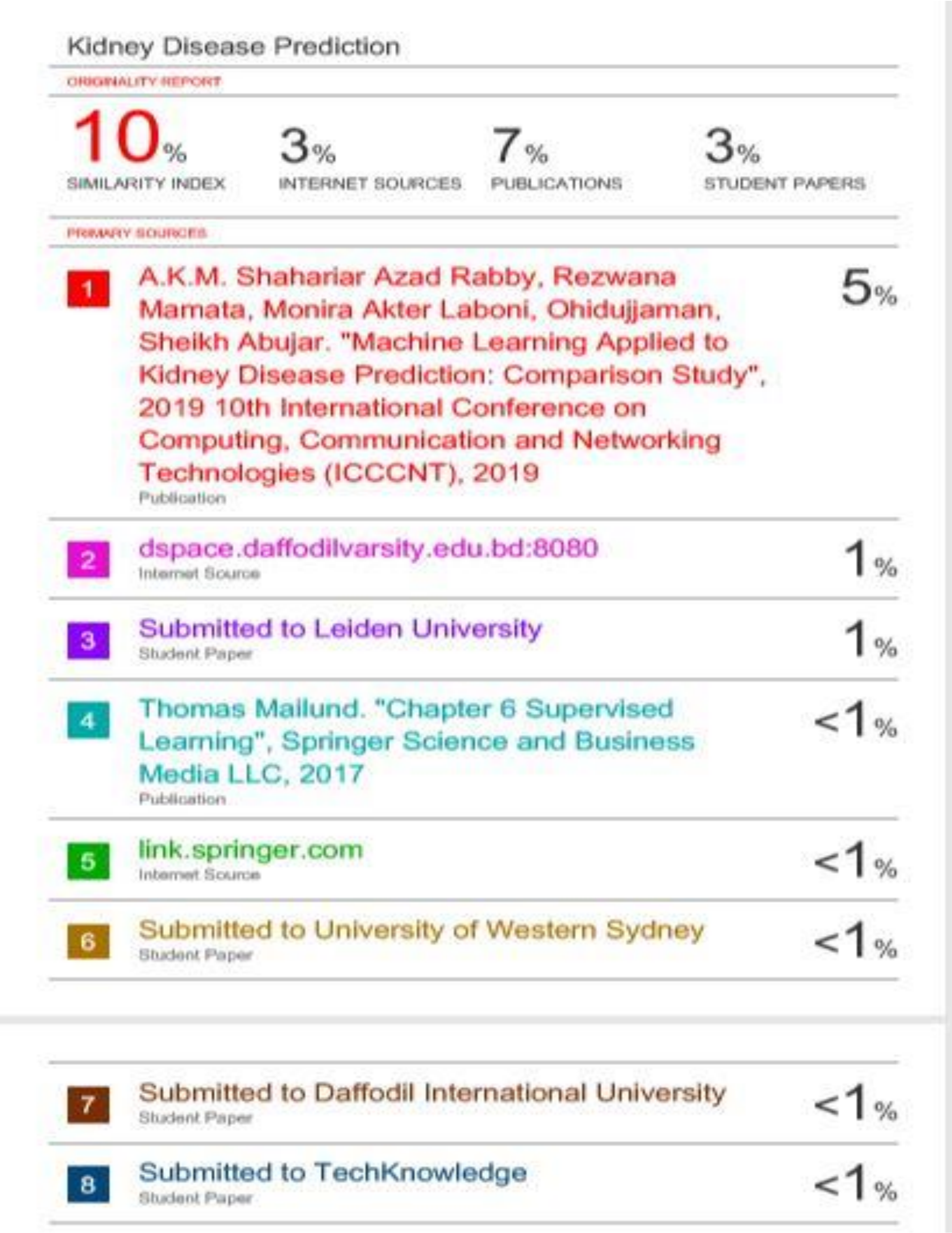
[16] Yafeng Ren, Hao Fei, Xiaohui Liang, Donghong Ji, Ming Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records", China Health Information Processing Conference, China, pp. 132-138, 2018

[17] Ren, Y., Fei, H., Liang, X., Ji, D., & Cheng, M. (2019). A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. *BMC medical informatics and decision making*, 2019

[18] 7 Things to know about kidney function, available at <<www.kidney.org/kidneydisease/howkidneyswrk#:~:text=Why%20Are%20the%20Kidneys%20So,of%20excretion%20and%20re%2Dabsorption>> last accessed on 01 January, 2021 at 02:48 AM

[19] Classification: ROC Curve and AUC | Machine Learning Crash Course, available at <<www.developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=An%20ROC%20curve%20(receiver%20operating,False%20Positive%20Rate>> last accessed on 29 December, 2020 at 08:23 PM

PLAGARISM REPORT



Kidney Disease Prediction

ORIGINALITY REPORT

10%	3%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	A.K.M. Shahariar Azad Rabby, Rezwana Mamata, Monira Akter Laboni, Ohidujjaman, Sheikh Abujar. "Machine Learning Applied to Kidney Disease Prediction: Comparison Study", 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019 Publication	5%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Submitted to Leiden University Student Paper	1%
4	Thomas Mailund. "Chapter 6 Supervised Learning", Springer Science and Business Media LLC, 2017 Publication	<1%
5	link.springer.com Internet Source	<1%
6	Submitted to University of Western Sydney Student Paper	<1%
7	Submitted to Daffodil International University Student Paper	<1%
8	Submitted to TechKnowledge Student Paper	<1%