A Project Report

ON

Prediction of Chronic Kidney Disease using Machine Learning Algorithms

Project report submitted in partial fulfilment of the requirement for the award of the Degree of

Master of Computer Applications (MCA)

By

Aashiq Hussain Kuchey	18045112026
Mohsin Manzoor Badam	18045112032
Amir Nisar	18045112040

Under Supervision of

Mr. Gulnawaz Gani

Department of Computer science

North Campus, University of Kashmir

Delina, Baramullah, Kashmir, 193103

Session: 2018-2022

Department of Computer science

North Campus, University of Kashmir

Delina, Baramullah, Kashmir, 193103

CERTIFICATE

Certified that, **Aashiq Hussain Kuchey, Mohsin Manzoor Badam, Amir Nisar** have Carried out the project work presented in this report entitled ***"Prediction of Chronic Kidney Disease"*** for the award of **Master of Computer Application (MCA)** in the session **2018-2022** from **University of Kashmir** under supervision. The report embodies result of work and studies carried out by student himself and the contents of the report do not form the basis for the award of any other degree to the candidate or to anybody else.

Supervisor

Gulnawaz Gani

Coordinator

Dr. Umar Farooq

Department of Computer Science
University of Kashmir (North Campus)

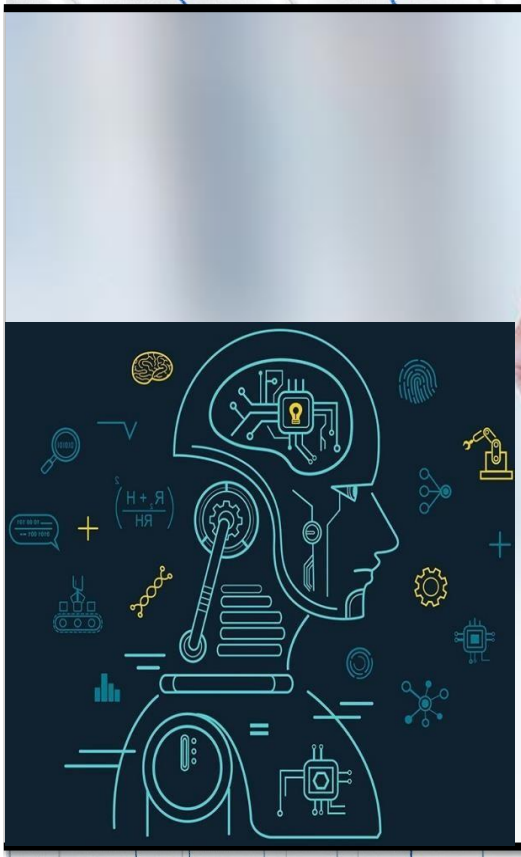
ACKNOWLEDGEMENT

In today's competitive world, there is a race of existence in which those having will to come forward, succeed. With this will we took this project. First of all, we would like to thank the supreme power Almighty Allah who has guided us and bestowed upon us the knowledge required for this project. Without his grace this project could not have become a reality. After Allah we are greatly indebted to our parents who brought us up with love, care, mercy and always encouraged and supported us in every sphere of life.

No words of thanks can sum up the gratitude that we owe to our project supervisor **Gulnawaz Gani, Department of Computer Science, University of Kashmir**, who has always been a key person for us. We thank him for his support, right guidance and valuable suggestions from time to time that helped us in bringing out this project within the stipulated time frame.

We would also like to thank UCI Machine Learning Repository for providing us the dataset for our project.

At last, but not the least we are thankful to all our teachers and friends who have always been helping and encouraging us. We have no valuable words to express our thanks, but our heart is still full of favors received from every person.



PREDICTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING ALGORITHMS

AASHIQ HUSSAIN KUCHEY

MOHSIN MANZOOR BADAM

AMIR NISAR

Table of Contents

ABSTRACT

Human life is precious and fragile and should be treated as such. Death is inevitable but with the advancement in the field of Science and Technology there are numerous ways which can be used to make the quality of life better for every person by the study and analysis of the various parameters of a healthy body. As we all know in today's world one in every 5 persons suffers from some sort of chronic disease. Chronic diseases require ongoing medical attention or limit activities of daily living or both. Sooner the disease is predicted for a person easier his/her life can become by the modification of his/her living-style. For our project we have taken a chronic disease "kidney disease", with the help of machine learning algorithms we will try to predict the various aspects of the above-mentioned disease.

Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Our kidneys filter wastes and excess fluids from our blood, which are then excreted in our urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in our body. In the early stages of chronic kidney disease, we may have few signs or symptoms. Chronic kidney disease may not become apparent until our kidney function is significantly impaired.

1 INTRODUCTION

1.1 Chronic Kidney Disease

Chronic kidney disease, also called chronic kidney failure, describes the gradual loss of kidney function. Kidneys filter wastes and excess fluids from the blood, which are then excreted in urine. When chronic kidney disease reaches an advanced stage, dangerous levels of fluid, electrolytes and wastes can build up in the body. In the early stages of chronic kidney disease, we may have few signs or symptoms. Chronic kidney disease may not become apparent until our kidney function is significantly impaired.

Treatment for chronic kidney disease focuses on slowing the progression of the kidney damage, usually by controlling the underlying cause. Chronic kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant.

1.1.1 Symptoms

Signs and symptoms of chronic kidney disease develop over time if kidney damage progresses slowly. Signs and symptoms of kidney disease may include:

- Nausea
- Vomiting
- Loss of appetite
- Fatigue and weakness
- Sleep problems
- Changes in how much we urinate
- Decreased mental sharpness
- Muscle twitches and cramps
- Swelling of feet and ankles
- Persistent itching
- Chest pain, if fluid builds up around the lining of the heart
- Shortness of breath, if fluid builds up in the lungs
- High blood pressure (hypertension) that's difficult to control

Signs and symptoms of kidney disease are often nonspecific, meaning they can also be caused by other illnesses. Because our kidneys are highly adaptable and able to compensate for lost function, signs and symptoms may not appear until irreversible damage has occurred.

1.1.2 Causes

Chronic kidney disease occurs when a disease or condition impairs kidney function, causing kidney damage to worsen over several months or years.

Diseases and conditions that cause chronic kidney disease include:

- Type 1 or type 2 diabetes
- High blood pressure
- Glomerulonephritis, an inflammation of the kidney's filtering units (glomeruli)
- Interstitial nephritis, an inflammation of the kidney's tubules and surrounding structures
- Polycystic kidney disease
- Prolonged obstruction of the urinary tract, from conditions such as enlarged prostate, kidney stones and some cancers
- Vesicoureteral reflux, a condition that causes urine to back up into our kidneys
- Recurrent kidney infection, also called pyelonephritis

1.1.3 Risk factors

Factors that may increase the risk of chronic kidney disease include:

- Diabetes
- High blood pressure
- Heart and blood vessel (cardiovascular) disease
- Smoking
- Obesity
- Being African-American, Native American or Asian-American
- Family history of kidney disease
- Abnormal kidney structure
- Older age

1.1.4 Stages

Chronic kidney disease (CKD) refers to all five stages of kidney damage, from very mild damage in stage 1 to complete kidney failure in stage 5. The stages of kidney disease are based on how well the kidneys can filter waste and extra fluid out of the blood. In the early stages of kidney disease, our kidneys are still able to filter out waste from our blood. In the later stages, our kidneys must work harder to get rid of waste and may stop working altogether.

The way doctors measure how well our kidneys filter waste from our blood is by the estimated glomerular filtration rate, or eGFR. eGFR is a number based on our blood test for creatinine, a waste product in our blood.

The stages of kidney disease are based on the eGFR number.

eGFR

eGFR is an estimate of how well our kidneys are working. The way eGFR is calculated will be changing. Currently the test considers our age, sex and whether we are African American, among other things. A task force led by the National Kidney Foundation and the American Society of Nephrology is working on recommendations that may remove African American race as a factor in the eGFR calculation. The task force has been seeking the input of stakeholders. AKF advised the task force that eGFR equations should be an unbiased estimate of kidney function. This would make sure that every person will receive appropriate and equitable care. When the NKF-ASN task force makes its recommendations, AKF will promptly review them and then update our educational materials.

Stage 1 CKD

Stage 1 CKD means, the person has mild kidney damage and an eGFR of 90 or greater. Most of the time, an eGFR of 90 or greater means your kidneys are healthy and working well, but you have other signs of kidney damage. Signs of kidney damage could be protein in your urine (pee) or physical damage to your kidneys.

Stage 2 CKD

Stage 2 CKD means, the person has mild kidney damage and an eGFR between 60 and 89. Most of the time, an eGFR between 60 and 89 means the kidneys are healthy and working well. But if the person has Stage 2 kidney disease, this means the person has other signs of kidney damage even though the eGFR is normal. Signs of kidney damage could be protein in the urine or physical damage to the kidneys

Stage 3 CKD

Stage 3 CKD means, the person has an eGFR between 30 and 59. An eGFR between 30 and 59 means that there is some damage to the kidneys and they are not working as well as they should.

Stage 3 is separated into two stages:

- Stage 3a means, the person has an eGFR between 45 and 59.
- Stage 3b means, the person has an eGFR between 30 and 44.

Many people with Stage 3 kidney disease do not have any symptoms. But if there are symptoms, there may be:

- Swelling in the hands and feet.
- Back pain.
- Urinating more or less than normal.

Stage 4 CKD

Stage 4 CKD means, the person has an eGFR between 15 and 29. An eGFR between 15 and 30 means the kidneys are moderately or severely damaged and are not working as they should. Stage 4 kidney disease should be taken very seriously – it is the last stage before kidney failure.

At Stage 4 kidney disease, many people have symptoms such as:

- Swelling in the hands and feet.
- Back pain.
- Urinating (peeing) more or less than normal.

Stage 4 kidney disease, is the time to start talking with the nephrologist about how to prepare for kidney failure. Once the kidneys have failed, the patient will need to start dialysis or have a kidney transplant to live.

Stage 5 CKD

Stage 5 CKD means, the person has an eGFR less than 15. An eGFR less than 15 means the kidneys are getting very close to failure or have completely failed. If the kidneys fail, waste builds up in the blood, which makes a person very sick.

Some of the symptoms of kidney failure are:

- Itching
- Muscle cramps

- Feeling sick and throwing up
- Not feeling hungry
- Swelling in your hands and feet
- Back pain
- Urinating (peeing) more or less than normal
- Trouble breathing
- Trouble sleeping

Once the kidneys have failed, the patient will need to start dialysis or have a kidney transplant to live.

1.2 Objective/Aim

The aim of this project was to build a model which would predict whether a patient would develop Chronic Kidney Disease or not based on his/her medical data. We did the same by implementing machine learning algorithms on the dataset. We implemented more than one algorithm on the dataset to find the best suitable algorithm based on the accuracy score.

1.3 Methodologies

Chronic kidney disease may not become apparent until the kidney function is significantly impaired. Chronic kidney disease can progress to end-stage kidney failure, which is fatal without artificial filtering (dialysis) or a kidney transplant. In today's world one in every 5 persons suffers from some sort of chronic disease. Chronic diseases require ongoing medical attention or limit activities of daily living or both.

Sooner this disease is predicted for a person easier his/her life can become by the modification of his/her living-style.

The latest advances in Machine Learning technologies can be applied for obtaining hidden patterns, which may diagnose Chronic Kidney Disease at an early phase. Our project presents a methodology for the prediction of Chronic Kidney Disease using adverse Machine Learning Algorithm(s).

We implemented the following Classification Algorithms on our dataset:

1. Decision Tree Classifier
2. K-Nearest Neighbors Classifier
3. Support Vector Machine Classifier
4. Random Forest Classifier
5. Logistic Regression
6. Ada Boost Classifier
7. Gaussian Naïve Bayes Classifier

1.4 Requirements

Hardware Requirements

- Laptop/Desktop/Smart Phone
- GB or more RAM
- 32 GB or more HDD

Software Requirements

1. OS(Windows, Linux, MAC OS, Android, IOS etc)
2. Python 3 and libraries:
    i. NumPy
    ii. Pandas
    iii. Matplotlib
    iv. Scikit-learn
    v. Seaborn
3. Google chrome
4. Google Colab

2 TOOLS USED

2.1 NumPy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

We used NumPy while plotting heatmap.

2.2 Pandas

Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions.

We used Pandas while:

a. Reading/Writing the data frame.

b. Changing datatypes of attributes.

c. Feature Selection.

2.3 Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications. A Python matplotlib script is structured so that a few lines of code are all that is required in most instances to generate a visual data plot.

We used Matplotlib while plotting heatmap and barplot.

2.4 Sklearn

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib.

2.5 Seaborn

Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data. One has to be familiar with Numpy and Matplotlib and Pandas to learn about Seaborn.

We used Seaborn while plotting heatmap and barplot.

2.6 Graphviz

Graphviz is open-source graph visualization software. Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks. It has important applications in networking,

bioinformatics, software engineering, database and web design, machine learning, and in visual interfaces for other technical domains.

We used graphviz while generating decision tree.

2.7 IO module

The io module provides Python's main facilities for dealing with various types of I/O. There are three main types of I/O: text I/O, binary I/O and raw I/O. These are generic categories, and various backing stores can be used for each of them. A concrete object belonging to any of these categories is called a file object. Other common terms are stream and file-like object.

Independent of its category, each concrete stream object will also have various capabilities: it can be read-only, write-only, or read-write. It can also allow arbitrary random access (seeking forwards or backwards to any location), or only sequential access (for example in the case of a socket or pipe).

We used IO module while reading the dataset.

3 THE DATASET

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
5	5	60.0	90.0	1.015	3.0	0.0	NaN	NaN	notpresent	notpresent	...	39	7800	4.4	yes	yes	no	good	yes	no	ckd
6	6	68.0	70.0	1.010	0.0	0.0	NaN	normal	notpresent	notpresent	...	36	NaN	NaN	no	no	no	good	no	no	ckd
7	7	24.0	NaN	1.015	2.0	4.0	normal	abnormal	notpresent	notpresent	...	44	6900	5	no	yes	no	good	yes	no	ckd
8	8	52.0	100.0	1.015	3.0	0.0	normal	abnormal	present	notpresent	...	33	9600	4.0	yes	yes	no	good	no	yes	ckd
9	9	53.0	90.0	1.020	2.0	0.0	abnormal	abnormal	present	notpresent	...	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
10	10	50.0	60.0	1.010	2.0	4.0	NaN	abnormal	present	notpresent	...	28	NaN	NaN	yes	yes	no	good	no	yes	ckd
11	11	63.0	70.0	1.010	3.0	0.0	abnormal	abnormal	present	notpresent	...	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
12	12	68.0	70.0	1.015	3.0	1.0	NaN	normal	present	notpresent	...	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
13	13	68.0	70.0	NaN	NaN	NaN	NaN	NaN	notpresent	notpresent	...	NaN	NaN	NaN	yes	yes	yes	poor	yes	no	ckd
14	14	68.0	80.0	1.010	3.0	2.0	normal	abnormal	present	present	...	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd

3.1 Information about Dataset

The Indians Chronic Kidney Disease (CKD) dataset consists of 400 instances and 24 attributes with 2 classes collected from UCI machine learning repository . The Attribute of this dataset consists of two types of attributes which are numeric and nominal attributes divided into 11 numeric attributes and 14 nominal attributes. This dataset is collected from the patients in Apollo Hospitals Indians. The dataset is divided into two groups, one for training and another for testing. The ratio of training and testing data is 70% and 30% respectively.

Attribute Information :

S. No	Name	Description	Data Type	Data
1.	age	Age	Numerical	Age in Years
2.	bp	Blood Pressure	Numerical	Mm/Hg
3.	sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020,1.025)
4.	al	Albumin	Nominal	(0,1,2,3,4,5)
5.	su	Sugar	Nominal	(0,1,2,3,4,5)
6.	rbc	Red Blood Cells	Nominal	normal, abnormal
7.	pc	Pus Cell	Nominal	normal, abnormal
8.	pcc	Pus Cell Clumps	Nominal	present, notpresent
9.	ba	Bacteria	Nominal	present, notpresent
10.	bgr	Blood Glucose Random	Numerical	mgs/dl
11.	bu	Blood Urea	Numerical	mgs/dl
12.	sc	Serum Creatinine	Numerical	mgs/dl
13.	sod	Sodium	Numerical	mEq/L
14.	pot	Potassium	Numerical	mEq/L

15.	hemo	Hemoglobin	Numerical	Gms
16.	pcv	Packed Cell Volume	Numerical	cells/cumm
17.	wc	White Blood Cell Count	Numerical	cells/cmm
18.	rc	Red Blood Cell Count	Numerical	yes, no
19.	htn	Hypertension	Nominal	yes, no
20.	dm	Diabetes Mellitus	Nominal	yes, no
21.	cad	Coronary Artery Disease	Nominal	yes, no
22.	appet	Appetite	Nominal	good, poor
23.	pe	Pedal Edema	Nominal	yes, no
24.	ane	Anemia	Nominal	yes, no
25.	class	Class	Nominal	ckd, notckd

1. Number of Instances: 400 (250 CKD, 150 notckd)
2. Number of Attributes: 24 + class = 25 (11 numeric, 14 nominal)
3. Class Distribution: (2 classes)

CLASS	NO. OF INSTANCES
CKD	250
NOTCKD	150

3.2 Dataset Attributes

1. Age

Older age is a key predictor of CKD. According to current estimates, CKD is more common in people aged 65 years or older (38%) than in people aged 45–64 years (12%) or 18–44 years (6%). CKD is slightly more common in women (14%) than men (12%). In addition to the natural aging of the kidneys, many conditions that damage the kidneys are more common in older people including diabetes, high blood pressure, and heart disease. According to recent estimates from researchers at Johns Hopkins University, more than 50 percent of seniors over the age of 75 are believed to have kidney disease. The relationship between kidney failure and age varies with age. We can conclude that it is an age-dependent effect. So, it is better to have the right kind of diet for your kidneys with increasing age.

2. Blood Pressure

High blood pressure is one of the major causes of chronic kidney disease.

And kidney disease can also cause high blood pressure. No matter which came first, having high blood pressure damages the tiny blood vessels in the kidneys. The nephrons in the kidneys are supplied with a dense network of blood vessels, and high volumes of blood flow through them. Over time, uncontrolled high blood pressure can cause arteries around the kidneys to narrow, weaken or harden. These

damaged arteries are not able to deliver enough blood to the kidney tissue. If you have high blood pressure, it is important to lower it. It is important to note that diseased kidneys are less able to help regulate blood pressure. As a result, blood pressure increases. If you have CKD, high blood pressure makes it more likely that your kidney disease will get worse and you will have heart problems.

High blood pressure is considered one of the two main causes of chronic kidney disease which is responsible for up to two-thirds of the total cases of CKD.

Relationship:

High blood pressure a chronic kidney disease.

3. Specific Gravity or Urine Specific Gravity (USG)

Formula

$$RD = \frac{\rho_{\text{substance}}}{\rho_{\text{reference}}}$$

RD = relative density

$\rho_{\text{substance}}$ = density of the substance being measured

$\rho_{\text{reference}}$ = Density of the reference

Measurement: Relative density can be calculated directly by measuring the density of a sample and dividing it by the (known) density of the reference substance. The density of the sample is simply its mass divided by its volume. A urine specific gravity test compares the density of urine to the density of water. This quick test can help determine how well your kidneys are diluting your urine. Urine that's too concentrated could mean that your kidneys aren't functioning properly or that you aren't drinking enough water.

Specific gravity is usually 1.010-1.025 (normal range: 1.003-1.030) and highest in the morning. A value >1.025 indicates normal concentrating ability. A value >1.035-

1.040 suggests possible contamination, very high levels of glucose.

4. Albumin

Albumin is a protein, found in animal sources such as meats, milk-products, and eggs. It is also found in plant sources such as beans, nuts, and seeds. Albumin provides the body with the protein needed to both maintain growth and repair tissues. It can also help with fluid removal during the dialysis. Many studies have shown that chronic kidney disease (CKD) patients with a low serum albumin have an increased risk for reaching kidney failure as compared to patients with a normal serum albumin. A damaged kidney lets some albumin pass into the urine. The less albumin in your urine, the better. Sometimes albuminuria is also called proteinuria. A healthy kidney doesn't let albumin pass into the urine.

- **What can cause your Albumin level to drop?**
  - **i.** Inadequate nutrition (not eating enough protein).
  - **ii.** Protein Loss.
  - **iii.** Albumin levels decrease when an inflammation is present.

So to maintain an acceptable level of albumin one must include proteinaceous food items in the diet.

Albuminuria categories in CKD		
Category	**ACR (mg/g)**	**Terms**
A1	< 30	Normal to mildly increased
A2	30-300	Moderately increased*
A3	> 300	Severely increased**
*Relative to young adult level. **ACR 30-300 mg/g for > 3 months indicates CKD.**		
**Including nephrotic syndrome (albumin excretion ACR > 2220 mg/g)		

5. **Sugar**

One cause of kidney failure is diabetes mellitus, a condition characterized by high blood glucose (sugar) levels. Over time, the high levels of sugar in the blood damage the millions of tiny filtering units within each kidney. This eventually leads to kidney failure. Around 20 to 30 per cent of people with diabetes develop kidney disease (diabetic nephropathy), although not all of these will progress to kidney failure. A person with diabetes is susceptible to nephropathy whether they use insulin or not. The risk is related to the length of time the person has diabetes. There is no cure for diabetic nephropathy, and treatment is lifelong. Another name for the condition is diabetic glomerulosclerosis. People with diabetes are also at risk of other kidney problems, including narrowing of the arteries to the kidneys, called renal artery stenosis or renovascular disease.

For people with diabetes, kidney problems are usually picked up during a check- up by their doctor. Occasionally, a person can have type 2 diabetes without knowing it. This means their unchecked high blood sugar levels may be slowly damaging their kidneys. At first, the only sign is high protein levels in the urine. Most people with early kidney damage do not have symptoms. The best way to find early kidney damage is to have a urine test once a year. This test checks for very small amounts of protein in the urine called albuminuria. It helps show kidney damage at an early stage in people with diabetes. Not everyone with kidney disease gets kidney failure. With the right treatment, you can prevent kidney disease from getting worse. It may be years before the kidneys are damaged severely enough to cause symptoms. Some of the symptoms may include:

- Fluid retention (oedema of the legs or face)
- Fatigue
- Headache
- Nausea
- Vomiting.

- **Diagnosis methods**

Diabetic nephropathy is diagnosed using a number of tests including:

- Urine tests - to check protein levels. An abnormally high level of protein in the urine is one of the first signs of diabetic nephropathy.
- Blood pressure - regular checks for raised blood pressure are necessary. Elevated blood pressure is caused by diabetic nephropathy and also contributes to its progression.
- Blood tests - to check the degree of kidney function.
- Biopsy - a small tag of tissue is removed from the kidney, via a slender needle, and examined in a laboratory. This is usually only performed when there is doubt about whether kidney damage is due to diabetes or to another cause.
- Kidney ultrasound - enables the size of the kidneys to be imaged and allows the arteries to the kidneys to be checked for narrowing that can cause decreased kidney function.

➢ **Risk reduction strategies**

A person with diabetes can reduce their risk of diabetic nephropathy, or at least delay its onset, in a number of ways including:

- Strictly controlling blood sugar levels
- Making sure that blood pressure is well controlled
- Avoiding non-steroidal anti-inflammatory drugs (NSAIDS)
- Treating urinary tract infections promptly with antibiotics
- Drinking plenty of non-alcoholic fluids, preferably water
- Avoiding medical treatments that stress the kidneys, such as x-rays requiring the injection of contrast dyes
- Having regular tests to ensure the health of your kidneys.

6. Red Blood Cells

Your kidneys make an important hormone called erythropoietin (EPO). Hormones are chemical messengers that travel to tissues and organs to help you stay healthy. EPO tells your body to make red blood cells. When you have kidney disease, your kidneys cannot make enough EPO. Low EPO levels cause your red blood cell count to drop and anemia to develop. Most people with kidney disease will develop anemia. Anemia can happen early in the course of kidney disease and grow worse as kidneys fail and can no longer make EPO.

Anemia is especially common if you:

- Have diabetes

- Are African-American/Black
- Have moderate or severe loss of kidney function (CKD stage 3 or 4)
- Have kidney failure (stage 5)
- Are female.

7. **Pus cells**

White blood cells (pus cells) are signs of infection. Presence of pus cells in urine may indicate the presence of urinary tract infection (UTI). A renal abscess can be caused by bacteria from an infection that's gotten to the kidneys. The bacteria can travel through the blood or in urine backing up into the kidney. In the kidney, the bacteria can spread to the kidney tissue. A renal abscess is not a common disease. A kidney abscess, once discovered, is usually treated with direct intervention. Typically, pus from the abscess is drained through a catheter inserted percutaneously (through the skin) or surgically implanted. Intravenous antibiotics are usually administered to clear the infection.

8. **Pus Cell Clumps**

Presence of puss cell clumps indicates that there is an inflammation or bacterial infection of the kidney and urinary tract. And the absence of puss cell clumps indicates that there is no bacterial infection in the kidney.

9. **Bacteria**

A Bacteria called Escherichia Coli (E Coli) causes about 90 percent of kidney infections. The bacteria migrate from the genitals through the urethra (the tube that removes urine from the body) into the bladder and up the tubes (ureters) that connect the bladder to the kidneys.

Most common ways by which bacterial infections cause renal dysfunction is AKI (acute kidney injury), which occurs as part of multi-organ dysfunction due to sepsis, SIRS, hypotension, hemolysis, or hepatorenal syndrome. Direct invasion of renal tissue by various bacteria either through ascending infection or, through hematogenous spread leads to urinary tract infection, which may lead to renal dysfunction in form of pyelonephritis. Occasionally, renal injury may be due to nephrotoxic effects of various antimicrobial agents used as a part of management. Elderly, diabetic, pregnancy, and immunocompromised patients are at increased risk of acute injury and carry high mortality and morbidity. As many as 20% of critically ill patient have irreversible renal damage due to acute cortical necrosis and another 40% have incomplete renal recovery, leading to CKD . It is not uncommon to observe that episodes of bacterial sepsis accelerate the rate of progression of pre-existing CKD by multiple mechanisms.

10. Blood Glucose

Hyperglycemia is a problem for people with diabetes, and it poses a significant health risk when you have chronic kidney disease (CKD). If your diabetes is not controlled, it can lead to increased loss of kidney function, cardiovascular disease, vision loss and other complications.

Over time, the high levels of sugar in the blood damage the millions of tiny filtering units within each kidney. This eventually leads to kidney failure. Around 20 to 30 per cent of people with diabetes develop kidney disease (diabetic nephropathy), although not all of these will progress to kidney failure. Kidney damage caused by diabetes usually occurs slowly, over many years. You can take steps to protect your kidneys and to prevent or delay kidney damage.

MEAN BLOOD GLUCOSE (FASTING)				
Level	mg/dL	mmol/L	Risk	Suggested action
Dangerous high	315+	11	Very high	Seek immediate medical attention
High	280	15.6	High	Seek medical attention
High	250	13.7	High	Seek medical attention
High	198	11	High	Seek medical attention
Borderline	180	10	Medium	Consult your doctor
Borderline	150	8.2	Medium	Consult your doctor
Borderline	120	7	Medium	Consult your doctor
Normal	108	6	No risk	No action needed
Normal	72	4	No risk	No action needed
Low	70	3.9	Medium	Consult your doctor
Dangerously Low	50	2.8	High	Seek medical Attention

vitamonk

Normal levels of blood sugar for *non-diabetics* range from 70-130 mg throughout the day. They are at their lowest (70-90 mg) in the morning and before meals, and at their highest about an hour after meals.

11. Blood Urea Or Blood Urea Nitrogen (BUN)

Urea is a waste product of metabolism that is excreted by the kidneys in urine. Kidney disease is associated with reduced urea excretion and consequent rise in blood concentration.

A BUN, or blood urea nitrogen test, can provide important information about your kidney function. The main job of your kidneys is to remove waste and extra fluid from your body. If you have kidney disease, this waste material can build up in your blood and may lead to serious health problems, including high blood pressure, anemia, and heart disease.

The test measures the amount of urea nitrogen in your blood. Urea nitrogen is one of the waste products removed from your blood by your kidneys. Higher than normal BUN levels may be a sign that your kidneys aren't working efficiently.

People with early kidney disease may not have any symptoms. A BUN test can help uncover kidney problems at an early stage when treatment can be more effective.

A BUN test is only one type of measurement of kidney function. If your health care provider suspects you have kidney disease, additional tests may be recommended. These may include a measurement of creatinine, which is another waste product filtered by your kidneys, and a test called a GFR (Glomerular Filtration Rate), which estimates how well your kidneys are filtering blood.

12. Serum Creatinine

Creatinine is a waste product that comes from the normal wear and tear on muscles of the body. Everyone has creatinine in their bloodstream.

Elevated creatinine level signifies impaired kidney function or kidney disease. As the kidneys become impaired for any reason, the creatinine level in the blood will rise due to poor clearance of creatinine by the kidneys. Abnormally high levels of creatinine thus warn of possible malfunction or failure of the kidneys. The serum creatinine level is an insensitive marker of GFR early in the course of CKD. A 33 percent decrease in GFR may raise the creatinine level from 0.8 to only 1.2 mg per dL (70.72 to 106.08 μmol per L). If the prior creatinine level is not known, this decrease in GFR may go unrecognized. When estimated GFR is suspected to be inaccurate—for example, in patients with severe malnutrition or paraplegia—a 24-hour urine collection should be performed to evaluate creatinine clearance.

Looking at how much creatinine is in your blood is not the best way to check your kidney health. That’s because the level of creatinine in your blood is affected by your age, race, gender, and body size. (In other words, what’s considered “normal” depends on these factors.) The best way to know if your kidneys are working properly is by looking at your glomerular filtration rate (GFR).

GFR is a routine lab that can be found on your blood work report.GFR is a calculation that includes your creatinine, along with your age, gender, race, and weight. Your GFR number will help your healthcare provider know if you have kidney disease. You may have kidney disease if your GFR number is:

Below 60 for three months

Above 60 with signs of kidney damage (having protein in the urine is a sign of kidney damage).

Another important test to check kidney function is a urine test. You will be asked to pee into a clean cup called a specimen cup. Only about two tablespoons of urine is needed to do the test. The urine will be sent to a laboratory, where a test called an ACR (albumin-to-creatinine ratio) is done. An ACR shows whether you have a type of protein called albumin in your urine. A normal amount of albumin in your urine is less than 30 mg/g. Anything above 30 mg/g may mean you have kidney disease, even if your GFR number is above 60. This test is also used to look at how likely it is that a person's kidney disease will get worse. This is called risk for progression. Having high amounts of albumin points to a higher risk. The above picture clearly depicts that serum creatinine alone does not help to discover the kidney problems but with GFR we clearly uncover the problems associated with the kidney.

THE SAME SERUM CREATININE: VERY DIFFERENT eGFR			
	 22-YR-OLD BLACK MAN	 58-YR-OLD WHITE MAN	 80-YR-OLD WHITE WOMAN
Serum creatinine	1.2 mg/dL	1.2 mg/dL	1.2 mg/dL
GFR as estimated by the MDRD equation	98 mL/min/1.73 m²	66 mL/min/1.73 m²	46 mL/min/1.73 m²
Kidney function	Normal GFR *or* stage 1 CKD if kidney damage is also present	Stage 2 CKD if kidney damage is also present	Stage 3 CKD

13. Sodium

Sodium is considered as an important mineral in maintaining the balance of body fluid. The concept that salt is a beneficial substance was so ingrained that although the earlier studies on the relationship between low sodium intake and reduction of blood pressure (BP) date back to 1948, it was only after nearly forty years Rthat the international community has recognized the role of salt intake in the pathophysiology of

hypertension. According to the World Health Organization, the restriction of sodium intake to less than 2.3 g/day of sodium corresponding to 5.8 g of salt (or 100 mmol) is one of the most cost-effective measures to improve public health. Cumulating evidence highlights that higher sodium consumption contributes to higher BP, thus increasing the risk of cardiovascular disease (CVD). However, recent studies have raised some concerns about the real benefit of a low salt diet in the healthy general population.

In Chronic Kidney Disease (CKD) patients, high BP is a frequent finding, which is traditionally considered as a direct consequence of sodium sensitivity. Hence, a low salt diet (LSD) is widely considered a cornerstone in the treatment of hypertension in CKD.

Formulas to convert sodium in salt (sodium chloride) and vice versa, according to units of measurement.

grams of sodium = mmol of sodium × 0.023

grams of salt = mmol of sodium × 0.058 (or mmol of sodium/17)

grams of sodium = grams of salt × 0.394

grams of salt = grams of sodium × 2.542

In End Stage Kidney Disease (ESKD) patients, sodium intake can be estimated with a dietary questionnaire, though several factors, such as high dialysate sodium concentration and sodium plasma concentration, can affect thirst and water intake in these patients, irrespective of their sodium intake.

14. Potassium (pot)

Potassium is a mineral that helps our body balance fluids and supports the function of our cells, nerves, and muscles. It is found in varying levels in many foods, especially fruits and vegetables. It is important to have the right balance of potassium in our blood. Levels should generally remain between 3.5 and 5.0 milliequivalents per liter (mEq/L). Getting enough potassium in our diet supports the muscles controlling our heartbeat and breathing.

High levels of potassium in the blood (called hyperkalemia) is unpredictable and can be life-threatening. It can cause serious heart problems and sudden death. There are often no warning signs, meaning a person can have high potassium without knowing it.

Blood potassium >5.0 indicates potassium imbalance. Arbitrary thresholds are used to indicate degree of severity, such as mild (>5.0), moderate (>5.5), and severe (>6.0). Clinical severity is determined by the speed of onset, magnitude of the severity, and the development of clinical findings.

Under normal circumstances, the kidneys are responsible for excreting 90% of the potassium that is consumed daily, with the remaining 10% excreted by feces.

People with chronic kidney disease (CKD) have a high risk for hyperkalemia, due in part to the effects of kidney dysfunction on potassium homeostasis.

A recent review reports hyperkalemia frequency as high as 40-50% in people with chronic kidney disease compared to 2-3% in the general population. CKD patients with the highest risk include those with diabetes, cardiovascular disease, advanced CKD, transplant recipients, and patients taking renin-angiotensin aldosterone system (RAAS) inhibitors. An episode of hyperkalemia in patients with CKD increases the odds of mortality within one day of the event.

15. Hemoglobin (hemo)

Hemoglobin is a protein in our red blood cells that carries oxygen to our body's organs and tissues and transports carbon dioxide from our organs and tissues back to our lungs. If a hemoglobin test reveals that your hemoglobin level is lower than normal, it means we have a low red blood cell count (anemia). Anemia can have many different causes, including vitamin deficiencies, bleeding and chronic diseases.

The risk of developing anemia grows as kidney disease progresses. A person may be at higher risk of anemia if he/she is:

- older than 60
- female
- on dialysis

Other factors may also increase the risk of developing anemia with CKD, including:

- diabetes
- heart disease

- high blood pressure
- kidney failure
- infection
- inflammation
- malnutrition
- blood loss, including from frequent blood draws or dialysis treatment

With anemia, our body isn't making enough red blood cells. If a person have anemia and CKD, the red blood cells may also have a shorter lifespan than usual. They can die off faster than the body can replace them.

16. Packed Cell Volume (pcv)

The hematocrit test, also known as a packed-cell volume (PCV) test, is a simple blood test.

Blood is a mixture of cells and plasma. The packed cell volume (PCV) is a measurement of the proportion of blood that is made up of cells. The value is expressed as a percentage or fraction of cells in blood. For example, a PCV of 40% means that there are 40 milliliters of cells in 100 milliliters of blood.

Red blood cells account for nearly all the cells in the blood. The PCV rises when the number of red blood cells increases or when the total blood volume is reduced, as in dehydration. The PCV falls to less than normal, indicating anaemia, when your body decreases its production of red blood cells or increases its destruction of red blood cells.

Your kidneys make an important hormone called erythropoietin (EPO). Hormones are chemical messengers that travel to tissues and organs to help you stay healthy. EPO tells your body to make red blood cells. When you have kidney disease, your kidneys cannot make enough erythropoietin EPO). Low EPO levels cause your red blood cell count to drop and anemia to develop. Most people with kidney disease will develop anemia. Anemia can happen early in the course of kidney disease and grow worse as kidneys fail and can no longer make EPO.

Most people with kidney disease will develop anemia. Anemia can happen early in the course of kidney disease and grow worse as kidneys fail and can no longer make EPO. Anemia is especially common if you:

- Have diabetes
- Are African-American/Black
- Have moderate or severe loss of kidney function (CKD stage 3 or 4)
- Have kidney failure (stage 5)
- Are female

17. White Blood Cell Count (wc)

Elevated white blood cell (WBC) count is a well-known predictor of chronic kidney disease (CKD) progression. However, elderly patients commonly fail to develop a high WBC count in response to several diseased states and may instead present a low WBC count.

18. Red Blood Cell Count (rc)

Reduced red blood cell (RBC) lifespan has been reported to be a contributory factor to anemia in patients with end-stage chronic kidney disease (CKD), there are limited data regarding RBC lifespan in early-stage of CKD. Serum erythropoietin (EPO) is considered a primary causative factor of renal anemia.

RBC lifespan has been shown to be abnormally short in patients with renal failure. RBC lifespan refers to the duration of time that RBCs survive in circulation after they are released from bone marrow. The commonly referenced normal human adult RBC lifespan of about 120 days was derived from the transfused allogeneic RBC survival time. However, monitoring of RBC lifespan has been complicated by the fact that established RBC lifespan measurement methods with radioactive isotope labeling are cumbersome and time-consuming. Moreover, prior studies of RBC lifespan with such techniques had small sample sizes and RBC lifespan data from patients in early-stage CKD are limited. Hence, the RBC lifespan shortening in association with renal anemia may be underestimated.

19. Hypertension (htn)

Hypertension and CKD are common chronic noncommunicable diseases strictly inter-related with each other; indeed, elevated BP is not only a frequent complication of CKD, but it can also act as the cause of

CKD. A recent meta-analysis showed that hypertensive patients have a 75% greater risk than normotensive individuals of development of de novo CKD (GFR <60 mL/min/1.73 m2), estimating a 10% increase of CKD onset for each increase of 10 mmHg of either BP component. Notably, even pre- hypertension (Systolic BP of 120–139 mm Hg and/or Diastolic BP of 80–89 mm Hg) was associated with a 25% higher risk of developing low GFR.

Furthermore, the prognostic role of lowering BP assumes greater importance in CKD patients if we bear in mind at least three basic points:

Higher prevalence of hypertension in CKD than in the general population, which increases progressively from 65% to 95% as GFR falls from 85 to 15 mL/min/1.73 m2.

Hypertension is the main known risk factor for CKD progression and for CV mortality.

Hypertension is often resistant to the treatment in CKD patients, resulting in worsening CV prognosis.

Salt and water retention play a key role for development of hypertension in CKD. In fact, according to the classical model, under normal conditions, high salt intake temporarily increases plasma sodium level, which is soon buffered by movement of water from the intracellular to the extracellular compartment. Thus, increased plasma sodium concentration also stimulates the thirst center, leading to an increase in water intake and secretion of antidiuretic hormone, which restores plasma sodium concentration to a normal level while increasing and maintaining extracellular fluid volume. On the other hand, high salt intake suppresses the renin-angiotensin-aldosterone system (RAAS), which consequently reduces sodium tubular reabsorption, thus contributing to re-establishing sodium and water homeostasis.

In CKD patients, external sodium balance is preserved by expansion of the extracellular volume (ECV), which however causes the persistence of high BP levels. Therefore, hypertension in CKD is an early manifestation of ECV expansion and, at the same time, a maladaptive mechanism aimed at limiting ECV expansion that corresponds to approximately 5% to 10% of body weight, generally without peripheral edema, when cardiac and hepatic function is normal and the transcapillary Starling forces are not disrupted. In spite of ECV expansion, RAAS is inappropriately activated in CKD, leading to vasoconstriction and sodium retention, which contribute significantly to the raising of BP levels.

20. Diabetes mellitus (dm)

Diabetic kidney disease is a type of kidney disease caused by diabetes. Diabetes is the leading cause of kidney disease. About 1 out of 3 adults with diabetes has kidney disease. The main job of the kidneys is to

filter wastes and extra water out of the blood to make urine. Kidneys also help control blood pressure and make hormones that the body needs to stay healthy. When kidneys are damaged, they can't filter blood like they should, which can cause wastes to build up in the body. Kidney damage can also cause other health problems. Kidney damage caused by diabetes usually occurs slowly, over many years. We can take steps to protect our kidneys and to prevent or delay kidney damage. High blood glucose, also called blood sugar, can damage the blood vessels in the kidneys. When the blood vessels are damaged, they don't work as well. Many people with diabetes also develop high blood pressure, which can also damage the kidneys.

Diabetes is a leading cause of CKD and kidney failure worldwide. In addition to the risk for kidney function decline, patients with diabetes and CKD have high cardiovascular risk.

21. Coronary artery disease (cad)

Chronic kidney disease (CKD) is an independent risk factor for coronary artery disease (CAD). Coronary artery disease is the leading cause of morbidity and mortality in patients with CKD. The outcomes of CAD are poorer in patients with CKD. In addition to traditional risk factors, several uremia-related risk factors such as inflammation, oxidative stress, endothelial dysfunction, coronary artery calcification, hyperhomocysteinemia, and immunosuppressants have been associated with accelerated atherosclerosis. A number of uremia-related biomarkers are identified as predictors of cardiac outcomes in CKD patients. The symptoms of CAD may not be typical in patients with CKD. Both dobutamine stress echocardiography and radionuclide myocardial perfusion imaging have moderate sensitivity and specificity in detecting obstructive CAD in CKD patients.

22. Appetite (appet)

The progressive decline of glomerular filtration rate in chronic kidney disease patients is associated with a significant reduction in food intake. Approximately one third of chronic dialysis patients complain of a fair or poor appetite and this is related directly to poor patient outcomes. Appetite regulation involves the gastrointestinal tract (ghrelin as an appetite stimulant, and cholecystokinin, glucagon-like peptide-1, and neuropeptide YY as appetite inhibitors); the adipose tissue with leptin, a potent appetite inhibitor; the vagal system; and the brain, which integrates the stimuli in the hypothalamus area. Satiety relies on the melanocortin receptors with serotonin as the main neurotransmitter and is challenged with hunger peptides, namely, neuropeptide Y and agouti-related peptide. In nondialyzed chronic renal failure patients and in maintenance dialysis patients, anorexia is related mainly to the accumulation of unidentified anorexigenic compounds, inflammatory cytokines, and alterations in appetite regulation, such as amino acid imbalance,

which increases the transport of free tryptophan across the blood-brain barrier. This creates a hyperserotoninergic state that is prone to low appetite. Treatment of anorexia involves counseling, starting dialysis treatments in uremic chronic kidney disease patients, increasing the dialysis dose.

23. Pedal Edema (pe)

Bilateral pedal edema is the most common symptom in chronic kidney disease patients. It occurs due to the loss of functioning of the kidney. This may lead to fluid accumulation in the body and also an accumulation of excretory products or waste products like creatinine, uric acid, urea levels are increases in blood.

24. Anemia (ane)

Anemia is a common complication of chronic kidney disease (CKD). CKD patients suffer from both absolute and functional iron deficiency. Absolute iron deficiency is defined by severely reduced or absent iron stores, while functional iron deficiency is defined by adequate iron stores but insufficient iron availability for incorporation into erythroid precursors. This is due to increased levels of hepcidin. Anemia in CKD is associated with an increased risk of morbidity and mortality. The association between anemia and mortality may be related to the severity of anemia. All CKD patients should be screened for anemia during the initial evaluation for CKD. Criteria used to define iron deficiency are different among CKD compared to normal renal function. Among CKD patients, absolute iron deficiency is defined when the transferrin saturation (TSAT) is

≤20% and the serum ferritin concentration is ≤100 ng/mL among predialysis and peritoneal dialysis patients or ≤200 ng/mL among hemodialysis patients. Functional iron deficiency, also known as iron-restricted erythropoiesis, is characterized by TSAT ≤20% and elevated ferritin levels. Iron supplementation is recommended for all CKD patients with anemia. There is general agreement according to guidelines that intravenous (i.v.) iron supplementation is the preferred method for CKD patients on dialysis (CKD stage 5D) and either i.v. or oral iron is recommended for patients with CKD ND (CKD stages 3– 5). Anemia in CKD has been shown to be associated with an increased risk of morbidity and mortality.

4 MACHINE LEARNING TECHNIQUES

4.1 Decision Tree

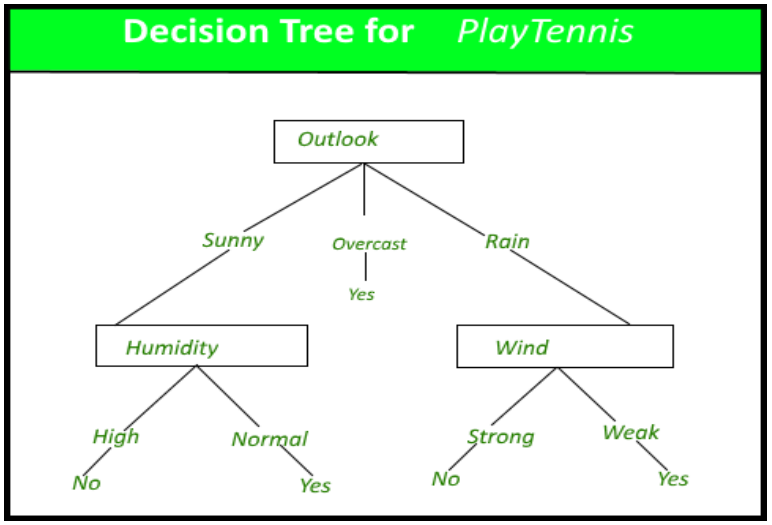
Decision Tree: Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

- **Construction of Decision Tree:**

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a

node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

- **Decision Tree Representation:**



Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.

The decision tree in above figure classifies a particular morning according to whether it is suitable for playing tennis and returning the classification associated with the particular leaf.(in this case Yes or No).

In other words we can say that decision tree represent a disjunction of conjunctions of constraints on the attribute values of instances.

Strengths and Weakness of Decision Tree approach

The strengths of decision tree methods are:

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

The weaknesses of decision tree methods:

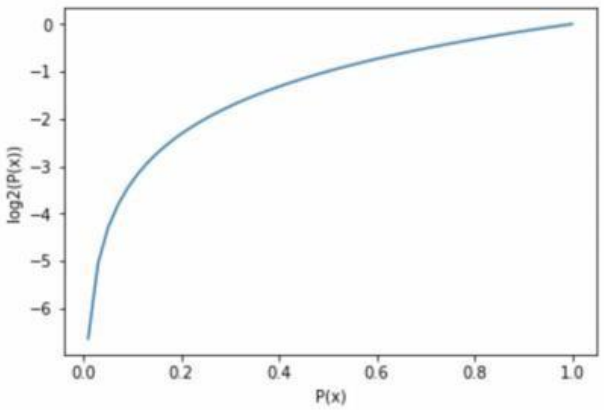
- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
- Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

4.1.1. Entropy

It is used to measure the impurity or randomness of a dataset. Imagine choosing a yellow ball from a box of just yellow balls (say 100 yellow balls). Then this box is said to have 0 entropy which implies 0 impurity or total purity.

Now, let's say 30 of these balls are replaced by red and 20 by blue. If we now draw another ball from the box, the probability of drawing a yellow ball will drop from 1.0 to

0.5. Since the impurity has increased, entropy has also increased while purity has decreased. Shannon's entropy model uses the logarithm function with base 2 (log2(P(x)) to measure the entropy because as the probability P(x) of randomly drawing a yellow ball increase, the result approaches closer to binary logarithm 1 as shown in the graph below.



When a target feature contains more than one type of element (balls of different colours in a box), it is useful to sum up the entropies of each possible target value and weigh it by the probability of getting these values assuming a random draw. This finally leads us to the formal definition of Shannon's entropy which serves as the baseline for the information gain calculation:

$$Entropy(x) = -\sum (P(x=k) * log_2(P(x=k))$$

Where P(x=k) is the probability that a target feature takes a specific value, k. Logarithm of fractions gives a negative value and hence a '-' sign is used in entropy formula to negate these negative values. The maximum value for entropy depends on the number of classes. For n classes, maximum entropy is log2n.

4.1.2. Information Gain

To find the best feature which serves as a root node in terms of information gain, we first use each descriptive feature and split the dataset along the values of these descriptive features and then calculate the entropy of the dataset. This gives us the remaining entropy once we have split the dataset along the feature values. Then, we subtract this value from the originally calculated entropy of the dataset to see how much this feature splitting reduces the original entropy which gives the information gain of a feature and is calculated as:

Information Gain(feature) = Entropy (Dataset) – Entropy (feature)

The feature with the largest information gain should be used as the root node to start building the decision tree. ID3 algorithm uses information gain for constructing the decision tree.

4.1.3. Gini Index

It is calculated by subtracting the sum of squared probabilities of each class from one. It favors larger partitions and easy to implement whereas information gain favors smaller partitions with distinct values.

$$Gini\ Index = 1 - \sum (P(x=k))^2$$

A feature with a lower Gini index is chosen for a split. The classic CART algorithm uses the Gini Index for constructing the decision tree.

4.1.4. Pruning

Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning processes can be divided into two types (pre- and post-pruning).

Pre-pruning procedures prevent a complete induction of the training set by replacing a stop() criterion in the induction algorithm (e.g., max. Tree depth or information gain (Attr) > minGain). Pre-pruning methods are considered to be more efficient because they do not induce an entire set, but rather trees remain small from the start. Pre-pruning methods share a common problem, the horizon effect. This is to be understood as the undesired premature termination of the induction by the stop () criterion.

Post-pruning (or just pruning) is the most common way of simplifying trees. Here, nodes and subtrees are replaced with leaves to reduce complexity. Pruning can not only significantly reduce the size but also improve the classification accuracy of unseen objects. It may be the case that the accuracy of the assignment on the train set deteriorates, but the accuracy of the classification properties of the tree increases overall.

The procedures are differentiated on the basis of their approach in the tree (top- down or bottom-up).

4.2 K-Nearest Neighbours

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

4.2.1. Euclidean Distance

This distance is the most widely used one as it is the default metric that SKlearn library of Python uses for K-Nearest Neighbour. It is a measure of the true straight- line distance between two points in Euclidean space.

It can be used by setting the value of p equal to 2 in Minkowski distance metric.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

4.3. Support Vector Machines

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n- dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM algorithm can be used for Face detection, image classification, text categorization, etc.

- **SVM can be of two types:**

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier

used is called as Non-linear SVM classifier.

4.3.1 Hyper planes

There can be multiple lines/decision boundaries to segregate the classes in n- dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM. The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane. We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

4.3.2 Support Vectors

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

4.4 Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Random forest works on the Bagging principle.

4.4.1 Bagging

Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

4.5 Logistic Regression

Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

$$Logistic\ function = \frac{1}{1+e^{-x}}$$

In the logistic function equation, x is the input variable.

4.6 Naïve Bayes

Naïve Bayes is a classification algorithm that is based on the Bayesian probability theorem. The classifier operates under the fundamental Naive Bayes assumptions which are independent and equal of feature contribution to the outcome, which means that feature presence or absence is unrelated to the presence or absence of any other feature. Naïve Bayes classifies an instance by calculating the probability of its belonging to each target class, where the instance will be considered as belonging to the target class with the highest
probability using the following rule:

$$P(A|B) = \frac{P(B|A).\ P(A)}{P(B)|}$$

In the rule above, A is the target class and B is the features vector describing an instance, P(A|B) is the probability of the instance B belonging to the target class, A. P(A) is the prior probability of the target class in the training set. P(B) is the probability of the features vector given the target class in the training set.

4.7 Ada Boot

In machine learning, boosting originated from the question of whether a set of weak classifiers could be converted to a strong classifier. A weak learner or classifier is a learner who is better than random guessing. This will be robust in over-fitting as in a large set of weak classifiers, each weak classifier being better than random. As a weak classifier, a simple threshold on a single feature is generally used. If the feature is above the threshold than predicted, it belongs to positive otherwise belongs to negative.

AdaBoost stands for 'Adaptive Boosting', which transforms weak learners or predictors to strong predictors in order to solve problems of classification.

4.8. Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. it is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC-ROC curves.

There are four ways to check if the predictions are right or wrong:

i. **TN / True Negative:** the case was negative and predicted negative

ii. **TP / True Positive:** the case was positive and predicted positive

iii. **FN / False Negative:** the case was positive but predicted negative

iv. **FP / False Positive:** the case was negative but predicted positive

4.9 Classification Metrices:

A Classification metrices is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False? More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report. The classification report visualizer displays the precision, recall, F1, and support scores for the model.

4.9.1 Recall

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall:- Fraction of positives that were correctly identified.

Recall = TP/(TP+FN)

4.9.2 Precision

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision:- Accuracy of positive predictions.

Precision = TP/(TP + FP)

4.9.3 F1-score

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their

computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4.9.4 Support:

Support is the number of actual occurrences of the class in the specified dataset.

Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

4.9.5 Macro-Average:

Macro-average is mean average precision/recall/f1 of all classes.

4.9.6 Weighted Average:

Weighted average is the total number of TP(True Positives of all classes)/total number of objects in all classes.

5 DATA CLEANSING AND DATA VISUALIZATION/ANALYSIS

Below is the snapshot of the output which displays the dataset.

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	395	55.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	47	6700	4.9	no	no	no	good	no	no	notckd
396	396	42.0	70.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	54	7800	6.2	no	no	no	good	no	no	notckd
397	397	12.0	80.0	1.020	0.0	0.0	normal	normal	notpresent	notpresent	...	49	6600	5.4	no	no	no	good	no	no	notckd
398	398	17.0	60.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	51	7200	5.9	no	no	no	good	no	no	notckd
399	399	58.0	80.0	1.025	0.0	0.0	normal	normal	notpresent	notpresent	...	53	6800	6.1	no	no	no	good	no	no	notckd

5.1 Deleting Unnecessary Column(s)

```
dataset.columns
```

```
Index(['id', 'age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr',
       'bu', 'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad',
       'appet', 'pe', 'ane', 'class'],
      dtype='object')
```

The column ‘id’ is not required since it is not a calculated value. It can be misleading indecision making; hence we removed this column from the dataset.

```
dataset=dataset.drop(['id'],axis=1)
dataset.columns
```

```
Index(['age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr', 'bu',
       'sc', 'sod', 'pot', 'hemo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad',
       'appet', 'pe', 'ane', 'class'],
      dtype='object')
```

The above image shows the columns/attributes after deleting the column.

5.2 Data Transformation

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 25 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   age     391 non-null    float64
 1   bp      388 non-null    float64
 2   sg      353 non-null    float64
 3   al      354 non-null    float64
 4   su      351 non-null    float64
 5   rbc     248 non-null    object 
 6   pc      335 non-null    object 
 7   pcc     396 non-null    object 
 8   ba      396 non-null    object 
 9   bgr     356 non-null    float64
 10  bu      381 non-null    float64
 11  sc      383 non-null    float64
 12  sod     313 non-null    float64
 13  pot     312 non-null    float64
 14  hemo    348 non-null    float64
 15  pcv     330 non-null    object 
 16  wc      295 non-null    object 
 17  rc      270 non-null    object 
 18  htn     398 non-null    object 
 19  dm      398 non-null    object 
 20  cad     398 non-null    object 
 21  appet   399 non-null    object 
 22  pe      399 non-null    object 
 23  ane     399 non-null    object 
 24  class   400 non-null    object 
dtypes: float64(11), object(14)
memory usage: 78.2+ KB
```

As shown in the above chart, 14 attributes are of object type. Out of these 14 attributes 11 have string values. It is not possible to process non-numeric(string) data, So we converted the string values to numeric values(float64).

5.2.1 Transforming attributes of object type

```
-> rbc  :[nan 'normal' 'abnormal']
-> pc   :['normal' 'abnormal' nan]
-> pcc  :['notpresent' 'present' nan]
-> ba   :['notpresent' 'present' nan]
-> pcv  :['44' '38' '31' '32' '35' '39' '36' '33' '29' '28' nan '16' '24' '37' '30'
 '34' '40' '45' '27' '48' '\t?' '52' '14' '22' '18' '42' '17' '46' '23'
 '19' '25' '41' '26' '15' '21' '43' '20' '\t43' '47' '9' '49' '50' '53'
 '51' '54']
-> wc   :['7800' '6000' '7500' '6700' '7300' nan '6900' '9600' '12100' '4500'
 '12200' '11000' '3800' '11400' '5300' '9200' '6200' '8300' '8400' '10300'
 '9800' '9100' '7900' '6400' '8600' '18900' '21600' '4300' '8500' '11300'
 '7200' '7700' '14600' '6300' '\t6200' '7100' '11800' '9400' '5500' '5800'
 '13200' '12500' '5600' '7000' '11900' '10400' '10700' '12700' '6800'
 '6500' '13600' '10200' '9000' '14900' '8200' '15200' '5000' '16300'
 '12400' '\t8400' '10500' '4200' '4700' '10900' '8100' '9500' '2200'
 '12800' '11200' '19100' '\t?' '12300' '16700' '2600' '26400' '8800'
 '7400' '4900' '8000' '12000' '15700' '4100' '5700' '11500' '5400' '10800'
 '9900' '5200' '5900' '9300' '9700' '5100' '6600']
-> rc   :['5.2' nan '3.9' '4.6' '4.4' '5' '4' '3.7' '3.8' '3.4' '2.6' '2.8' '4.3'
 '3.2' '3.6' '4.1' '4.9' '2.5' '4.2' '4.5' '3.1' '4.7' '3.5' '6' '2.1'
 '5.6' '2.3' '2.9' '2.7' '8' '3.3' '3' '2.4' '4.8' '\t?' '5.4' '6.1' '6.2'
 '6.3' '5.1' '5.8' '5.5' '5.3' '6.4' '5.7' '5.9' '6.5']
-> htn  :['yes' 'no' nan]
-> dm   :['yes' 'no' ' yes' '\tno' '\tyes' nan]
-> cad  :['no' 'yes' '\tno' nan]
-> appet        :['good' 'poor' nan]
-> pe   :['no' 'yes' nan]
-> ane  :['no' 'yes' nan]
-> class        :['ckd' 'ckd\t' 'notckd']
```

As shown in the above figure, three attributes having object type contain numeric values. We first changed the datatype of these three attributes viz pcv, wc and rc.

We changed the values of the following attributes:

Attribute Name	Previous Values	New Values
htn, dm, cad, pe, ane	yes, no	1, 0
rbc, pc	normal, abnormal	1, 0
pcc, ba	present, notpresent	1, 0
Appet	good, poor	1, 0
Classification	ckd, notckd	1, 0

Below is the chart showing datatypes of all attributes after the datatype conversion and data transformation.

```
dataset.dtypes
```

```
age        float64
bp         float64
sg         float64
al         float64
su         float64
rbc        float64
pc         float64
pcc        float64
ba         float64
bgr        float64
bu         float64
sc         float64
sod        float64
pot        float64
hemo       float64
pcv        float64
wc         float64
rc         float64
htn        float64
dm         float64
cad        float64
appet      float64
pe         float64
ane        float64
class        int64
dtype: object
```

	count	mean	std	min	25%	50%	75%	max
age	391.0	51.483376	17.169714	2.000	42.00	55.00	64.50	90.000
bp	388.0	76.469072	13.683637	50.000	70.00	80.00	80.00	180.000
sg	353.0	1.017408	0.005717	1.005	1.01	1.02	1.02	1.025
al	354.0	1.016949	1.352679	0.000	0.00	0.00	2.00	5.000
su	351.0	0.450142	1.099191	0.000	0.00	0.00	0.00	5.000
rbc	248.0	0.810484	0.392711	0.000	1.00	1.00	1.00	1.000
pc	335.0	0.773134	0.419431	0.000	1.00	1.00	1.00	1.000
pcc	396.0	0.106061	0.308305	0.000	0.00	0.00	0.00	1.000
ba	396.0	0.055556	0.229351	0.000	0.00	0.00	0.00	1.000
bgr	356.0	148.036517	79.281714	22.000	99.00	121.00	163.00	490.000
bu	381.0	57.425722	50.503006	1.500	27.00	42.00	66.00	391.000
sc	383.0	3.072454	5.741126	0.400	0.90	1.30	2.80	76.000
sod	313.0	137.528754	10.408752	4.500	135.00	138.00	142.00	163.000
pot	312.0	4.627244	3.193904	2.500	3.80	4.40	4.90	47.000
hemo	348.0	12.526437	2.912587	3.100	10.30	12.65	15.00	17.800
pcv	329.0	38.884498	8.990105	9.000	32.00	40.00	45.00	54.000
wc	294.0	8406.122449	2944.474190	2200.000	6500.00	8000.00	9800.00	26400.000
rc	269.0	4.707435	1.025323	2.100	3.90	4.80	5.40	8.000
htn	398.0	0.369347	0.483235	0.000	0.00	0.00	1.00	1.000
dm	398.0	0.344221	0.475712	0.000	0.00	0.00	1.00	1.000
cad	398.0	0.085427	0.279868	0.000	0.00	0.00	0.00	1.000
appet	399.0	0.794486	0.404584	0.000	1.00	1.00	1.00	1.000
pe	399.0	0.190476	0.393170	0.000	0.00	0.00	0.00	1.000
ane	399.0	0.150376	0.357888	0.000	0.00	0.00	0.00	1.000
class	400.0	0.625000	0.484729	0.000	0.00	1.00	1.00	1.000

The above chart shows the following information about the attributes:

1. **Count:** Number of Non-Null values in the attribute.
2. **Mean:** Mean value of the attribute.
3. **Std:** Standard Deviation value of the attribute.
4. **Min:** Minimum value in the attribute.
5. **25%:** 25% quartile deviation value of the attribute.
6. **50%:** 50% quartile deviation value of the attribute.
7. **75%:** 75% quartile deviation value of the attribute.
8. Max: Maximum value in the attribute

5.3 Univariate Plot and Analysis

5.3.1 Plotting Histogram

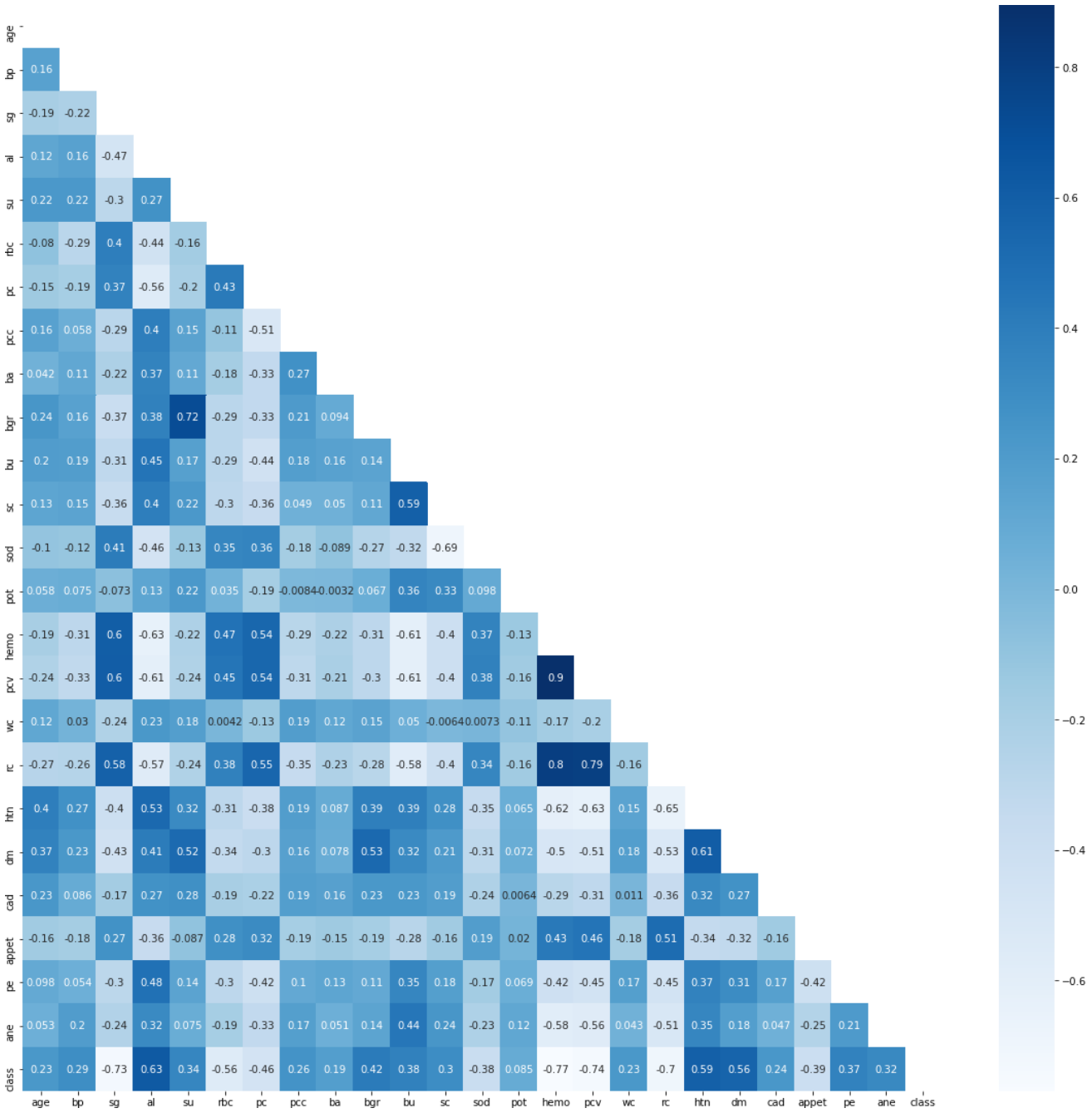
The above graphs show the density of values in the attributes. The x-axisrepresents the values of the attribute(s) and the y-axis represent the count.

Observations:

1. age looks a bit left skewed
2. Blood glucose random is right skewed
3. Blood Urea is also a bit right skewed
4. Rest of the features are lightly skewed
5. A few features having binary values have imbalanced graph.

5.4 Multivariate Plot and Analysis

5.4.1 Plotting Heatmap using Pearson's Correlation



Positive Correlation	Negative Corelation
hemo -> pcv	hemo -> class
hemo -> rc	pcv -> class
pcv -> rc	sg -> class
su -> bgr	rc -> class

6 DATA PRE-PROCESSING AND DATA VISUALIZATION

6.1. Handing Null Values

Chart A

```
dataset.isnull().sum()
```

```
age        9
bp        12
sg        47
al        46
su        49
rbc      152
pc        65
pcc        4
ba         4
bgr       44
bu        19
sc        17
sod       87
pot       88
hemo      52
pcv       71
wc       106
rc       131
htn        2
dm         2
cad        2
appet      1
pe         1
ane        1
class      0
dtype: int64
```

Chart B

```
dataset.isnull().sum()
```

```
age      0
bp       0
sg       0
al       0
su       0
rbc      0
pc       0
pcc      0
ba       0
bgr      0
bu       0
sc       0
sod      0
pot      0
hemo     0
pcv      0
wc       0
rc       0
htn      0
dm       0
cad      0
appet    0
pe       0
ane      0
class    0
dtype: int64
```

Chart A shows the total number of null values in each attribute.

The null/missing values were replaced by the **mean** value(s) of correspondingattribute(s).

Chart B shows the number of null values in each attribute after replacement.

6.2. Feature Selection

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

- Simplification of models to make them easier to interpret by researchers/users
- Shorter training times
- To avoid the curse of dimensionality
- Improve data's compatibility with a learning model class
- Encode inherent symmetries present in the input space.

The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

Benefits of performing feature selection before modeling

- **Reduces Over fitting**: Less redundant data means less opportunity to makedecisions based on noise.
- **Improves Accuracy**: Less misleading data means modeling accuracy improves.
- **Reduces Training Time**: fewer data points reduce algorithm complexity and algorithms train faster.

6.2.1. SelectKBest Feature Selection:

The SelectKBest class just scores the features using a function (by default f_classifbut could be others) and then removes all but the k highest scoring features. So its kind of a wrapper, the important thing here is the function you use to score the features.

6.2.2. Chi2 (chi-square) distribution:

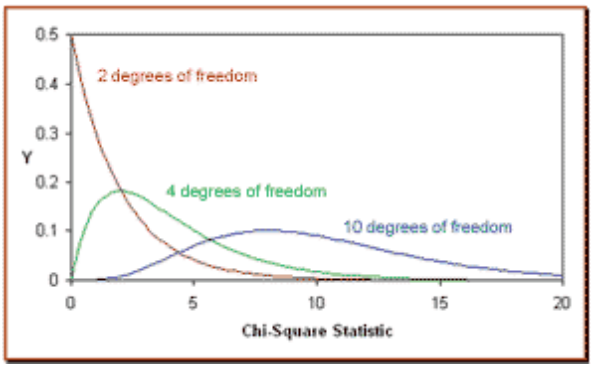
A random variable □ follows chi-square distribution if it can be written as a sum ofsquared standard normal variables.

$$\chi^2 = \sum Zi^2$$

Z_1, Z_2 are standard normal variables

6.2.2.1. Degrees of freedom:

Degrees of freedom refers to the maximum number of logically independent values, which have the freedom to vary. In simple words, it can be defined as the total numberof observations minus the number of independent constraints imposed on the observations.



6.2.3. Chi2 (chi-squrare) test for Feature Selection:

A chi-square test is used in statistics to test the independence of two events. Giventhe data of two variables, we can get observed count O and expected count E. Chi-Square measures how expected count E and observed count O deviates each other.

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

Let's consider a scenario where we need to determine the relationship between the independent

category feature (predictor) and dependent category feature(response). In feature selection, we aim to select the features which are highly dependent on the response.

When two features are independent, the observed count is close to the expectedcount, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square valuethe feature is more dependent on the response and it can be selected for model training.

Below are the 10 attributes which we got after feature selection.

```
['wc', 'bgr', 'bu', 'sc', 'pcv', 'al', 'hemo', 'age', 'su', 'htn']
```

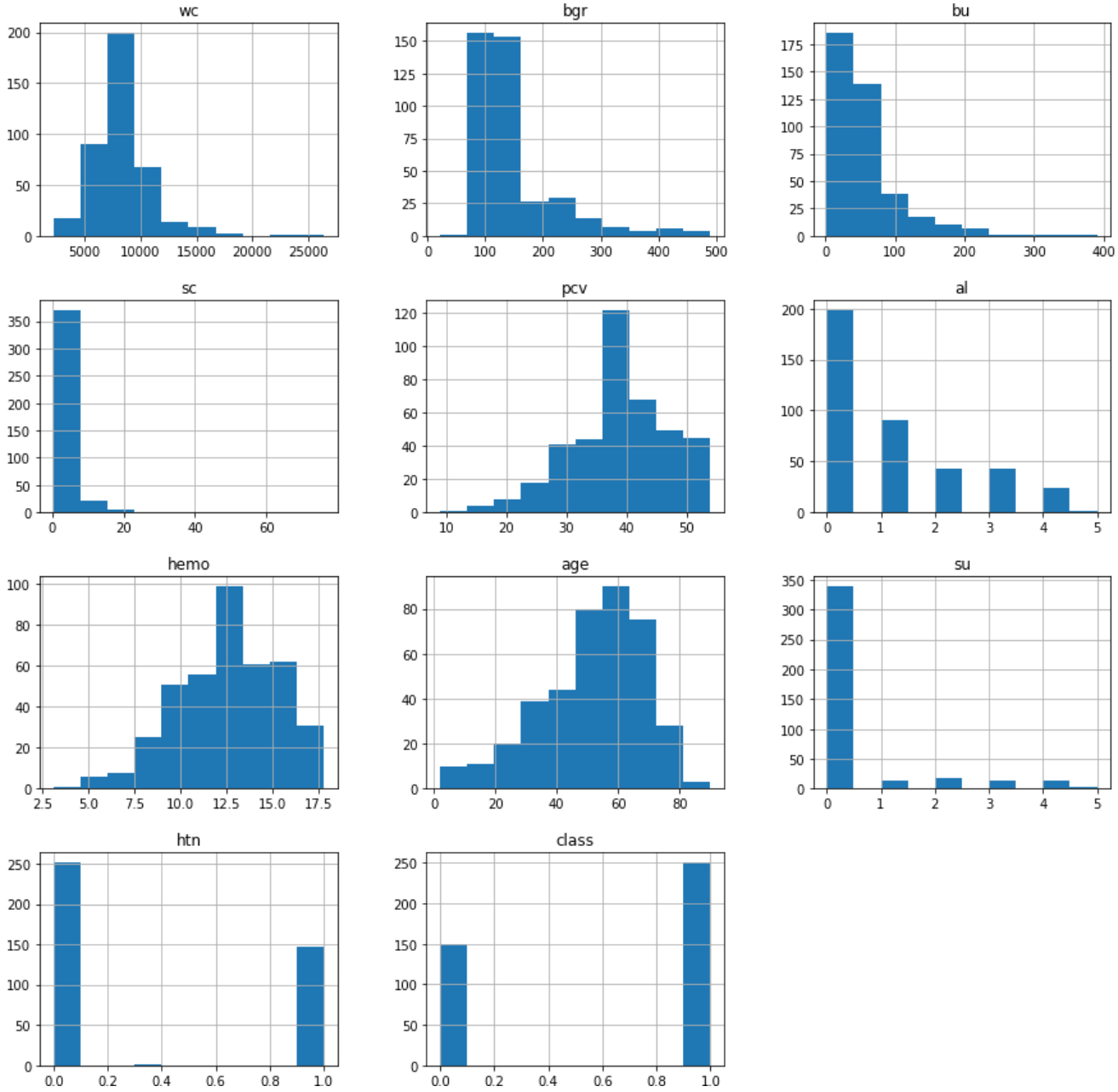
1. White Blood Cell Count
2. Blood Glucose Random
3. Blood Urea
4. Serum Creatinine
5. Packed Cell Volume
6. Albumin
7. Hemoglobin
8. Age
9. Sugar
10. Hypertension

- **Dataset after Feature Selection**

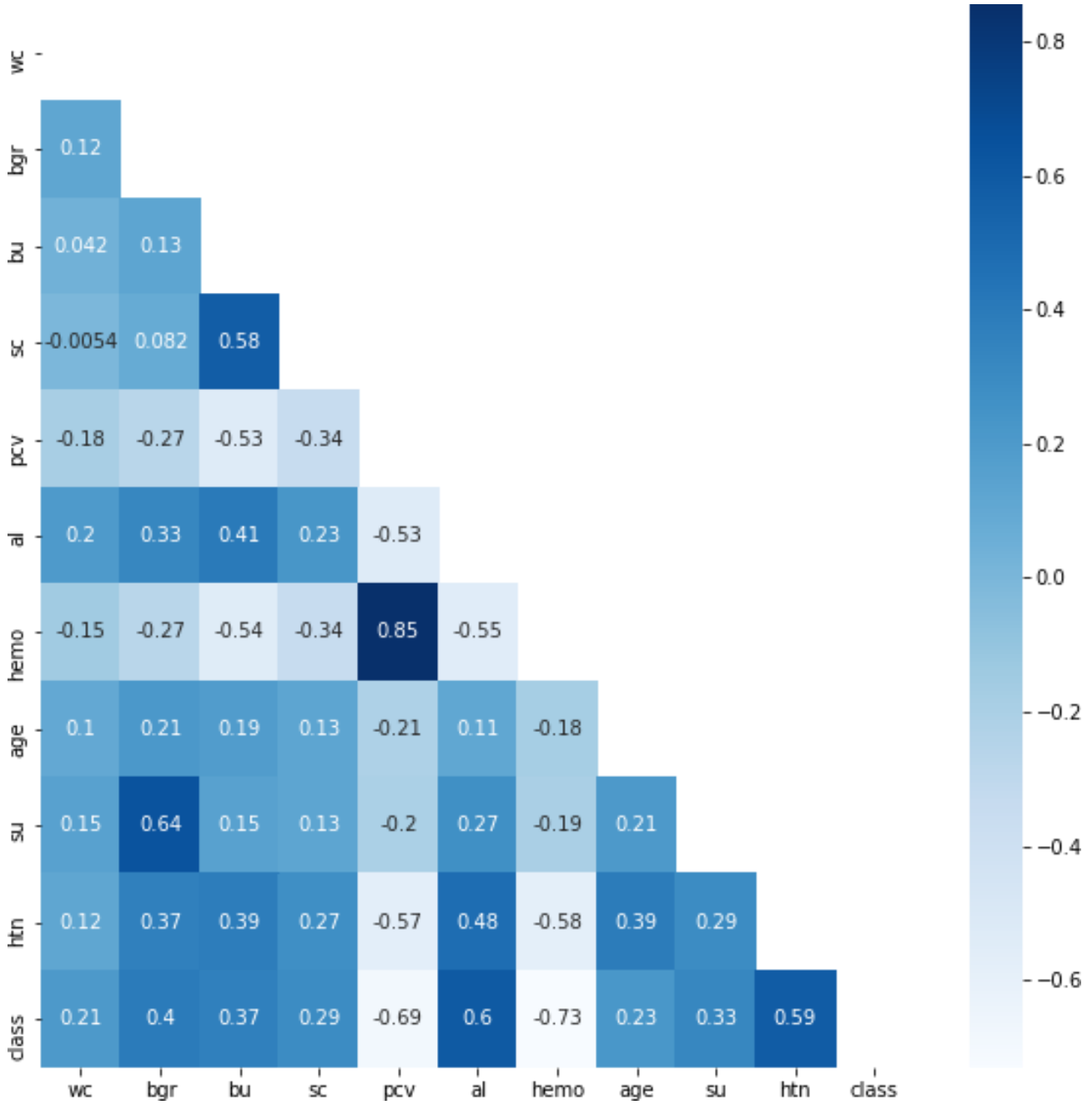
	wc	bgr	bu	sc	pcv	al	hemo	age	su	htn	class
0	7800.0	121.000000	36.0	1.2	44.0	1.0	15.4	48.0	0.0	1.0	1
1	6000.0	148.036517	18.0	0.8	38.0	4.0	11.3	7.0	0.0	0.0	1
2	7500.0	423.000000	53.0	1.8	31.0	2.0	9.6	62.0	3.0	0.0	1
3	6700.0	117.000000	56.0	3.8	32.0	4.0	11.2	48.0	0.0	1.0	1
4	7300.0	106.000000	26.0	1.4	35.0	2.0	11.6	51.0	0.0	0.0	1
...	...	...	...	...	...	...	...	...	...	...	...
395	6700.0	140.000000	49.0	0.5	47.0	0.0	15.7	55.0	0.0	0.0	0
396	7800.0	75.000000	31.0	1.2	54.0	0.0	16.5	42.0	0.0	0.0	0
397	6600.0	100.000000	26.0	0.6	49.0	0.0	15.8	12.0	0.0	0.0	0
398	7200.0	114.000000	50.0	1.0	51.0	0.0	14.2	17.0	0.0	0.0	0
399	6800.0	131.000000	18.0	1.1	53.0	0.0	15.8	58.0	0.0	0.0	0

400 rows × 11 columns

6.3. Univariate Plot after Pre-processing



6.4. Multivariate Plot after Pre-Processing



7 EXPERIMENTAL RESULTS

The dataset was sorted on the basis of ‘class’ column i.e., the first 250 records had a value of ‘ckd’ or ‘1’ and the last 150 had ‘nockd’ or ‘0’ value for the attribute ‘class’.

Before model building and implementation, we **shuffled the dataset**.

❖ **Dataset after shuffle**

	wc	bgr	bu	sc	pcv	al	hemo	age	su	htn	class
341	7300.000000	130.000000	37.000000	0.900000	41.000000	0.0	13.400000	63.0	0.0	0.000000	0
5	7800.000000	74.000000	25.000000	1.100000	39.000000	3.0	12.200000	60.0	0.0	1.000000	1
323	7800.000000	130.000000	30.000000	1.100000	45.000000	0.0	15.900000	43.0	0.0	0.000000	0
111	9000.000000	294.000000	71.000000	4.400000	32.000000	3.0	10.000000	65.0	3.0	1.000000	1
216	8406.122449	107.000000	15.000000	3.072454	38.000000	0.0	12.800000	64.0	0.0	0.000000	1
...	...	...	...	...	...	...	...	...	...	...	...
288	11000.000000	70.000000	46.000000	1.200000	50.000000	0.0	15.900000	56.0	0.0	0.369347	0
361	5400.000000	70.000000	16.000000	0.700000	54.000000	0.0	13.700000	29.0	0.0	0.000000	0
113	9800.000000	148.036517	57.425722	3.072454	38.884498	0.0	12.526437	61.0	2.0	0.000000	1
32	9600.000000	159.000000	39.000000	1.500000	34.000000	1.0	11.300000	61.0	1.0	1.000000	1
141	6500.000000	148.036517	106.000000	6.000000	19.000000	1.0	6.100000	67.0	0.0	1.000000	1

400 rows × 11 columns

7.1 Model Building

We split the dataset into training set and test set. The ratio for the split is 70:30.

	RATIO	ROWS	COLUMNS
TRAINING SET	70 %	280	10
TEST SET	30 %	120	10

7.2 Decision Tree Classifier

After implementing Decision Tree Classifier Algorithm on the dataset, we got:

7.2.1 Classification Metrices

```
classification report is:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97        45
           1       0.99      0.97      0.98        75

    accuracy                           0.97       120
   macro avg       0.97      0.98      0.97       120
weighted avg       0.98      0.97      0.98       120
```

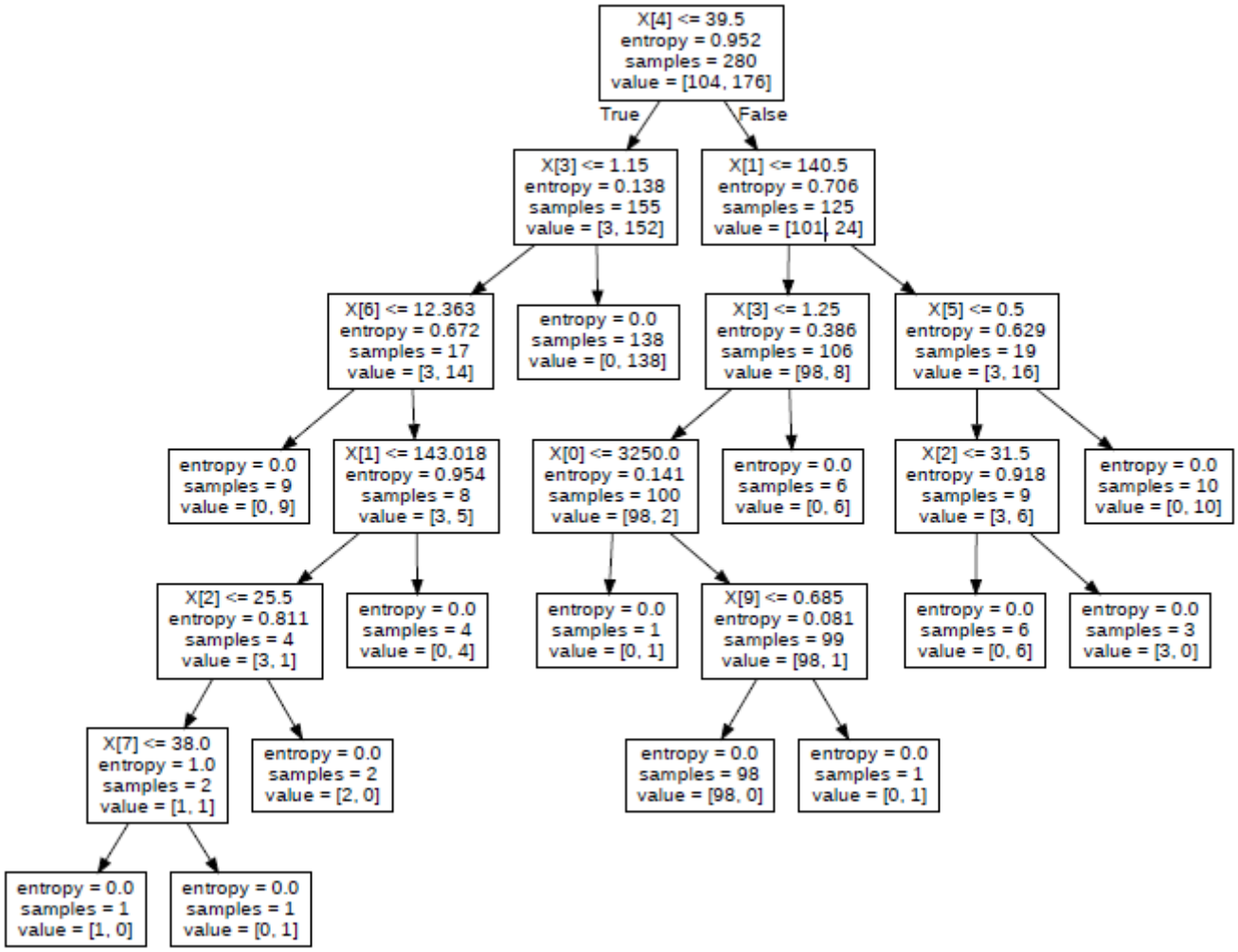
7.2.2 Confusion Matrix

```
Confusion matrix:
 [[44  1]
 [ 2 73]]
```

Analysis

Criterion	Entropy
Max Depth	7
Test Size	120
True Positives	73
True Negatives	44
False Positives	2
False Negatives	1
Accuracy	97%

Decision Tree



7.3 K-Nearest Neighbor Classifier

After implementing K-Nearest Neighbour Classifier Algorithm on the dataset, we got:

7.3.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.65      0.93      0.76        45
           1       0.95      0.69      0.80        75

    accuracy                           0.78       120
   macro avg       0.80      0.81      0.78       120
weighted avg       0.83      0.78      0.79       120
```

7.3.2 Confusion Matrix

```
confusion matrix:
 [[42  3]
 [23 52]]
```

Analysis

Neighbours	**5**
Test Size	120
True Positives	52
True Negatives	42
False Positives	23
False Negatives	3
Accuracy	78%

7.4 Support Vector Machine Classifier

After implementing Support Vector Classifier Algorithm on the dataset, we got:

7.4.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.77      0.44      0.56        45
           1       0.73      0.92      0.82        75

    accuracy                           0.74       120
   macro avg       0.75      0.68      0.69       120
weighted avg       0.75      0.74      0.72       120
```

7.4.2 Confusion Matrix

```
confusion matrix:
 [[20 25]
 [ 6 69]]
```

Analysis

Test Size	**120**
True Positives	69
True Negatives	20
False Positives	6
False Negatives	25
Accuracy	74%

7.5 Random Forest Classifier

After implementing Random Forest Classifier Algorithm on the dataset, we got:

7.5.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        45
           1       1.00      0.96      0.98        75

    accuracy                           0.97       120
   macro avg       0.97      0.98      0.97       120
weighted avg       0.98      0.97      0.98       120
```

7.5.2 Confusion matrix

```
confusion matrix:
 [[45  0]
 [ 3 72]]
```

Analysis

Test Size	**120**
True Positives	72
True Negatives	45
False Positives	3
False Negatives	0
Accuracy	97%

7.6 Logistic Regression

After implementing Logistic Regression Algorithm on the dataset, we got:

7.6.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.91      0.96      0.93        45
           1       0.97      0.95      0.96        75

    accuracy                           0.95       120
   macro avg       0.94      0.95      0.95       120
weighted avg       0.95      0.95      0.95       120
```

7.6.2 Confusion Matrix

```
confusion matrix:
 [[43  2]
 [ 4 71]]
```

Analysis

Test Size	120
True Positives	71
True Negatives	43
False Positives	4
False Negatives	2
Accuracy	95%

7.7 Gaussian Naïve Bayes Classifier

After implementing Gaussian Naïve Bayes Classifier Algorithm on the dataset, we got:

7.7.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.85      0.98      0.91        45
           1       0.99      0.89      0.94        75

    accuracy                           0.93       120
   macro avg       0.92      0.94      0.92       120
weighted avg       0.93      0.93      0.93       120
```

7.7.2 Confusion Matrix

```
confusion matrix:
 [[44  1]
 [ 8 67]]
```

Analysis

Test Size	120
True Positives	67
True Negatives	44
False Positives	8
False Negatives	1
Accuracy	93%

7.8 Ada Boot Classifier

After implementing Ada Boot Classifier Algorithm on the dataset, we got:

7.8.1 Classification Metrices

```
classification report is::
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        45
           1       1.00      0.95      0.97        75

    accuracy                           0.97       120
   macro avg       0.96      0.97      0.97       120
weighted avg       0.97      0.97      0.97       120
```

7.8.2 Confusion Matrix

```
confusion matrix:
 [[45  0]
 [ 4 71]]
```

Analysis

Test Size	120
True Positives	71
True Negatives	45
False Positives	4
False Negatives	0
Accuracy	97%

8 CONCLUSIONS

Machine learning algorithms have shown good classification performance on our project **"Prediction of Chronic Kidney Disease using Machine Learning Algorithms"**. The algorithms that we used, work on existing data to gather information in order to predict. That's why machine-learning methods are not as widely used as one might expect. We argue that our model can at least partially be attributed to uncertainties regarding the diagnosis of Chronic Kidney Disease. The experimental results can assist healthcare personals to make early decisions to protect a person from Chronic Kidney Disease and if a person has chronic Kidney disease, it will help them to curb this disease so that it won't advance to other stages which will in turn help them to reduce the number of Kidney failures due to CKD. The main aim of this project was to design and implement a model that will make prediction about whether a patient has Chronic Kidney Disease or not using machine learning algorithms on a relevant dataset.

By using a Machine Learning Technique called **Feature Selection**, we came to theconclusion that among 24 attributes, the top 10 attributes which contribute to/affect our decision the most are:

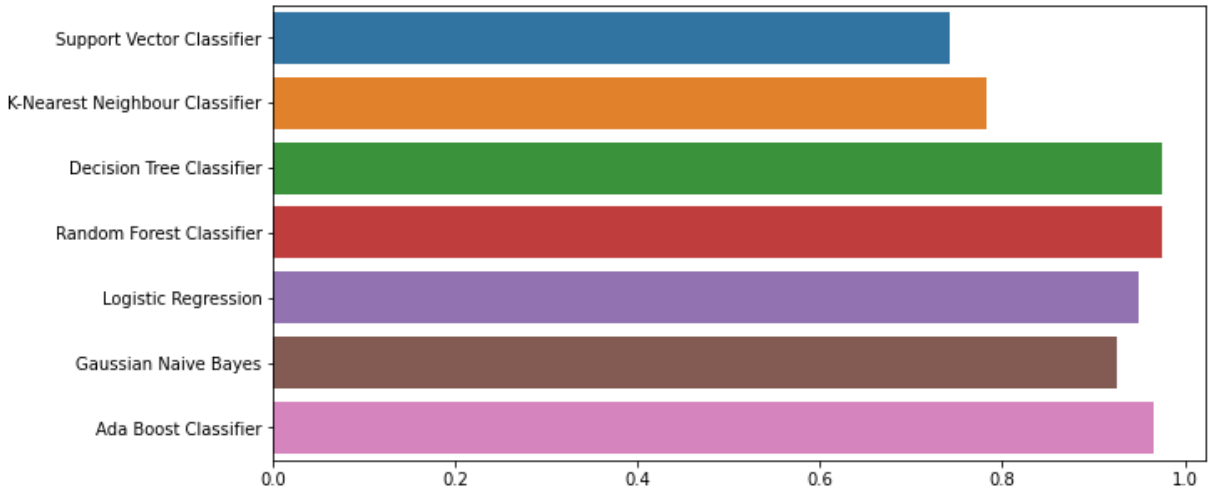
1. White Blood Cell Count
2. Blood Glucose Random
3. Blood Urea
4. Serum Creatinine
5. Packed Cell Volume
6. Albumin
7. Hemoglobin
8. Age
9. Sugar
10. Hypertension

These 10 attributes are the most important parameters among all the 24 attributes for the prediction of Chronic Kidney Disease according to the dataset and our research.

Since our project was classification based, so we implemented seven classifications on our dataset. We implemented Decision Tree Classifier, K-Nearest Neighbour, Support Vector Classifier, Ada Boost Classifier, Gaussian Naïve bayes, Random Forest classifier and Logistics regression classification algorithms. Among these seven algorithms, Random Forest Classifier, Ada Boost and Decision Tree have the highest accuracy of **97%** approximately. The KNN, SVM, Ada Boost, GNB, Decision Tree, Random Forest and Logistics Regression have accuracy of **78%**, 74**%, 97%, 93%, 97%, 97%, 95%** respectively. The different performance measures that are being compared are Accuracy, F1-score, Precision and Recall.

Below is the graph and table showing algorithms with their accuracies mentioned.

Graph



Table

	model	score
0	Support Vector Classifier	0.741667
1	K-Nearest Neighbour Classifier	0.783333
5	Gaussian Naive Bayes	0.925000
4	Logistic Regression	0.950000
6	Ada Boost Classifier	0.966667
2	Decision Tree Classifier	0.975000
3	Random Forest Classifier	0.975000

Score Table

Based on the accuracy scores of the above algorithms, we came to conclusion that **Random Forest Classifier** is the best suited algorithm for this dataset having accuracy of **97%** approx.

9 References

1. Predictive analytics for chronic kidney disease using machine learning techniques: https://ieeexplore.ieee.org/abstract/document/8025242
2. Early prediction of chronic kidney disease using machine learning supported by predictive analytics: https://ieeexplore.ieee.org/abstract/document/8477876
3. Chronic kidney disease prediction using machine learning models: https://www.researchgate.net/profile/Revathy-Ramesh-3/publication/341398109_Chronic_Kidney_Disease_Prediction_using_machine_Learning_Models/links/5ebe42b1458515626ca85977/Chronic-Kidney-Disease-Prediction-using-machine-Learning-Models.pdf
4. Optimization of prediction method of chronic kidney disease using machine learning algorithm: https://ieeexplore.ieee.org/abstract/document/9376787
5. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression: https://link.springer.com/article/10.1186/s12967-019-1860-0
6. Chronic kidney disease prediction using machine learning: https://www.academia.edu/download/56493051/41_Paper_310318109_IJCSIS_Camera_Ready_pp.308-311.pdf
7. Detection of chronic kidney disease using machine learning algorithms with least number of predictors: http://doras.dcu.ie/23782/1/SAI_PAPER_FORMAT_MARWA_CKD.pdf
8. Preemptive diagnosis of chronic kidney disease using machine learning techniques: https://ieeexplore.ieee.org/abstract/document/8606040
9. Chronic kidney disease prediction using machine learning methods: https://ieeexplore.ieee.org/abstract/document/9185249
10. Risk prediction of chronic kidney disease using machine learning algorithms: https://ieeexplore.ieee.org/abstract/document/9225548
11. https://archive.ics.uci.edu/ml/index.php/
12. https://scikit-learn.org/
13. https://www.kaggle.com/
14. https://www.kidney.org/atoz/content/about-chronic-kidney-disease
15. https://www.mayoclinic.org/diseases-conditions/chronic-kidney-
16. disease/symptoms-causes/syc-20354521
17. https://www.kidneyfund.org/kidney-disease/chronic-kidney-disease-ckd/

18. www.google.com
19. Mathematics for Machine Learning by Marc Peter Deisenroth, A. Aldo Faisal and Cheng Soon Ong.
20. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools and Techniques to build Intelligent Systems by Aurelien Geron.
21. Introduction to Machine Learning with Python: A Guide for Data Scientists by Andreas C. Muller & Sarah Guido.
22. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn and TensorFlow 2 by Sebastian Raschka and Vahid Mirjalili.
23. Understanding Machine Learning: From Theory to Algorithms by Shai Shalev- Shwartz and Shai Ben-David.