Comprehensive Analysis of Factors Influencing Chronic Kidney Disease: Data Driven Insights

By Statistics Team-2 | PGDM Section-B (2025-27)

Jaspreeth Kour (2501145)
Meghna Panniru (2501117)
Pavan Kumar Nakka (2501123)
Sai Lalith Reddy (2501104)
Vinay Chandra (2501121)
Miduthuri Karthikeya (2501118)
Nitesh Ray (2501124)
Yashaswini Mishra (2501144)
Pinjarla Varun Yadav (2501128)
Krushnal Patil (2501110)
Aditya Kakde (2501083)

Under the guidance of Dr. Shaheen

Index

Descriptive Analysis of Numeric Data

Dataset Overview

Dataset

data

Variables

7

Observations

1973

Numerical Statistical Analysis by Variable

stats	New.Age	BP	SG	Chl	Sod	Pot	Bun
min	40.0000	120.0100	1.0100	150.0900	135.0000	3.5000	10.0000
1st Qu	49.0000	130.4500	1.0150	174.3500	137.6800	3.8500	15.0800
mean	58.0243	140.2345	1.0200	199.2490	140.1147	4.2387	20.0735
median	58.0000	140.7000	1.0200	199.5000	140.2100	4.2400	20.0600
3rd Qu	67.0000	150.0500	1.0250	224.4500	142.5800	4.6200	25.0400
max	78.0000	159.9700	1.0300	249.7500	144.9900	5.0000	29.9900
sd	10.4566	11.5307	0.0057	28.6516	2.8705	0.4340	5.7462
skew	0.0879	-0.0448	0.0284	0.0085	-0.0478	-5.0201e-05	0.0138
kurtosi	-1.1703	-1.1854	-1.1460	-1.2338	-1.1958	-1.2155	-1.1950
cv	0.1802	0.0822	0.0056	0.1438	0.0205	0.1024	0.2863
var	109.3412	132.9562	3.2739e-05	820.9154	8.2399	0.1883	33.0186
n	1973.0000	1973.0000	1973.0000	1973.0000	1973.0000	1973.0000	1973.0000

*Data generated using Bluesky Statistics

Interpretation of Descriptive Analysis

	Explanation
New Age	It ranges from **40 to 78 years**, with an **average of 58 years** – most participants are middle-aged to elderly.
Blood Pressure (BP)	It averages around **140 mmHg**, indicating that on average participants are near the high blood pressure range.
Specific Gravity (SG)	SG of urine is stable (mean **~1.020**), suggesting most values are within normal kidney function range.
Cholesterol (Chl)	Chl levels average around **199 mg/dL**, which is close to the upper limit of normal.
Sodium (Sod)	Sod averages **~140 mmol/L**, within the normal blood sodium range.
Potassium (Pot)	It averages **~4.24 mmol/L**, also in the normal range.
Blood Urea Nitrogen (BUN)	It averages **~20**, which is on the higher side of normal kidney function values.

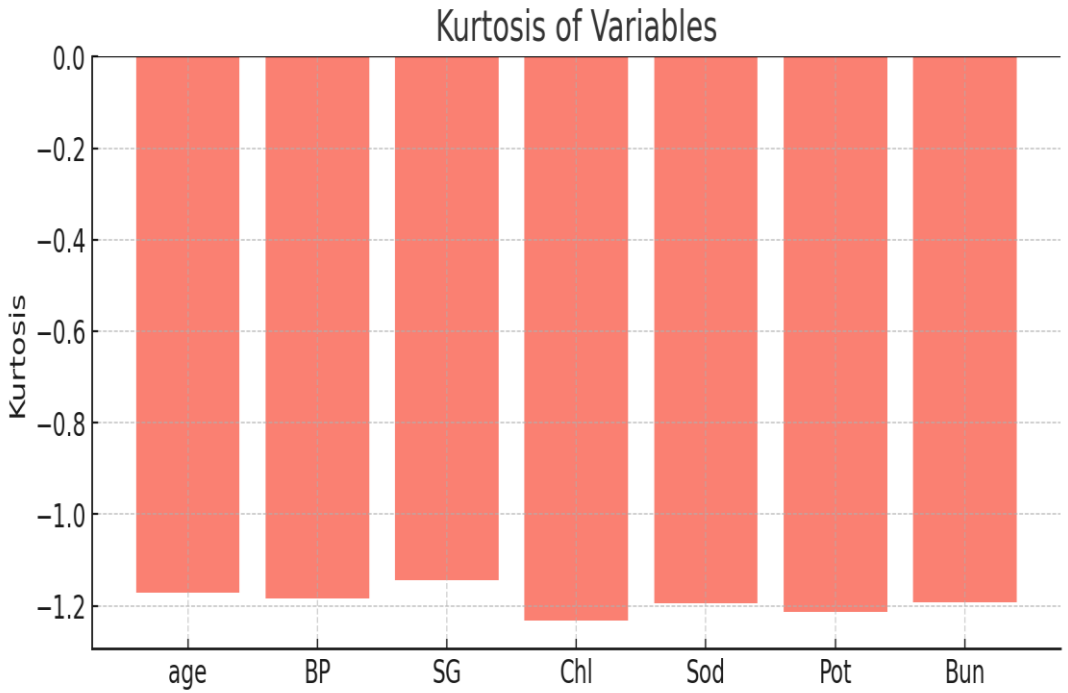
NOTE

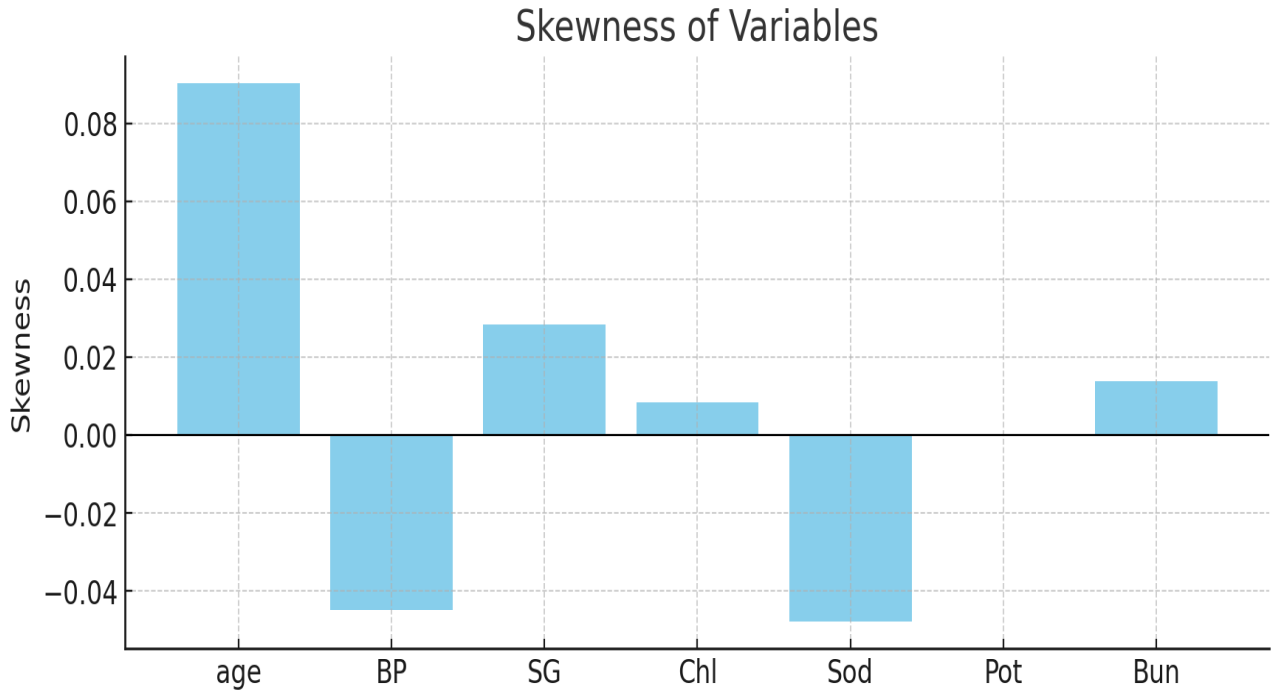
Very low **skewness** mean the data is fairly symmetric and not heavily skewed.

Very low **kurtosis** mean the data has flattened peaks and therefore too much variation in data.

Coefficient of variation (CV) is highest for **BUN** (0.2863), meaning it varies more among people compared to other measures.

Kurtosis and Skewness of Numerical Data





Kurtosis

All values are **negative** (around -1.17 to -1.23) → **platykurtic** shape.

This means the data has **flatter peaks** and **lighter tails** than a normal curve (values are more spread out).

Skewness

All variables have skewness values close to **0** (between -0.05 and 0.09)

This means the data is **fairly symmetrical**.

Frequency Distribution Table for Numeric Data

Dataset Overview

Dataset	Variables	Nominals	Observations
data	7	0	1973

Summary by Variables

New.Age	BP	SG	Chl	Sod	Pot	Bun
Min. :40.00	Min. :120.0	Min. :1.010	Min. :150.1	Min. :135.0	Min. :3.500	Min. :10.00
1st Qu.:49.00	1st Qu.:130.4	1st Qu.:1.015	1st Qu.:174.3	1st Qu.:137.7	1st Qu.:3.850	1st Qu.:15.08
Median :58.00	Median :140.7	Median :1.020	Median :199.5	Median :140.2	Median :4.240	Median :20.06
Mean :58.02	Mean :140.2	Mean :1.020	Mean :199.2	Mean :140.1	Mean :4.239	Mean :20.07
3rd Qu.:67.00	3rd Qu.:150.1	3rd Qu.:1.025	3rd Qu.:224.4	3rd Qu.:142.6	3rd Qu.:4.620	3rd Qu.:25.04
Max. :78.00	Max. :160.0	Max. :1.030	Max. :249.8	Max. :145.0	Max. :5.000	Max. :29.99

Frequency table for New.Age (Starting Rows)

New.Age	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
44	85	4.3082	12.1642	4.3082	12.1642
59	72	3.6493	55.9047	3.6493	55.9047
43	69	3.4972	7.8561	3.4972	7.8561
45	69	3.4972	15.6614	3.4972	15.6614
47	69	3.4972	22.5038	3.4972	22.5038
46	66	3.3452	19.0066	3.3452	19.0066
58	64	3.2438	52.2554	3.2438	52.2554
52	62	3.1424	34.4653	3.1424	34.4653
74	61	3.0917	94.1206	3.0917	94.1206
55	60	3.0411	43.1830	3.0411	43.1830
57	59	2.9904	49.0117	2.9904	49.0117
68	59	2.9904	79.3208	2.9904	79.3208
75	59	2.9904	97.1110	2.9904	97.1110

*Data generated using Bluesky Statistics

Interpretation from Frequency Distribution Table of Numeric Data

Observation-1

Age group **44 years** has the highest count (**85 people, ~4.3%**).

Other common ages: **59** (72 people), **43**, **45**, **47 years** (69 people each).

Observation-2

The distribution is spread out, with many participants between **40–75 years**.

Only **5 people** are aged **78**, the highest age recorded.

Observation-3

The **middle range** (around 58 years) appears to have the largest concentration of people.

Frequency Distribution Table of Categorical Data

Dataset Overview

Dataset	Variables	Nominals	Observations
data	4	4	1973

Summary of variables

gender	Ckd_status	Smoking.Status	SES
Female: 963	No CKD :367	Current: 441	High :363
Male :1010	Stage 1 :100	Former : 425	Low :611
NA	Stage 2 :433	Never :1107	Middle:999
NA	Stage 3a:633	NA	NA
NA	Stage 3b:430	NA	NA
NA	Stage 4 : 10	NA	NA

Frequency Table of gender

gender	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Male	1010	51.1911	100.0000	51.1911	100.0000
Female	963	48.8089	48.8089	48.8089	48.8089
NA	0	0.0000	100.0000		

Frequency Table of CKD_status

Ckd_status	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Stage 3a	633	32.0831	77.6989	32.0831	77.6989
Stage 2	433	21.9463	45.6158	21.9463	45.6158
Stage 3b	430	21.7942	99.4932	21.7942	99.4932
No CKD	367	18.6011	18.6011	18.6011	18.6011
Stage 1	100	5.0684	23.6695	5.0684	23.6695
Stage 4	10	0.5068	100.0000	0.5068	100.0000
NA	0	0.0000	100.0000		

Frequency Table for smoking_status

Smoking.Status	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Never	1107	56.1074	100.0000	56.1075	100.0000
Current	441	22.3517	22.3517	22.3517	22.3517
Former	425	21.5408	43.8926	21.5408	43.8925
NA	0	0.0000	100.0000		

*Data generated using Bluesky Statistics

Interpretation from Frequency Distribution Table of Categorical Data

Summary of variables

SES	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Middle	999	50.6336	100.0000	50.6336	100.0000
Low	611	30.9681	49.3664	30.9681	49.3664
High	363	18.3984	18.3984	18.3984	18.3984
NA	0	0.0000	100.0000		

GENDER

Almost equal distribution:

- **Male (51.2%)**
- **Female (48.8%)**

CKD (Chronic Kidney Disease) Status

- Most common stages: **Stage 3a (32.1%)**, **Stage 2 (21.9%)**, and **Stage 3b (21.8%)**.
- Only **18.6%** have **No CKD**, meaning most people in the dataset have some stage of CKD.
- **Stage 4** is rare (0.5%).

Smoking Status

Almost equal distribution:

- **Never smoked**: 56.1% (majority)
- **Current smokers**: 22.3%
- **Former smokers**: 21.5%

Socio-Economic Status (SES)

- **Middle SES**: 50.6% (half of participants)
- **Low SES**: 31%
- **High SES**: 18.4%

*Data generated using Bluesky Statistics

Frequency Distribution Table of Categorical Data

Dataset Overview

Dataset	Variables	Nominals	Observations
data	4	4	1973

Summary of variables

Htn	Dm	Ane	Hd
No : 967	No : 914	No:1973	No:1973
Yes:1006	Yes:1059	NA	NA

Frequency Table of Htn (Hyper Tension)

Htn	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Yes	1006	50.9883	100.0000	50.9883	100.0000
No	967	49.0117	49.0117	49.0117	49.0117
NA	0	0.0000	100.0000		

Frequency Table of Dm (Diabetes Mellitus)

Dm	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
Yes	1059	53.6746	100.0000	53.6746	100.0000
No	914	46.3254	46.3254	46.3254	46.3254
NA	0	0.0000	100.0000		

Frequency Table for Ane (Anaemia)

Ane	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
No	1973	100	100	100	100
NA	0	0	100		

Frequency Table for Hd (Heart Disease)

Hd	Frequency	Percent	CumPercent	Valid Percent	Valid CumPercent
No	1973	100	100	100	100
NA	0	0	100		

*Data generated using Bluesky Statistics

Interpretation from Frequency Distribution Table of Categorical Data

Originally, the four variables mentioned consists of numeric data as 0 and 1. 0 means No and 1 means Yes. As the values are fixed and not continuous, this report converted those numeric data into categorical data and then conducted the interpretation.

Hypertension (Htn)

50.99% of the participants have hypertension. Since high blood pressure is both a cause and effect of CKD, this high prevalence suggests a strong link between hypertension and kidney disease risk in the dataset.

Diabetes (Dm)

53.67% of participants have diabetes. Diabetes is a leading cause of CKD, and its slightly higher prevalence than hypertension indicates it may be equally, if not more, important in this population.

Anaemia (Ane)

100% of the entries show “No” for anemia. This means the dataset does not capture variation for anemia, so it’s not contributing to CKD identification in this data — even though medically, anemia is a known complication of CKD.

Heart Disease (Hd)

100% of the entries show “No” for heart disease. Similarly, no variation is recorded for heart disease in this dataset, so it cannot influence CKD prediction here, even though CKD and cardiovascular disease are strongly interconnected in reality.

Why we took the numerical variables?

Numeric

These are continuous measurements from lab tests or vitals — crucial because CKD often shows up as abnormal readings.

- Age → CKD risk increases with age.
- BP (Blood Pressure) → High BP damages kidneys over time and is both a cause and effect of CKD.
- SG (Specific Gravity) → Measures kidney's ability to concentrate urine; abnormal SG can signal kidney problems.
- Chl (Cholesterol) → High cholesterol can indicate vascular issues that also damage kidneys.
- Sod (Sodium) → Imbalances are common in CKD because kidneys can't regulate them properly.
- Pot (Potassium) → Imbalances are common in CKD because kidneys can't regulate them properly.
- BUN (Blood Ure Nitrogen) → Elevated BUN is a direct indicator of reduced kidney function.

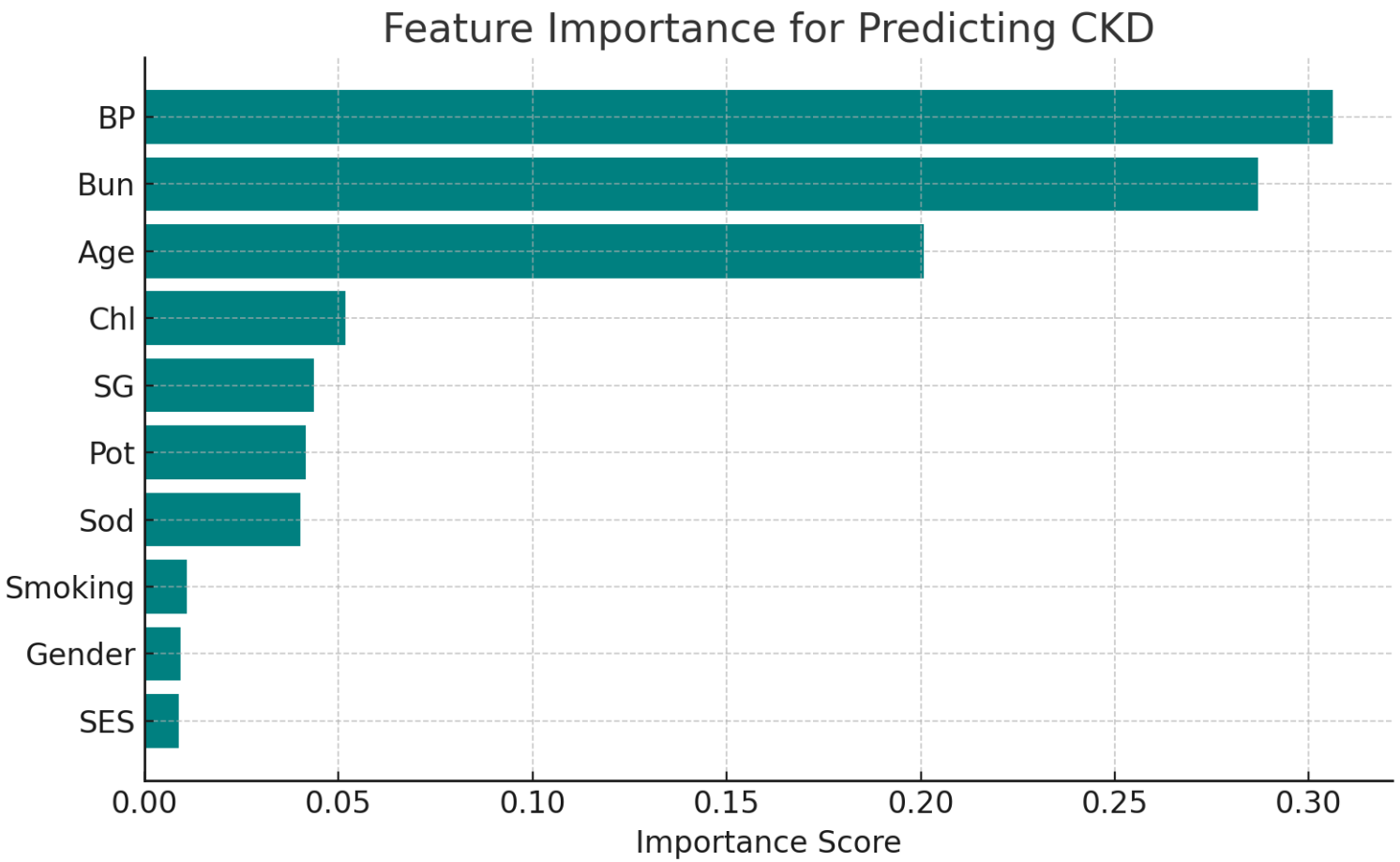
Why we took the categorical variables?

Categorical

These describe patient characteristics or habits that influence CKD risk.

- Gender → May show patterns (e.g., some studies find CKD more in females, others in males).
- CKD Status → Direct classification of the disease stage.
- Smoking Status → Smoking reduces kidney blood flow, accelerating CKD progression.
- Socio-Economic Status (SES) → Affects diet, healthcare access, and risk factors for CKD.
- Htn (Hypertension) → High blood pressure is both a *cause* and *complication* of CKD. Damaged kidneys can't regulate blood pressure well, and high BP further harms kidney blood vessels.
- Dm (Diabetes Mellitus) → High blood sugar damages kidney filtering units (nephrons), making diabetes the **leading cause** of CKD globally.
- Ane (Anemia) → CKD often reduces the kidney's production of erythropoietin (needed for red blood cell formation), leading to anemia, which can be an early complication.
- Hd (Heart Disease) → CKD and heart disease are closely linked; poor kidney function increases heart disease risk and vice versa.

Variable Importance Description w.r.t CKD Status



Variable Importance

The most important factors for identifying **CKD status** were:

Blood Pressure (BP) → ~30.6% importance

Blood Urea Nitrogen (BUN) → ~28.7% importance

Age → ~20.1% importance

Cholesterol (Chl) → ~5.2% importance

Specific Gravity (SG) → ~4.4% importance

Potassium (Pot) → ~4.1% importance

Sodium (Sod) → ~4.0% importance

Smoking Status → ~1.1% importance

Gender → ~0.9% importance

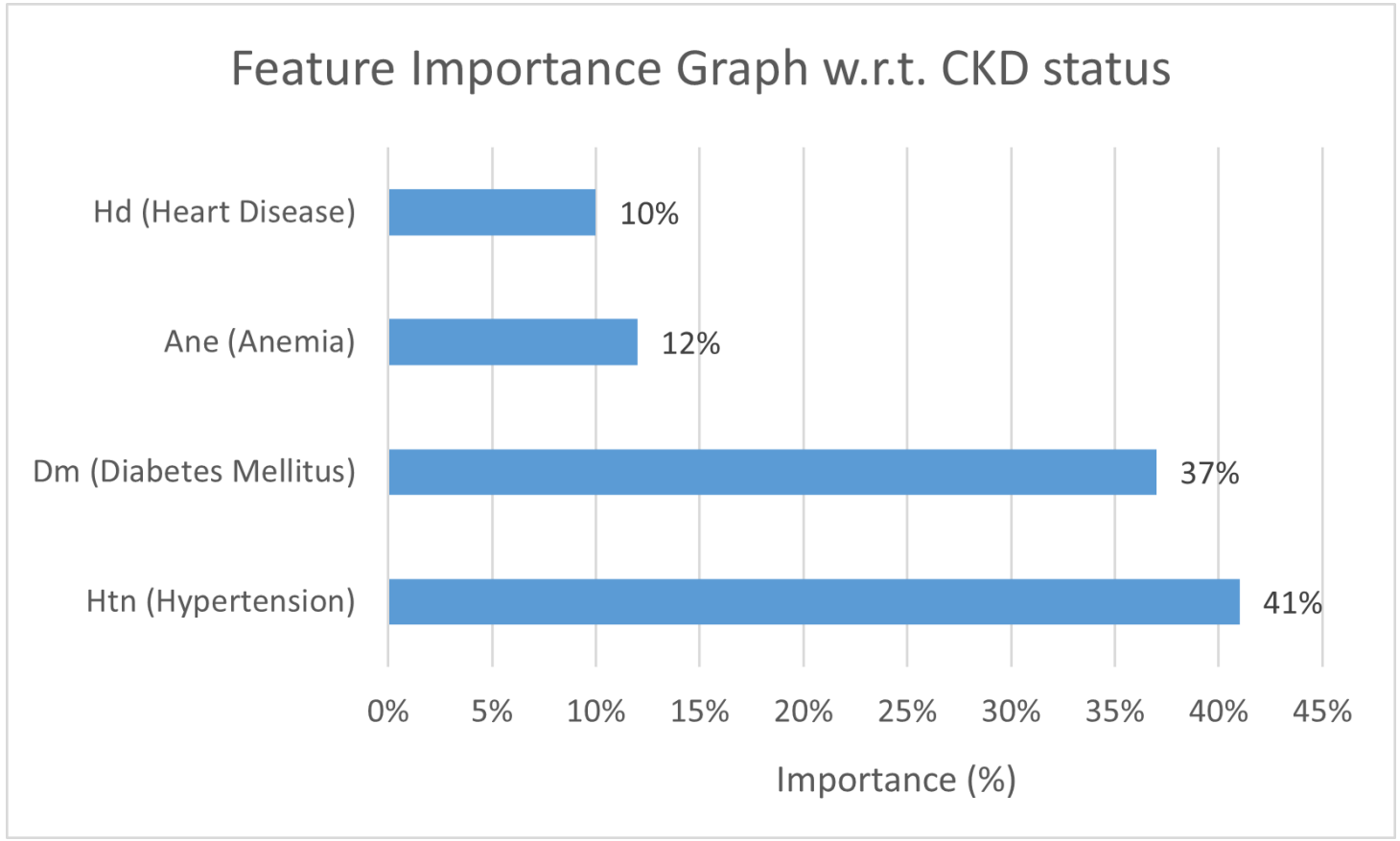
Socio-Economic Status (SES) → ~0.9% importance

Interpretation

The top three — **BP, BUN,** and **Age** — are the strongest predictors in this dataset for whether someone has CKD.

Lifestyle/demographic variables (Smoking, Gender, SES) contribute less directly but still add context to the prediction.

Variable Importance Description w.r.t CKD Status



Variable Importance

The most important factors for identifying **CKD status** were:

Hyertension (Htn) → ~41% importance

Diabetes (Dm) → ~37% importance

Anaemia (Ane) → ~12% importance

Heart Disease (Hd) → ~10% importance

Interpretation

Hypertension and diabetes together account for nearly **78%** of the predictive power for CKD status in this variable set, making them the most critical factors to monitor. Anemia and heart disease add complementary information, helping refine CKD detection, especially in later stages.

Exploratory Data Analysis

Age Distribution

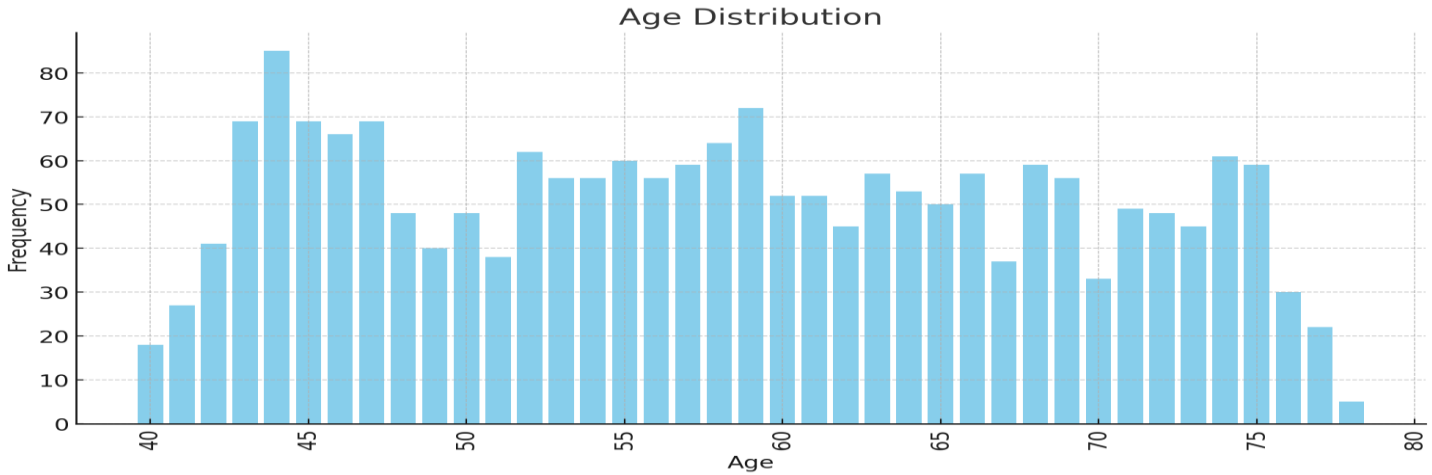


Fig-1: Most participants are between 40–75 years, with peaks around 44, 59, and mid-40s.

CKD Status Distribution

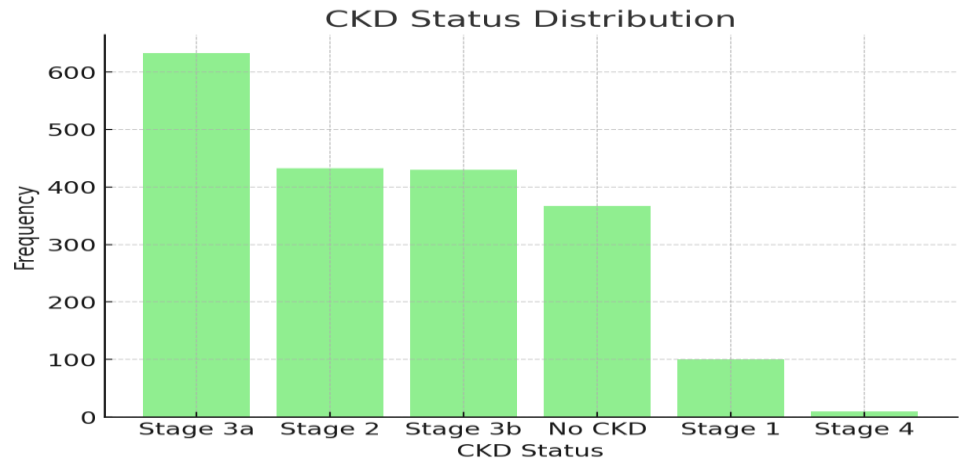


Fig-2: Stages 3a, 2, and 3b dominate; very few have Stage 4.

Smoking Status Distribution

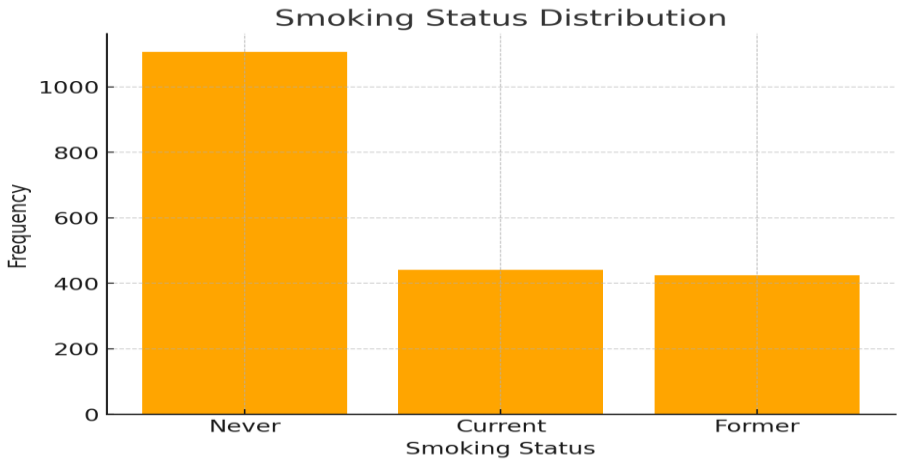


Fig-3: Majority never smoked, followed by current and former.

SES Distribution

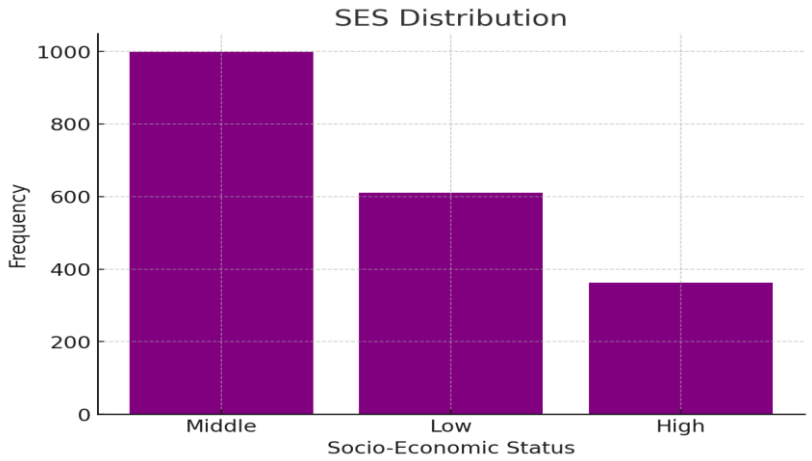


Fig-4: Middle class forms the majority, then low and high.

Gender Distribution

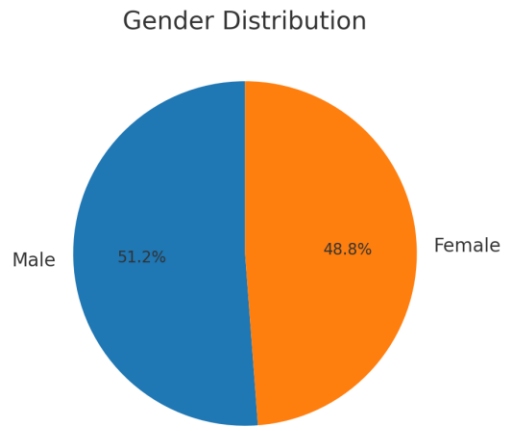


Fig-5: Almost equal male–female ratio.

THANK YOU