SICP Assignment 1

Master of Science in Computer Science
Course Name: Machine Learning
Course Code: MCS 7102
Lecturer: Dr. Joyce Nabende

Kasule John Trevor
2024/HD05/21930U

September 11, 2024

Introduction

Why this dataset: The Chronic Kidney Disease (CKD) dataset provides various health metrics to differentiate patients with CKD from those without. The main objective is to explore which health factors most predict CKD.

Data Loading

The dataset was loaded using the `pandas` library in Python. After loading, I examined the first few rows to understand its structure. I also identified missing values and non-numeric entries in columns that should have numeric data.

Dataset Overview

The dataset contains 400 entries/rows and 26 columns with features related to kidney disease diagnosis. Below is a brief description of some key columns:

- **id:** A unique identifier for each record.
- **age:** The age of the patient.
- **bp:** Blood pressure.
- **sg:** Specific gravity of urine.
- **al:** Albumin levels.
- **su:** Sugar levels.
- **rbc:** Red blood cells.
- **pc:** Pus cell levels.
- **pcc:** Pus cell clumps.
- **ba:** Bacteria levels.
- **pcv:** Packed cell volume.
- **wc:** White blood cell count.
- **rc:** Red blood cell count.
- **htn:** Hypertension.
- **dm:** Diabetes mellitus.
- **cad:** Coronary artery disease.
- **appet:** Appetite level.
- **pe:** Pedal edema.

- **ane:** Anemia.
- **classification:** The target label indicating whether the patient has chronic kidney disease (ckd) or not.

Some columns had notable missing values. Some of these columns include;

- **rbc:** (Red Blood Cells) with 152 missing values,
- **pcv:** (Packed Cell Volume) with 70 missing values
- **wc:** (White Blood Cell Count) with 105 missing values,
- **rc:** (Red Blood Cell Count) with 130 missing values.

Data Types: There are a mix of numerical and categorical columns. Some columns like pcv, wc, and rc should be numeric but are stored as objects, likely due to data entry issues.
Target variable: The classification column is the target variable, containing categories like "ckd" (chronic kidney disease).

Data Rectification

These inconsistencies were identified, such as:

- Non-numeric values in numeric columns like `pcv`, `wc`, and `rc`.
- Non-standard values in categorical columns like `dm` and `cad`.

These issues were corrected to ensure the dataset was clean for analysis. I cleaned the invalid entries in pcv, wc, rc, and classification columns. The cleaned columns were converted into appropriate numeric types. After this process, pcv, wc, and rc now had numeric data types, with erroneous values removed. The classification column now contained only valid values: ckd and notckd.

Exploratory Data Analysis (EDA)

The EDA process involved visualizing the distribution of the target variable (`classification`) and analyzing the relationships between health metrics and CKD.

Target Variable Distribution

The distribution of the target variable `classification` shows more patients with CKD than without CKD. A count plot was used to visualize this distribution as shown in

Age, Blood Pressure, and Packed Cell Volume Analysis

Box plots were used to analyse the distribution of age, blood pressure, and packed cell volume (PCV) based on CKD classification. These plots revealed that patients with CKD tend to have different distributions for these metrics compared to those without CKD.

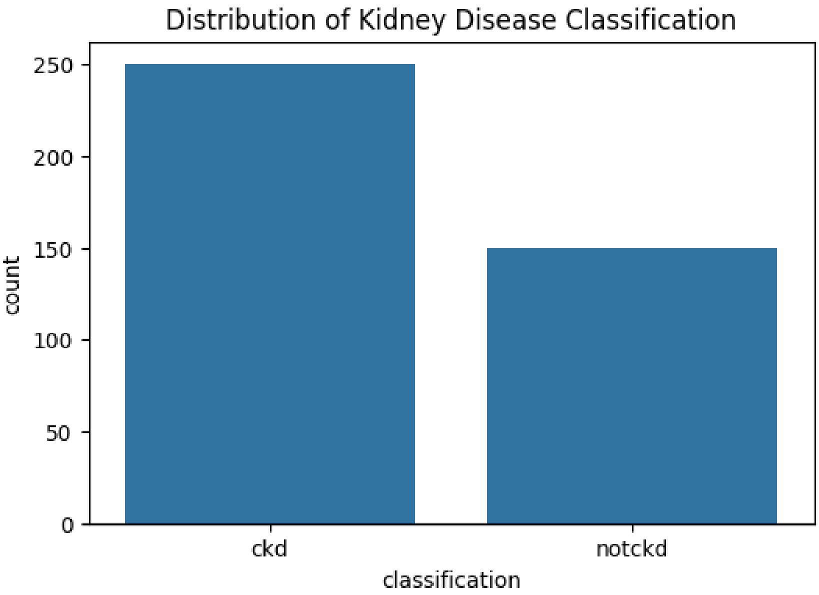


Figure 1: Distribution of the target class

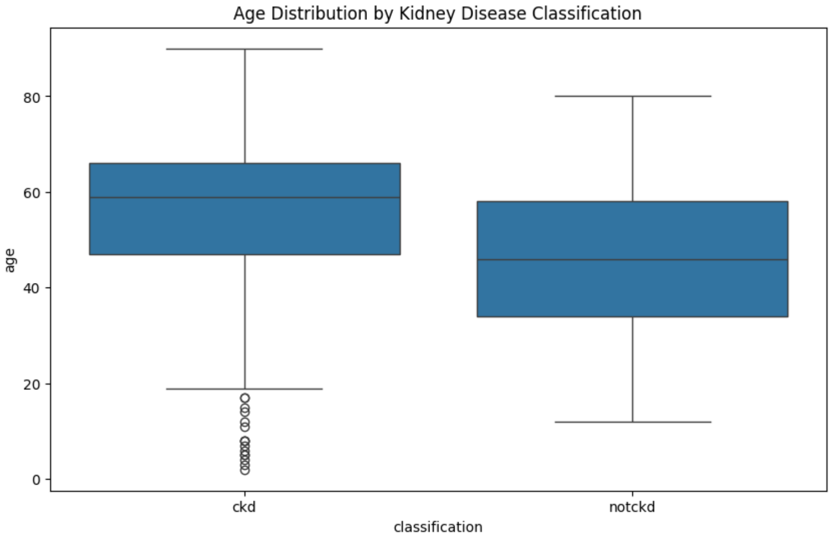


Figure 2: Distribution of Kidney Disease Classification by age

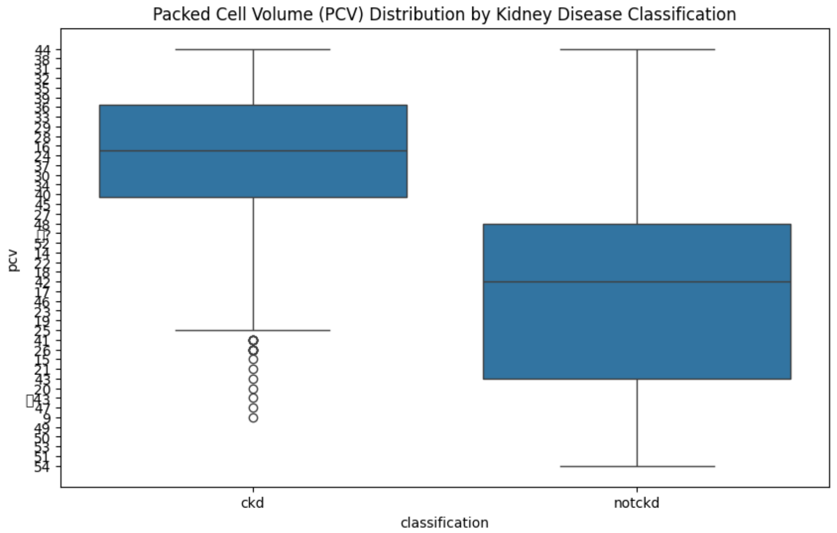


Figure 3: Age Distribution by kidney Disease Classification

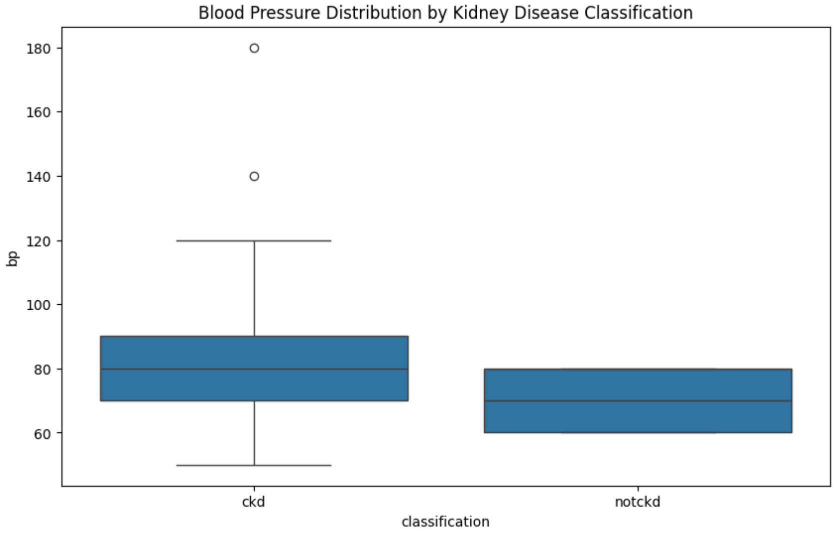


Figure 4: Packed cell Volume Distribution by Kidney Disease Classification

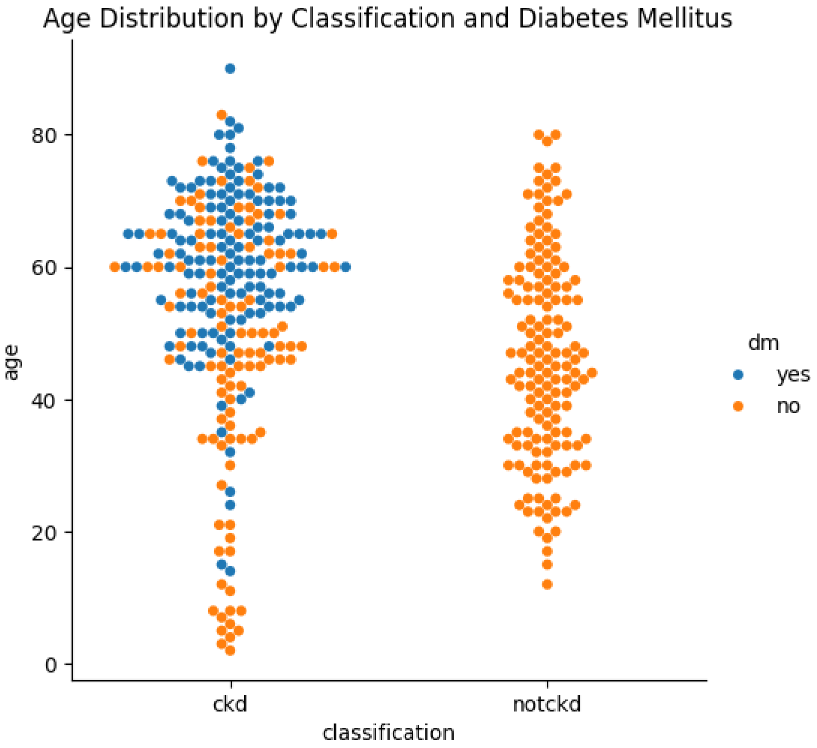


Figure 5: Age Distribution by Classification and Diabetes Mellitus

Analyzing Categorical Variables

I also explored the relationships between categorical variables such as `dm` (Diabetes Mellitus), `htn` (Hypertension), and `cad` (Coronary Artery Disease) with CKD classification. Swarm plots were used to visualize how these factors differ between CKD and non-CKD patients.

From Figure 5, it was observed that:

- For all cases where CKD = 0, there is no Diabetes Mellitus (i.e., `dm` is `no`).
- For all cases where CKD = 1, `dm` can be either `yes` or `no`.

From Figure 6, it was observed that:

- For all cases where CKD = 0, there is no Hypertension (i.e., `htn` is `no`).
- For all cases where CKD = 1, `htn` can be either `yes` or `no`.

From Figure 7, it was observed that:

- For all cases where CKD = 0, there is no Coronary Artery Disease (i.e., `cad` is `no`).
- For all cases where CKD = 1, `cad` can be either `yes` or `no`.

Correlation Matrix

The correlation matrix was plotted to show the relationships between numeric variables in the dataset. I selected only numeric columns from the dataset to calculate the correlation matrix, excluding non-numeric columns like categorical data (e.g., ”normal”).

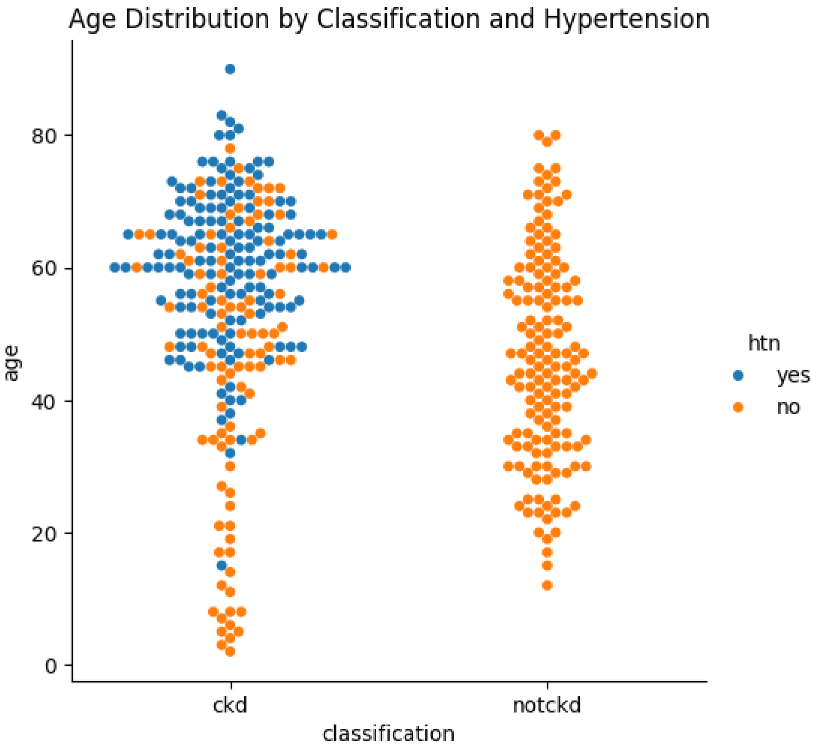


Figure 6: Age Distribution by Classification and Hypertension

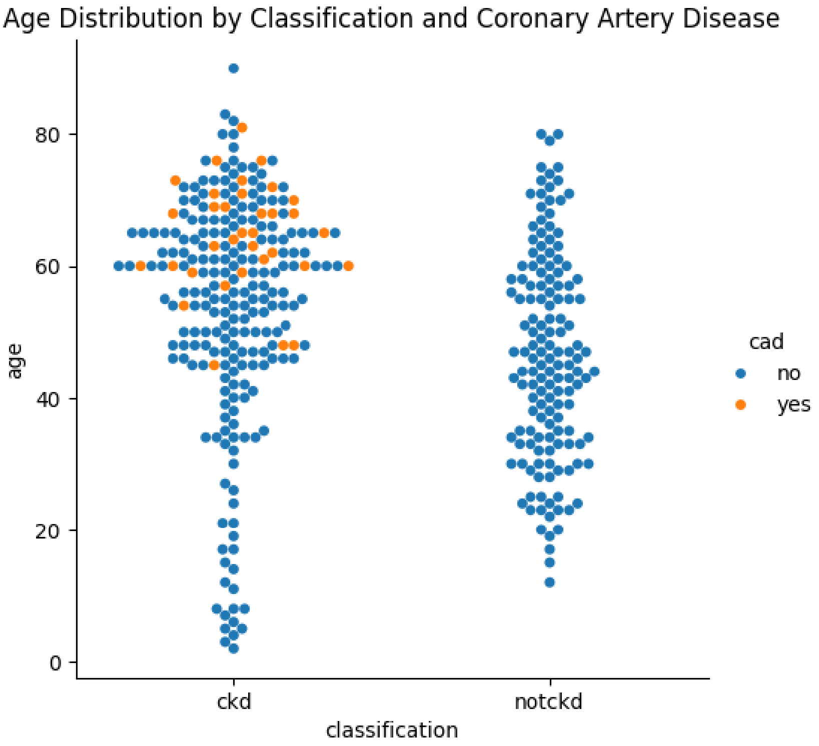
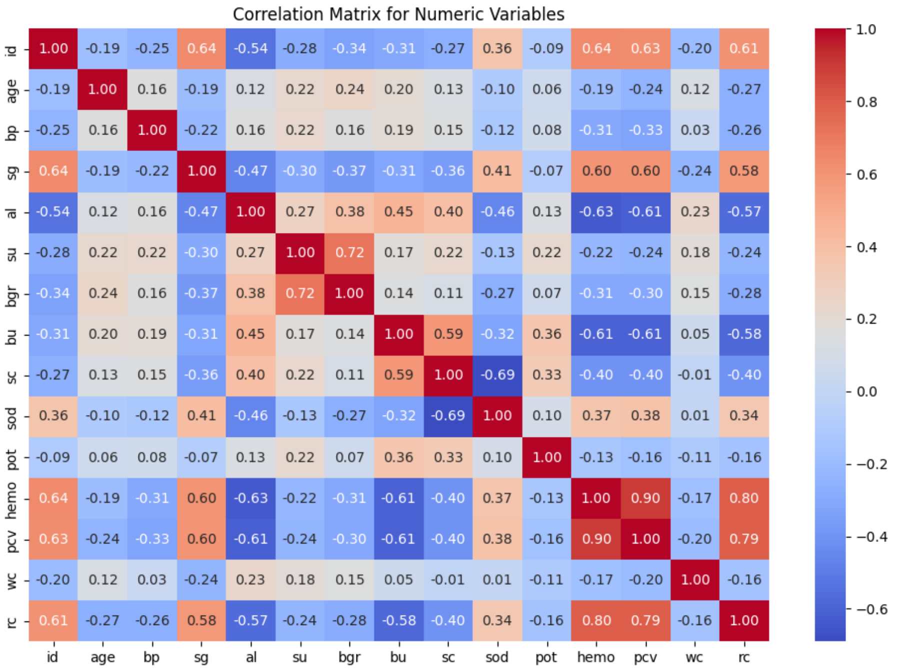


Figure 7: Age Distribution by Classification and Coronary Artery Disease



	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc
id	1.00	-0.19	-0.25	0.64	-0.54	-0.28	-0.34	-0.31	-0.27	0.36	-0.09	0.64	0.63	-0.20	0.61
age	-0.19	1.00	0.16	-0.19	0.12	0.22	0.24	0.20	0.13	-0.10	0.06	-0.19	-0.24	0.12	-0.27
bp	-0.25	0.16	1.00	-0.22	0.16	0.22	0.16	0.19	0.15	-0.12	0.08	-0.31	-0.33	0.03	-0.26
sg	0.64	-0.19	-0.22	1.00	-0.47	-0.30	-0.37	-0.31	-0.36	0.41	-0.07	0.60	0.60	-0.24	0.58
al	-0.54	0.12	0.16	-0.47	1.00	0.27	0.38	0.45	0.40	-0.46	0.13	-0.63	-0.61	0.23	-0.57
su	-0.28	0.22	0.22	-0.30	0.27	1.00	0.72	0.17	0.22	-0.13	0.22	-0.22	-0.24	0.18	-0.24
bgr	-0.34	0.24	0.16	-0.37	0.38	0.72	1.00	0.14	0.11	-0.27	0.07	-0.31	-0.30	0.15	-0.28
bu	-0.31	0.20	0.19	-0.31	0.45	0.17	0.14	1.00	0.59	-0.32	0.36	-0.61	-0.61	0.05	-0.58
sc	-0.27	0.13	0.15	-0.36	0.40	0.22	0.11	0.59	1.00	-0.69	0.33	-0.40	-0.40	-0.01	-0.40
sod	0.36	-0.10	-0.12	0.41	-0.46	-0.13	-0.27	-0.32	-0.69	1.00	0.10	0.37	0.38	0.01	0.34
pot	-0.09	0.06	0.08	-0.07	0.13	0.22	0.07	0.36	0.33	0.10	1.00	-0.13	-0.16	-0.11	-0.16
hemo	0.64	-0.19	-0.31	0.60	-0.63	-0.22	-0.31	-0.61	-0.40	0.37	-0.13	1.00	0.90	-0.17	0.80
pcv	0.63	-0.24	-0.33	0.60	-0.61	-0.24	-0.30	-0.61	-0.40	0.38	-0.16	0.90	1.00	-0.20	0.79
wc	-0.20	0.12	0.03	-0.24	0.23	0.18	0.15	0.05	-0.01	0.01	-0.11	-0.17	-0.20	1.00	-0.16
rc	0.61	-0.27	-0.26	0.58	-0.57	-0.24	-0.28	-0.58	-0.40	0.34	-0.16	0.80	0.79	-0.16	1.00

Figure 8: Correlation matrix

Highly Correlated Variables

Here are the pairs of numeric variables with high correlations (above 0.7 or below -0.7):

- **Blood Glucose Random (bgr) and Sugar (su):** Correlation = 0.72
  A strong positive correlation, which is expected as both are measures related to blood sugar levels.

- **Packed Cell Volume (pcv) and Hemoglobin (hemo):** Correlation = 0.90
  A very strong positive correlation, which makes sense as hemoglobin is a component of red blood cells, and packed cell volume measures the proportion of blood occupied by red blood cells.

- **Red Blood Cell Count (rc) and Hemoglobin (hemo):** Correlation = 0.80
  Another strong positive correlation, since red blood cells contain hemoglobin.

- **Red Blood Cell Count (rc) and Packed Cell Volume (pcv):** Correlation = 0.79
  This strong positive correlation is biologically logical, as both variables measure aspects of red blood cell function and volume.

These strong correlations indicate that some of these features may be redundant in terms of the information they provide. This could guide feature selection, particularly if you want to reduce multicollinearity in a model.

Missing Data Imputation

Missing data was handled through imputation:

- **Numerical Columns:** Missing values were imputed using the median to prevent skewing the data.
- **Categorical Columns:** Missing values were filled with the mode, i.e. the most frequent category.

I checked the data after imputation to ensure there were no missing values.

Conclusions

From the EDA process, several insights were gained:

- Patients with diabetes and hypertension show different age distributions in CKD classification, indicating a higher risk for these groups.
- Certain variables such as `pcv` and `hemo` are highly correlated, indicating potential redundancy.
- The dataset is slightly imbalanced, with more CKD cases than non-CKD.