STATS 3DA3

Project Chronic Kidney Disease Classification Challenge

Group 3
Xiangdong Wang (400335790)
Student Name (Student ID)
Student Name (Student ID)

2024-04-18

1. Classification Problem Identification

Dataset is used from the Early Stage of Indians Chronic Kidney Disease (CKD) project, which comprises data on 250 early-stage CKD patients and 150 healthy controls.
In this assignment, machine learning (ML) techniques have been deployed to predict, diagnose, and treat chronic kidney disease (CKD).

```
pip install ucimlrepo
```

```
Requirement already satisfied: ucimlrepo in /Library/Frameworks/Python.framework/Versions/3.11/
Note: you may need to restart the kernel to use updated packages.
```

```
# import dataset
from ucimlrepo import fetch_ucirepo


# fetch dataset
chronic_kidney_disease = fetch_ucirepo(id=336)
```

```
import pandas as pd
```

```
# metadata
print(chronic_kidney_disease.metadata)
```

```
{'uci_id': 336, 'name': 'Chronic Kidney Disease', 'repository_url': 'https://archive.ics.uci.ed
```

```
data_url = 'https://archive.ics.uci.edu/static/public/336/data.csv'
df = pd.read_csv(data_url)
df.head()
```

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbc
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44.0	7800.0	5.2
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38.0	6000.0	Na

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbc
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	31.0	7500.0	Na
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	32.0	6700.0	3.9
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	35.0	7300.0	4.6

```
# data (as pandas dataframes)
X = chronic_kidney_disease.data.features
y = chronic_kidney_disease.data.targets
```

```
X.head()
```

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	hemo	pcv	wbc
0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	15.4	44.0	780
1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	11.3	38.0	600
2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	423.0	...	9.6	31.0	750
3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	117.0	...	11.2	32.0	670
4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	106.0	...	11.6	35.0	730

```
y.head()
```

	class
0	ckd
1	ckd
2	ckd
3	ckd
4	ckd

The classification problem for assginemnt is to determine whether a patient has early-stage CKD based on various medical measurements included in the dataset.

2. Variable Transformation

```
data_url = 'https://archive.ics.uci.edu/static/public/336/data.csv'
chronic_kidney_disease_df = pd.read_csv(data_url)
chronic_kidney_disease_df.dtypes
```

From the dictionary, variables `sg`, `al`, `su` are Categorical

`age`, `bp`, `bgr`, `bu`, `sod`, `pcv`, `wbcc` are Integer

`rbc`, `pc`, `pcc`, `ba`, `htn`, `dm`, `cad`, `appet`, `pe`, `ane`, `class` are Binary

therefore we need to do transformation