STATS 3DA3

Project Chronic Kidney Disease Classification Challenge

Group 3
Xiangdong Wang (400335790)
Lingyun Huang (Student ID)
Jingyang Li (Student ID)

2024-04-18

```
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.
```

```
from ucimlrepo import fetch_ucirepo
import pandas as pd
```

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).

```
## Load Dataset
data_url = 'https://archive.ics.uci.edu/static/public/336/data.csv'
df = pd.read_csv(data_url)
df.head(2)
```

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wbcc	rbcc	ht
0	48.0	80.0	1.02	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	44.0	7800.0	5.2	ye
1	7.0	50.0	1.02	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	38.0	6000.0	NaN	no

```
# fetch dataset
chronic_kidney_disease = fetch_ucirepo(id=336)
# metadata
print(chronic_kidney_disease.metadata)
```

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

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

```
# Features
X.head(2)
```

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	hemo	pcv	wbcc	r
0	48.0	80.0	1.02	1.0	0.0	NaN	normal	notpresent	notpresent	121.0	...	15.4	44.0	7800.0	5
1	7.0	50.0	1.02	4.0	0.0	NaN	normal	notpresent	notpresent	NaN	...	11.3	38.0	6000.0	N

```
# Target
y.head(2)
```

	class
0	ckd
1	ckd

The classification problem is determining whether a patient has early-stage CKD based on various medical measurements included in the dataset. There are two classes here: Early-stage Indian CKD patients and Healthy patients.

2. Variable Transformation

```
df.dtypes
```

```
age      float64
bp       float64
sg       float64
al       float64
```

```
su       float64
rbc       object
pc        object
pcc       object
ba        object
bgr      float64
bu       float64
sc       float64
sod      float64
pot      float64
hemo     float64
pcv      float64
wbcc     float64
rbcc     float64
htn       object
dm        object
cad       object
appet     object
pe        object
ane       object
class     object
dtype: object
```

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

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

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

Therefore, we need to transform the data types.