BAN210 Predictive Analytics

Workshop 1 ¶

1. Download “Bank.csv” dataset and load it as ‘data’. (1 marks)

In [1]:
```
import pandas as pd
```

In [3]:
```
data = pd.read_csv("/content/bank.csv")
```

2. Display the first three rows in this dataset. (1 marks)

In [4]:
```
data.head(3)
```

Out[4]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr

3. Display the shape of data set. (1 marks)

In [5]:
```
data.shape
```

Out[5]:
```
(4521, 17)
```

4. Check the duplicate records. If you have duplicate records, please remove the duplicate records. (1 marks)

In [7]:
```
data.duplicated()
```

Out[7]:
```
0       False
1       False
2       False
3       False
4       False
        ...
4516    False
4517    False
4518    False
4519    False
4520    False
Length: 4521, dtype: bool
```

5. Check the null value in the data set. If the data set contains the null value, please replace the null values with appropriate value. (1 marks)

In [8]:
```
data.isnull()
```

Out[8]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
4516	False	False	False	False	False	False	False	False	False	False	False
4517	False	False	False	False	False	False	False	False	False	False	False
4518	False	False	False	False	False	False	False	False	False	False	False
4519	False	False	False	False	False	False	False	False	False	False	False
4520	False	False	False	False	False	False	False	False	False	False	False

4521 rows × 17 columns

In [10]:

```
data.isnull().sum()
#there are no null values in the dataset.
```

Out[10]:

```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
dtype: int64
```

6. Check the Data type of each attribute if its not correct, please modify the data type of the attribute. (1 marks)

In [11]:

```
data.dtypes
```

Out[11]:

```
age           int64
job          object
marital      object
education    object
default      object
balance       int64
housing      object
loan         object
contact      object
day           int64
month        object
duration      int64
campaign      int64
pdays         int64
previous      int64
poutcome     object
y            object
dtype: object
```

In [12]:
```
data.head(5)
```

Out[12]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may

7. Print the descriptive statistics of the admission data to understand the data a little better (min, max, mean, median, 1st and 3rd quartiles). (2 marks)

In [13]:
```
data.describe()
```

Out[13]:

	age	balance	day	duration	campaign	pdays	previ
count	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000000	4521.000
mean	41.170095	1422.657819	15.915284	263.961292	2.793630	39.766645	0.542
std	10.576211	3009.638142	8.247667	259.856633	3.109807	100.121124	1.693
min	19.000000	-3313.000000	1.000000	4.000000	1.000000	-1.000000	0.000
25%	33.000000	69.000000	9.000000	104.000000	1.000000	-1.000000	0.000
50%	39.000000	444.000000	16.000000	185.000000	2.000000	-1.000000	0.000
75%	49.000000	1480.000000	21.000000	329.000000	3.000000	-1.000000	0.000
max	87.000000	71188.000000	31.000000	3025.000000	50.000000	871.000000	25.000

8. Convert the categorical variable to a numeric variable using the one hot encoding method. (2 marks)

In [17]:
```
df = data
```

In [19]:
```
df.head(2)
```

Out[19]:

	age	job	marital	education	default	balance	housing	loan	contact	day	month
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may

In [20]:
```
dfx =pd.get_dummies(dfx)
```

In [21]:
```
dfx.head(2)
```

Out[21]:

	age	balance	day	duration	campaign	pdays	previous	job_admin.	job_blue-collar	job_entrepre
0	30	1787	19	79	1	-1	0	0	0	
1	33	4789	11	220	1	339	4	0	0	

2 rows × 53 columns

This is the end of Workshop 1

Savita Seharawat, PhD

In [ ]: