Forward
School

## Program Code: J620-002-4:2020

## Program Name: FRONT-END SOFTWARE DEVELOPMENT

## Title : Exe19 - Decision Tree Exercise 1

#### Name: Ooi Caaron

#### IC Number: 990701-07-5837

#### Date :21/7/23

#### Introduction : Decision Tree algorithm partitions the data into subsets by repeatedly asking questions about the features of the data points.

#### Conclusion : Still need to practice more and do revision

Section 1

Reference: https://www.kaggle.com/vinicius150987/bank-full-machine-learning/notebook (https://www.kaggle.com/vinicius150987/bank-full-machine-learning/notebook)

Decision Tree

In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
from sklearn.model_selection import train_test_split # Import train_test_split function
from sklearn import metrics, tree #Import scikit-learn metrics module for accuracy calcu
```

Read "bank-full.csv"

In [2]:

```
df = pd.read_csv("bank-full.csv", delimiter=';')
df
```

Out[2]:

	age	job	marital	education	default	balance	housing	loan	contact	day
0	58	management	married	tertiary	no	2143	yes	no	unknown	5
1	44	technician	single	secondary	no	29	yes	no	unknown	5
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5
4	33	unknown	single	unknown	no	1	no	no	unknown	5
...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	married	tertiary	no	825	no	no	cellular	17
45207	71	retired	divorced	primary	no	1729	no	no	cellular	17
45208	72	retired	married	secondary	no	5715	no	no	cellular	17
45209	57	blue-collar	married	secondary	no	668	no	no	telephone	17
45210	37	entrepreneur	married	secondary	no	2971	no	no	cellular	17

45211 rows × 17 columns

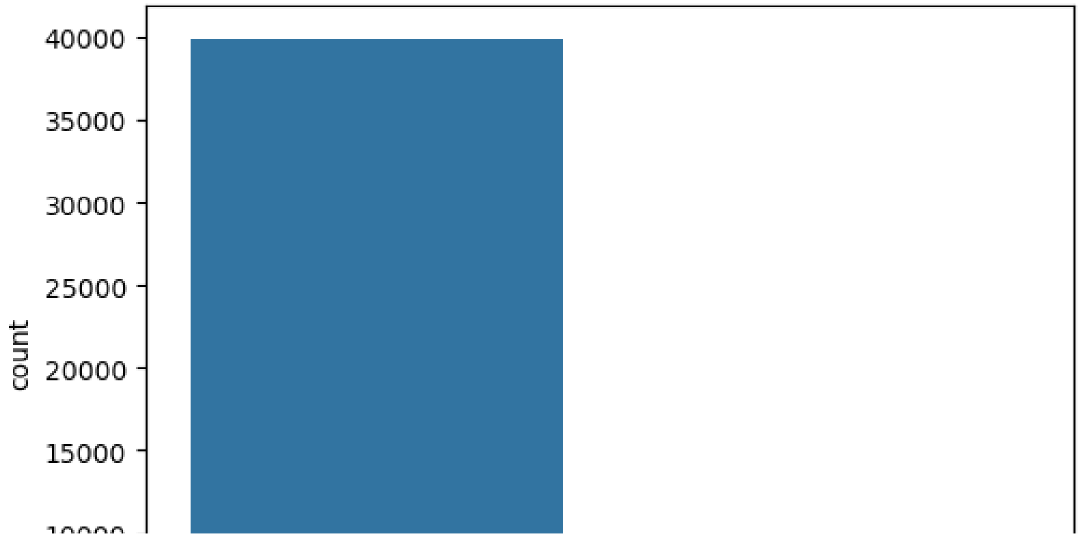
Check the distribution of labels (‘yes’, ‘no’) are distributed.

In [43]:

```
sns.countplot(x=df['y'], data=df)
```

Out[43]:

```
<Axes: xlabel='y', ylabel='count'>
```



In [3]:

```
yes = df[df['y'] == 'yes']
no = df[df['y'] == 'no']
yes.head()
```

Out[3]:

	age	job	marital	education	default	balance	housing	loan	contact	day	mor
83	59	admin.	married	secondary	no	2343	yes	no	unknown	5	m
86	56	admin.	married	secondary	no	45	no	no	unknown	5	m
87	41	technician	married	secondary	no	1270	yes	no	unknown	5	m
129	55	services	married	secondary	no	2476	yes	no	unknown	5	m
168	54	admin.	married	tertiary	no	184	no	no	unknown	5	m

In [4]:

```
no.head()
```

Out[4]:

	age	job	marital	education	default	balance	housing	loan	contact	day	mo
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	n
1	44	technician	single	secondary	no	29	yes	no	unknown	5	n
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	n
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	n
4	33	unknown	single	unknown	no	1	no	no	unknown	5	n

Counts of "yes" and "no" with "age"

In [5]:

```
age = df.groupby(['y', 'age']).size().reset_index(name='Count')
age
```

Out[5]:

	y	age	Count
0	no	18	5
1	no	19	24
2	no	20	35
3	no	21	57
4	no	22	89
...	...	...	...
143	yes	87	3
144	yes	90	2
145	yes	92	2
146	yes	93	2
147	yes	95	1

148 rows × 3 columns

Correlation between the data

In [6]:

```
cor = df.corr()
cor
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_17500\3057578215.py:1: FutureWa
rning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns o
r specify the value of numeric_only to silence this warning.
  cor = df.corr()
```

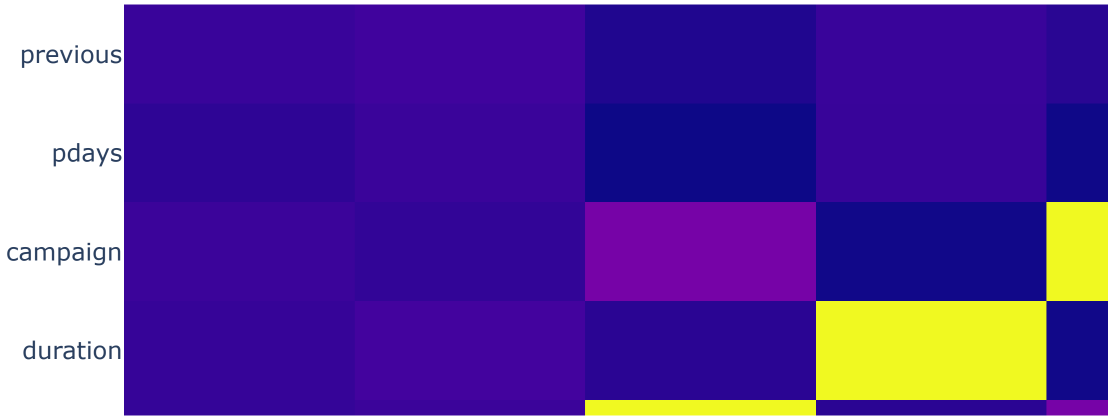
Out[6]:

	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.097783	-0.009120	-0.004648	0.004760	-0.023758	0.001288
balance	0.097783	1.000000	0.004503	0.021560	-0.014578	0.003435	0.016674
day	-0.009120	0.004503	1.000000	-0.030206	0.162490	-0.093044	-0.051710
duration	-0.004648	0.021560	-0.030206	1.000000	-0.084570	-0.001565	0.001203
campaign	0.004760	-0.014578	0.162490	-0.084570	1.000000	-0.088628	-0.032855
pdays	-0.023758	0.003435	-0.093044	-0.001565	-0.088628	1.000000	0.454820
previous	0.001288	0.016674	-0.051710	0.001203	-0.032855	0.454820	1.000000

Plot the heatmap

In [7]:

```
import plotly.graph_objects as go
from plotly.offline import plot
fig = go.Figure(data=go.Heatmap(
    z=cor.values,
    x=cor.columns,
    y=cor.columns
))

fig.show()
```



Convert categorical data into numerical

In [9]:

```
reset = {'yes':1, 'no':2}
df = df.replace({'y':reset})
df = df.replace({'loan':reset})
df = df.replace({'housing':reset})
df = df.replace({'marital':{'single':1, 'married':2,'divorced':3}})
df = df.replace({'contact':{'unknown':None, 'telephone':1,'cellular':2}})
df = df.replace({'education':{'unknown':None, 'primary':1,'secondary':2,'tertiary':3}})
df = df.dropna()
df
```

Out[9]:

	age	job	marital	education	default	balance	housing	loan	contact	day
12657	27	management	1	2.0	no	35	2	2	2.0	4
12658	54	blue-collar	2	1.0	no	466	2	2	2.0	4
12659	43	blue-collar	2	2.0	no	105	2	1	2.0	4
12660	31	technician	1	2.0	no	19	2	2	1.0	4
12661	27	technician	1	2.0	no	126	1	1	2.0	4
...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	2	3.0	no	825	2	2	2.0	17
45207	71	retired	3	1.0	no	1729	2	2	2.0	17
45208	72	retired	2	2.0	no	5715	2	2	2.0	17
45209	57	blue-collar	2	2.0	no	668	2	2	1.0	17
45210	37	entrepreneur	2	2.0	no	2971	2	2	2.0	17

31011 rows × 17 columns

Next step is to select features and labels

In [10]:

```
feature_cols = ['marital', 'education', 'housing', 'loan','balance','duration','campaign
X = df[feature_cols] # Features
y = df.y# Target variable
```

Drop "poutcome"

In [11]:

```
df.drop('poutcome', axis=1)
```

Out[11]:

	age	job	marital	education	default	balance	housing	loan	contact	day
12657	27	management	1	2.0	no	35	2	2	2.0	4
12658	54	blue-collar	2	1.0	no	466	2	2	2.0	4
12659	43	blue-collar	2	2.0	no	105	2	1	2.0	4
12660	31	technician	1	2.0	no	19	2	2	1.0	4
12661	27	technician	1	2.0	no	126	1	1	2.0	4
...	...	...	...	...	...	...	...	...	...	...
45206	51	technician	2	3.0	no	825	2	2	2.0	17
45207	71	retired	3	1.0	no	1729	2	2	2.0	17
45208	72	retired	2	2.0	no	5715	2	2	2.0	17
45209	57	blue-collar	2	2.0	no	668	2	2	1.0	17
45210	37	entrepreneur	2	2.0	no	2971	2	2	2.0	17

31011 rows × 16 columns

Split the data into train and test

In [12]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Applying Decision Tree Classifier:

Next, I created a pipeline of StandardScaler (standardize the features) and DT Classifier (see a note below regarding Standardization of features). We can import DT classifier as from sklearn.tree import DecisionTreeClassifier from Scikit-Learn. To determine the best parameters (criterion of split and maximum tree depth) for DT classifier, I also used Grid Search Cross Validation. The code snippet below is self-explanatory.

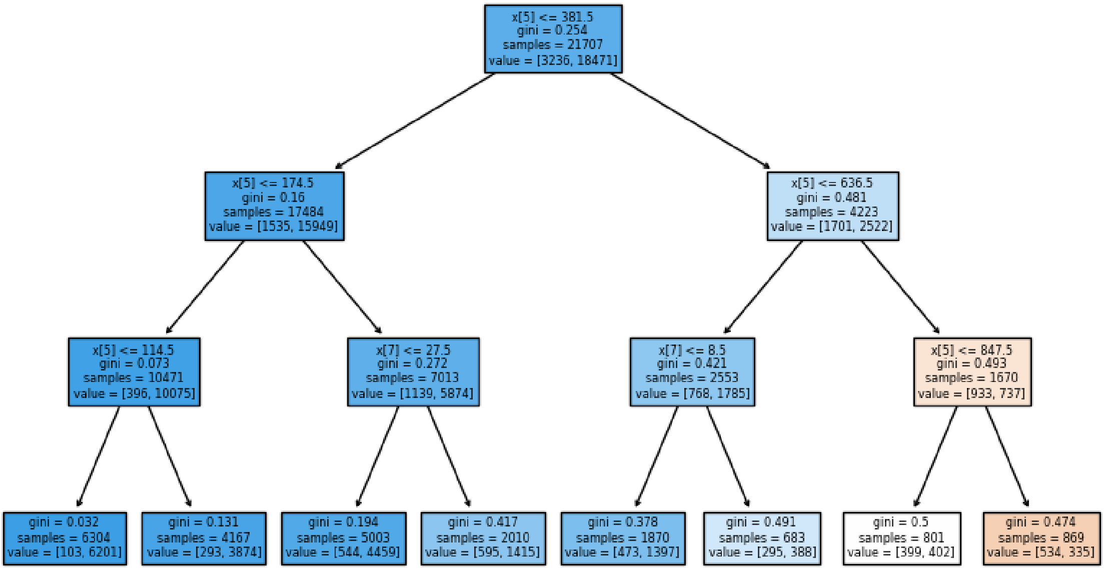
In [13]:

```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier(max_depth = 3)

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

To display

In [14]:

```
plt.figure(figsize=(10,6))
tree.plot_tree(clf, filled=True)
plt.show()
```



The number of nodes and the maximum depth

In [15]:

```
print(clf.tree_.node_count, clf.tree_.max_depth)
```

```
15 3
```

Accuracy measurement

In [16]:

```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8696259673258814
```

Prediction

In [18]:

```
y_pred = clf.predict(X_test)
pd.DataFrame({'Predicted':y_pred})
```

Out[18]:

	Predicted
0	2
1	2
2	2
3	2
4	2
...	...
9299	2
9300	2
9301	2
9302	2
9303	2

9304 rows × 1 columns

Grid Search

In [58]:

```
from sklearn.metrics import confusion_matrix
print(str(confusion_matrix(y_test, y_pred)))
```

```
[[ 196 1096]
 [ 117 7895]]
```

Display the best features

In [60]:

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=100)
classifier.fit(X_train, y_train)

feature_importances_df = pd.DataFrame(
    {"feature": list(X.columns), "importances": classifier.feature_importances_}
).sort_values("importances", ascending=False)

feature_importances_df
```

Out[60]:

	feature	importances
5	duration	0.428659
4	balance	0.256204
7	pdays	0.098139
6	campaign	0.057552
8	previous	0.040699
2	housing	0.035376
1	education	0.034884
0	marital	0.032886
3	loan	0.015602

Run DecisionTreeClassifier using the obtained features

In [ ]:

```
optimized_c = Decision
```

Concat train test results

In [61]:

```
metrics.accuracy_score(y2_test, y2_pred)
```

```
---------------------------------------------------------------------------
-
NameError                                 Traceback (most recent call las
t)
Cell In[61], line 1
----> 1 metrics.accuracy_score(y2_test, y2_pred)

NameError: name 'y2_test' is not defined
```

Section 2

1. Read "petrol_consumption.csv" file

In [62]:

```
df = pd.read_csv("petrol_consumption.csv")
df
```

Out[62]:

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Consu
0	9.00	3571	1976	0.525	
1	9.00	4092	1250	0.572	
2	9.00	3865	1586	0.580	
3	7.50	4870	2351	0.529	
4	8.00	4399	431	0.544	
5	10.00	5342	1333	0.571	
6	8.00	5319	11868	0.451	
7	8.00	5126	2138	0.553	
8	8.00	4447	8577	0.529	
9	7.00	4512	8507	0.552	
10	8.00	4391	5939	0.530	
11	7.50	5126	14186	0.525	
12	7.00	4817	6930	0.574	
13	7.00	4207	6580	0.545	
14	7.00	4332	8159	0.608	
15	7.00	4318	10340	0.586	
16	7.00	4206	8508	0.572	
17	7.00	3718	4725	0.540	
18	7.00	4716	5915	0.724	
19	8.50	4341	6010	0.677	
20	7.00	4593	7834	0.663	
21	8.00	4983	602	0.602	
22	9.00	4897	2449	0.511	
23	9.00	4258	4686	0.517	
24	8.50	4574	2619	0.551	
25	9.00	3721	4746	0.544	
26	8.00	3448	5399	0.548	
27	7.50	3846	9061	0.579	
28	8.00	4188	5975	0.563	
29	9.00	3601	4650	0.493	
30	7.00	3640	6905	0.518	
31	7.00	3333	6594	0.513	
32	8.00	3063	6524	0.578	
33	7.50	3357	4121	0.547	
34	8.00	3528	3495	0.487	
35	6.58	3802	7834	0.629	
36	5.00	4045	17782	0.566	

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Consu
37	7.00	3897	6385	0.586	
38	8.50	3635	3274	0.663	
39	7.00	4345	3905	0.672	
40	7.00	4449	4639	0.626	
41	7.00	3656	3985	0.563	
42	7.00	4300	3635	0.603	
43	7.00	3745	2611	0.508	
44	6.00	5215	2302	0.672	
45	9.00	4476	3942	0.571	
46	7.00	4296	4083	0.623	
47	7.00	5002	9794	0.593	

2. Display the first 5 records

In [63]:

```
df.head()
```

Out[63]:

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Consun
0	9.0	3571	1976	0.525	
1	9.0	4092	1250	0.572	
2	9.0	3865	1586	0.580	
3	7.5	4870	2351	0.529	
4	8.0	4399	431	0.544	

4. Identify the label (Petrol_Consumption)

In [64]:

```
y = df['Petrol_Consumption']
y
```

Out[64]:

```
0     541
1     524
2     561
3     414
4     410
5     457
6     344
7     467
8     464
9     498
10    580
11    471
12    525
13    508
14    566
15    635
16    603
17    714
18    865
19    640
20    649
21    540
22    464
23    547
24    460
25    566
26    577
27    631
28    574
29    534
30    571
31    554
32    577
33    628
34    487
35    644
36    640
37    704
38    648
39    968
40    587
41    699
42    632
43    591
44    782
45    510
46    610
47    524
Name: Petrol_Consumption, dtype: int64
```

5. Identify the features.

In [69]:

```
X = df.drop('Petrol_Consumption', axis=1)
X
```

Out[69]:

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)
0	9.00	3571	1976	0.525
1	9.00	4092	1250	0.572
2	9.00	3865	1586	0.580
3	7.50	4870	2351	0.529
4	8.00	4399	431	0.544
5	10.00	5342	1333	0.571
6	8.00	5319	11868	0.451
7	8.00	5126	2138	0.553
8	8.00	4447	8577	0.529
9	7.00	4512	8507	0.552
10	8.00	4391	5939	0.530
11	7.50	5126	14186	0.525
12	7.00	4817	6930	0.574
13	7.00	4207	6580	0.545
14	7.00	4332	8159	0.608
15	7.00	4318	10340	0.586
16	7.00	4206	8508	0.572
17	7.00	3718	4725	0.540
18	7.00	4716	5915	0.724
19	8.50	4341	6010	0.677
20	7.00	4593	7834	0.663
21	8.00	4983	602	0.602
22	9.00	4897	2449	0.511
23	9.00	4258	4686	0.517
24	8.50	4574	2619	0.551
25	9.00	3721	4746	0.544
26	8.00	3448	5399	0.548
27	7.50	3846	9061	0.579
28	8.00	4188	5975	0.563
29	9.00	3601	4650	0.493
30	7.00	3640	6905	0.518
31	7.00	3333	6594	0.513
32	8.00	3063	6524	0.578
33	7.50	3357	4121	0.547
34	8.00	3528	3495	0.487
35	6.58	3802	7834	0.629
36	5.00	4045	17782	0.566

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)
37	7.00	3897	6385	0.586
38	8.50	3635	3274	0.663
39	7.00	4345	3905	0.672
40	7.00	4449	4639	0.626
41	7.00	3656	3985	0.563
42	7.00	4300	3635	0.603
43	7.00	3745	2611	0.508
44	6.00	5215	2302	0.672
45	9.00	4476	3942	0.571
46	7.00	4296	4083	0.623
47	7.00	5002	9794	0.593

6. Use of describe method to describe the dataset.

In [70]:

```
df.describe()
```

Out[70]:

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)	Petrol_Co
count	48.000000	48.000000	48.000000	48.000000	
mean	7.668333	4241.833333	5565.416667	0.570333	
std	0.950770	573.623768	3491.507166	0.055470	
min	5.000000	3063.000000	431.000000	0.451000	
25%	7.000000	3739.000000	3110.250000	0.529750	
50%	7.500000	4298.000000	4735.500000	0.564500	
75%	8.125000	4578.750000	7156.000000	0.595250	
max	10.000000	5342.000000	17782.000000	0.724000	

7. Display the first 5 records of the features

In [71]:

```
X.head()
```

Out[71]:

	Petrol_tax	Average_income	Paved_Highways	Population_Driver_licence(%)
0	9.0	3571	1976	0.525
1	9.0	4092	1250	0.572
2	9.0	3865	1586	0.580
3	7.5	4870	2351	0.529
4	8.0	4399	431	0.544

8. Split the data into training (80%) and testing (20%) sets.

In [73]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

9. Build your model and train the training data

In [75]:

```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier(max_depth = 3)

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

10. Prediction using the testing set

In [76]:

```
clf.predict(X_test)
```

Out[76]:

```
array([577, 464, 644, 524, 524, 464, 534, 577, 587, 524], dtype=int64)
```

11. Display Actual and Predictied price side by side in df

In [78]:

```
 compare_df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred}).reset_index(drop=Tru
compare_df
```

Out[78]:

	Actual	Predicted
0	628	577
1	547	464
2	648	644
3	640	524
4	561	524
5	414	464
6	554	534
7	577	577
8	782	587
9	631	524

12. Evaluate the model using mean_absulate_error

In [80]:

```
from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(y_test, y_pred)
mae
```

Out[80]:

```
66.3
```

13. Display the predicted output using first 5 features.

In [81]:

```
clf.predict(X[:5])
```

Out[81]:

```
array([534, 524, 524, 464, 524], dtype=int64)
```

In [ ]:

```

```