Bài toán: Dự đoán khả năng trả nợ của khách hàng vay vốn tại ngân hàng

Bước 1: Khai báo thư viện

In [125]:
```
import pandas as pd
from sklearn import tree
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Bước 2: Đọc dữ liệu

In [126]:
```
df = pd.read_csv("Chuong 4. bank_data.csv")
```

Bước 3: Hiểu dữ liệu

In [127]:
```
# Xem 5 dòng dữ liệu đầu tiên
df.head()
```

Out[127]:

	ID	Tuoi	Gioi_Tinh	Khu_Vuc	Thu_Nhap	Ket_Hon	So_Con	O_To	TK_Tiet_Kiem	TK_Thanh_
0	1	48	Nu	Thanh Pho	17546.0	No	1	No	No	
1	2	40	Nam	Thi Tran	30085.1	Yes	3	Yes	No	
2	3	51	Nu	Thanh Pho	16575.4	Yes	0	Yes	Yes	
3	4	23	Nu	Thi Tran	20375.4	Yes	3	No	No	
4	5	57	Nu	Nong Thon	50576.3	Yes	0	No	Yes	

In [128]:
```
# Xem 5 dòng dữ liệu cuối cùng
df.tail()
```

Out[128]:

	ID	Tuoi	Gioi_Tinh	Khu_Vuc	Thu_Nhap	Ket_Hon	So_Con	O_To	TK_Tiet_Kiem	TK_Th
595	596	61	Nu	Thanh Pho	47025.00	No	2	Yes	Yes	
596	597	30	Nu	Thanh Pho	9672.25	Yes	0	Yes	Yes	
597	598	31	Nu	Thi Tran	15976.30	Yes	0	Yes	Yes	
598	599	29	Nam	Thanh Pho	14711.80	Yes	0	No	Yes	
599	600	38	Nam	Thi Tran	26671.60	No	0	Yes	No	

In [129]:
```
# Số lượng mẫu và số thuộc tính
print(df.shape)

# Số lượng mẫu
print(df.shape[0])

# số lượng thuộc tính
print(df.shape[1])
```

```
(600, 12)
600
12
```

In [130]:
```
# Thống kê các giá trị định lượng
df.describe()
```

Out[130]:

	ID	Tuoi	Thu_Nhap	So_Con
count	600.000000	600.000000	600.000000	600.000000
mean	300.500000	42.395000	27524.031217	1.011667
std	173.349358	14.424947	12899.468246	1.056752
min	1.000000	18.000000	5014.210000	0.000000
25%	150.750000	30.000000	17264.500000	0.000000
50%	300.500000	42.000000	24925.300000	1.000000
75%	450.250000	55.250000	36172.675000	2.000000
max	600.000000	67.000000	63130.100000	3.000000

In [131]:
```
# Hiển thị kiểu dữ liệu của các thuộc tính
df.dtypes
```

Out[131]:
```
ID                int64
Tuoi              int64
Gioi_Tinh        object
Khu_Vuc          object
Thu_Nhap        float64
Ket_Hon          object
So_Con            int64
O_To             object
TK_Tiet_Kiem     object
TK_Thanh_Toan    object
The_Chap         object
Tra_No           object
dtype: object
```

In [132]:
```
# Đổi dữ liệu từ dạng định danh (object) về dạng số
from sklearn.preprocessing import LabelEncoder
lb_make = LabelEncoder()
df["Gioi_Tinh"] = lb_make.fit_transform(df["Gioi_Tinh"])
df["Khu_Vuc"] = lb_make.fit_transform(df["Khu_Vuc"])
df["Ket_Hon"] = lb_make.fit_transform(df["Ket_Hon"])
df["O_To"] = lb_make.fit_transform(df["O_To"])
df["TK_Tiet_Kiem"] = lb_make.fit_transform(df["TK_Tiet_Kiem"])
df["TK_Thanh_Toan"] = lb_make.fit_transform(df["TK_Thanh_Toan"])
df["The_Chap"] = lb_make.fit_transform(df["The_Chap"])
df["Tra_No"] = lb_make.fit_transform(df["Tra_No"])
df.head(10)
```

Out[132]:

	ID	Tuoi	Gioi_Tinh	Khu_Vuc	Thu_Nhap	Ket_Hon	So_Con	O_To	TK_Tiet_Kiem	TK_Thanh
0	1	48	1	2	17546.00	0	1	0	0	
1	2	40	0	3	30085.10	1	3	1	0	
2	3	51	1	2	16575.40	1	0	1	1	
3	4	23	1	3	20375.40	1	3	0	0	
4	5	57	1	1	50576.30	1	0	0	1	
5	6	57	1	3	37869.60	1	2	0	1	
6	7	22	0	1	8877.07	0	0	0	0	
7	8	58	0	3	24946.60	1	0	1	1	
8	9	37	1	0	25304.30	1	2	1	0	
9	10	54	0	3	24212.10	1	2	1	1	

Bước 4: Xây dựng model dự đoán khả năng trả nợ của khách hàng

In [133]:
```
# Xác định thuộc tính mô tả X và thuộc tính dự đoán y
features = ['Tuoi','Gioi_Tinh','Khu_Vuc','Thu_Nhap','Ket_Hon','So_Con','O_To',
'TK_Tiet_Kiem','TK_Thanh_Toan','The_Chap']
target = ['Tra_No']
X = df[features]
y = df[target]
print(X)
print(y)
```

```
     Tuoi  Gioi_Tinh  Khu_Vuc  Thu_Nhap  Ket_Hon  So_Con  O_To  TK_Tiet_Kiem
\
0      48          1        2  17546.00        0       1     0             0
1      40          0        3  30085.10        1       3     1             0
2      51          1        2  16575.40        1       0     1             1
3      23          1        3  20375.40        1       3     0             0
4      57          1        1  50576.30        1       0     0             1
..    ...        ...      ...       ...      ...     ...   ...           ...
595    61          1        2  47025.00        0       2     1             1
596    30          1        2   9672.25        1       0     1             1
597    31          1        3  15976.30        1       0     1             1
598    29          0        2  14711.80        1       0     0             1
599    38          0        3  26671.60        0       0     1             0

     TK_Thanh_Toan  The_Chap
0                0         0
1                1         1
2                1         0
3                1         0
4                0         0
..             ...       ...
595              1         1
596              1         0
597              0         0
598              0         1
599              1         1

[600 rows x 10 columns]
     Tra_No
0         1
1         0
2         0
3         0
4         0
..      ...
595       0
596       0
597       1
598       0
599       1

[600 rows x 1 columns]
```

In [134]:
```
# Chia bộ dữ liệu thành hai tập train và test theo tỉ lệ 80% train,20% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
print(X_train)
print(y_train)
print(X_test)
print(y_test)
```

```
     Tuoi  Gioi_Tinh  Khu_Vuc  Thu_Nhap  Ket_Hon  So_Con  O_To  TK_Tiet_Kiem
\
269    20          0        3   16672.8        0       3     1             0
152    63          0        3   54618.8        1       2     0             1
365    27          0        2   17364.8        1       2     1             1
251    45          1        3   36057.8        1       1     1             1
344    34          0        1   23638.1        1       2     1             1
..    ...        ...      ...       ...      ...     ...   ...           ...
286    39          0        2   24675.7        1       1     1             1
200    46          0        3   26077.8        1       1     1             1
416    36          0        2   12533.2        0       1     0             1
14     36          0        1   19474.6        1       0     0             1
62     47          0        2   27022.6        1       2     0             1

     TK_Thanh_Toan  The_Chap
269              1         1
152              0         1
365              0         0
251              1         1
344              1         0
..             ...       ...
286              1         1
200              1         0
416              0         1
14               1         1
62               1         0

[480 rows x 10 columns]
     Tra_No
269       0
152       1
365       1
251       1
344       0
..      ...
286       1
200       1
416       0
14        0
62        0

[480 rows x 1 columns]
     Tuoi  Gioi_Tinh  Khu_Vuc  Thu_Nhap  Ket_Hon  So_Con  O_To  TK_Tiet_Kiem
\
443    38          0        3   33302.8        0       0     1             0
345    65          1        3   42378.2        1       1     0             1
177    41          1        2   30099.3        1       0     1             1
306    63          0        2   52117.3        0       2     1             1
541    39          0        1   37389.0        1       2     0             1
..    ...        ...      ...       ...      ...     ...   ...           ...
104    64          1        2   34513.6        1       1     0             1
448    53          0        1   48971.6        1       3     1             1
430    48          0        3   28920.6        1       0     0             1
173    34          0        1   26999.4        1       1     1             1
452    59          1        2   27045.1        0       0     0             0

     TK_Thanh_Toan  The_Chap
```

```
443              1         1
345              1         0
177              1         1
306              1         0
541              1         0
..             ...       ...
104              1         0
448              0         0
430              0         1
173              1         0
452              1         0

[120 rows x 10 columns]
     Tra_No
443       1
345       1
177       1
306       1
541       1
..      ...
104       1
448       1
430       0
173       1
452       1

[120 rows x 1 columns]
```

In [135]:
```
# Khai báo mô hình cây quyết định
model=tree.DecisionTreeClassifier(criterion = "entropy",
                                  random_state = 100,
                                  max_depth = 3,
                                  min_samples_leaf = 5)
```

In [136]:
```
# Thực thi mô hình
model = model.fit(X_train,y_train)
```

Bước 5: Xác định độ chính xác của mô hình

In [137]:
```
# Kiểm thử mô hình
y_pred = model.predict(X_test)
print("Giá trị nhãn mô hình dự đoán được:\n " ,y_pred)
```

```
Giá trị nhãn mô hình dự đoán được:
  [0 1 0 1 1 0 1 0 1 0 0 0 1 0 0 1 1 1 0 1 0 0 1 1 0 1 1 0 0 0 1 1 0 0 0 1 0
 1 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 1 1
 0 0 0 1 1 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1 0 1 0 1 0
 0 0 0 0 1 1 0 1 0]
```

In [138]:
```
# Xác định ma trần nhầm lẫn
print("Confusion Matrix: \n", confusion_matrix(y_test, y_pred))
```

```
Confusion Matrix:
 [[46 13]
 [23 38]]
```

In [139]:
```
# Độ chính xác của mô hình
print ("Accuracy : ", accuracy_score(y_test,y_pred)*100)
```

```
Accuracy :  70.0
```

Bước 6: Sử dụng mô hình

In [140]:
```
# Sử dụng mô hình dự đoán khả năng trả nợ của khách hàng có các thông tin sau
# Tuoi 42, Gioi_Tinh nữ 1, Khu_Vuc thị trấn 3, Thu_Nhap 30527, có Ket_Hon 1
# So_Con 2, có O_To 1, có TK_Tiet_Kiem 1, không có TK_Thanh_Toan 0, có The_Cha
p 1

x=[[42,1,3,30527,1,2,1,1,0,1]]
y = model.predict(x)
if y==1:
    print("Khách hàng có khả năng trả nợ")
else:
    print("Khách hàng không có khả năng trả nợ")
```

```
Khách hàng có khả năng trả nợ
```

In [ ]:
```

```