BÀI TUẦN 06: SVM và Random Forest

1. Thông tin sinh viên

- **Họ và Tên: Dương Minh Lượng**
- **MSSV: 18521071**

2. Dataset

Tập dữ liệu gồm 26187 điểm dữ liệu, mỗi điểm dữ liệu gồm 23 thuộc tính:

- **ID:** Mã số định danh.
- **NUMBER_OF_INSTALMENTS:** số lần trả góp**.**
- **TYPE_OF_LEASE:** loại hình thuê xe có 2 giá trị FL và OL.
- **CAR_PRICE:** giá xe.
- **DOWNPAYMENT_TO_CAR_PRICE:** giá tiền đặt cọc xe.
- **RESIDUAL_VALUE_TO_CAR_PRICE:** giá trị tiền còn lại.
- **LEGAL_FORM_group:** nhóm hợp pháp (capital_company, personal_company, civil_partnership, freelancer, other)**.**
- **WHETHER_CUSTOMER_BEFORE:** biết khách hàng trước đó hay không (YES/NO).
- **CUSTOMER_FOR:** khách hàng mấy lần.
- **NET_ASSETS:** mạng lưới tài sản.
- **ANNUAL_TURNOVER_LAST_YEAR:** doanh thu cuối năm.
- **ANNUAL_COSTS:** chi phí 1 năm.
- **ANNUAL_INCOME_LAST_YEAR:** thu nhập cuối năm.
- **NUMBER_OF_EMPLOYEES:** số lượng nhân viên.
- **COMPANY_AGE**: tuổi công ty.
- **REGION:** khu vực.
- **PKD_group:**(sale, production, services, education, building)
- **PKD_section:** (A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U)
- **EMAIL:** (YES/NO)
- **PHONE:** (YES/NO)
- **TYPE_OF_VEHICLE:** loại hình xe cộ (LKW, PKW).
- **VEHICLE_AGE:** tuổi của xe.
- **FRAUD:** có gian lận hay không (YES/NO).

3. Source

```
4. import pandas as pd
5. # Importing the dataset
```

```
from sklearn.model_selection import ShuffleSplit
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
dataset = pd.read_csv('/content/drive/My Drive/a/mmc2_Sample.csv',delimiter=";")

      dataset.drop(dataset.columns[[0,1,2,3,4,5,7,10,11,12,13,14,15,16,18,19]], axis=1, inplace=True)
      dataset.drop(dataset[dataset['PKD_section']=='?'].index, inplace=True)
      dataset=dataset.replace("?",0)
      dataset=dataset.replace("YES",1)
      dataset=dataset.replace("NO",0)
      y = dataset.iloc[:, -1].values
      del dataset['FRAUD']
      dataset=pd.get_dummies(dataset,columns=(['LEGAL_FORM_group','PKD_section','TYPE_OF_VEHICLE']))
      #PKD section NET ASSETS, LEGAL FORM group,TYPE OF VEHICLE,NUMBER OF EMPLOYEES,CUSTOMER FOR
      X = dataset.iloc[:].values
      from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
      from sklearn.preprocessing import StandardScaler
      sc = StandardScaler()
      X_train = sc.fit_transform(X_train)
      X_test = sc.transform(X_test)
      k=[ 'poly', 'rbf', 'sigmoid']
      from sklearn.svm import SVC
      print("SVM")
      for a in k:
          classifier = SVC(kernel = a)
          classifier.fit(X_train, y_train)
          y_pred = classifier.predict(X_test)
          cm1 = confusion_matrix(y_test, y_pred)
          print("Test\n",cm1)
          print("Accuracy= ",classifier.score(X_test, y_test))
          scores = cross_val_score(classifier, X, y, cv=3)
          print('K-fold cross validation score '+ a +' : ',scores)
      cr=['gini', 'entropy']
      from sklearn.ensemble import RandomForestClassifier
      print("RandomForestClassifier")
      for a in cr:
```

```
43.          classifier1 = RandomForestClassifier(n_estimators =
    10, criterion = a, random_state = 0)
44.          classifier1.fit(X_train, y_train)
45.          y_pred = classifier1.predict(X_test)
46.          cm = confusion_matrix(y_test, y_pred)
47.          print("Test\n",cm)
48.          print("Accuracy= ",classifier1.score(X_test, y_test
   ))
49.          scores = cross_val_score(classifier1, X, y, cv=3)
50.          print('K-
   fold cross validation score '+ a +' : ',scores)
51.
```

52. Kết quả

```
SVM
Test
 [[1684    0]
 [  40    0]]
Accuracy=  0.9767981438515081
K-fold cross validation score poly :  [0.97667943 0.97667943 0.97667943]
Test
 [[1684    0]
 [  40    0]]
Accuracy=  0.9767981438515081
K-fold cross validation score rbf :  [0.97667943 0.97667943 0.97667943]
Test
 [[1681    3]
 [  40    0]]
Accuracy=  0.9750580046403712
K-fold cross validation score sigmoid :  [0.97215454 0.97424295 0.97250261]
RandomForestClassifier
Test
 [[1675    9]
 [  35    5]]
Accuracy=  0.974477958236659
K-fold cross validation score gini :  [0.97215454 0.96832579 0.97215454]
Test
 [[1676    8]
 [  35    5]]
Accuracy=  0.9750580046403712
K-fold cross validation score entropy :  [0.97145841 0.96797772 0.97250261]
```

Nhận xét:

- Với SVM thì tham số poly và rbf cho kết quả như nhau và tốt nhất.
- Với RandomForest thì tham số gini cho kết quả tốt nhất so với entropy là 2/3.