القسم العاشر: مكتبة SKlearn

A. Data Preparation

1. Data files from SKlearn
2. Data cleaning
3. Metrics module
4. Feature selection
5. Data Scaling
6. Data Split

B. ML Algorithms

1. Linear Regression
2. Logistic Regression
3. Neural Network
4. SVR
5. SVC
6. K-means
7. PCA
8. Decision Tree
9. Ensemble Regression
10. Ensemble Classifier
11. K Nearest Neighbors
12. Naïve Bayes
13. LDA , QDA
14. Hierarchical Clusters
15. DbScan
16. NLP
17. Apriori

C. Algorithm Evaluation :

1. Model Check
2. Grid Search
3. Pipeline
4. Model Save

D. Time Series

1.4) Feature Selection

و هي خاصة باختيار الـ features المطلوبة و المؤثرة و استبعاد الباقين ويتم اختيارها بناء علي مدي ارتباطها بالمخرج y ,
وتتم عبر الموديول feature_selection

1.4.1 feature_selection.SelectPercentile
1.4.2 feature_selection.GenericUnivariateSelect
1.4.3 feature_selection.SelectKBest
1.4.4 feature_selection.SelectFromModel

1.4.1) Select Percentile

يتم استخدام اداة selectpercentile واللي بتختار اهم فيتشرز مرتبطة بالناتج حسب النسبة المئوية المعطاة , ويتم تحديد مقدار الاهمية بطرق عديدة , مثل أداة f_classif او أداة chi2

الصيغة :

```
# Import Libraries
from sklearn.feature_selection import SelectPercentile
from sklearn.feature_selection import chi2 , f_classif
#-------------------------------------------------

#Feature Selection by Percentile
print('Original X Shape is ' , X.shape)
FeatureSelection = SelectPercentile(score_func = chi2, percentile=20) # score_func can = f_classif
X = FeatureSelection.fit_transform(X, y)

#showing X Dimension
```

```
print('X Shape is ' , X.shape)
print('Selected Features are : ' , FeatureSelection.get_support())
```

مثال

```
from sklearn.datasets import load_digits
from sklearn.feature_selection import SelectPercentile, chi2
X, y = load_digits(return_X_y=True)
```

هنا يتم اظهار عدد فيتشرز الداتا و هي 64

```
X.shape
```

نقوم بجعله يختار اهم 10 % منهم اي 7 فيشترز

```
X_new = SelectPercentile(score_func =chi2, percentile=10).fit_transform(X, y)

print(X_new.shape)
```

مثال آخر

```
from sklearn.datasets import load_breast_cancer
from sklearn.feature_selection import SelectPercentile   , chi2
```

```
data = load_breast_cancer()
X = data.data
y = data.target
X.shape
sel = SelectPercentile(score_func = chi2 , percentile = 20).fit_transform(X,y)
sel.shape
```

و اذا اردنا معرفة الفيتشرز المختارة و المستبعدة

```
from sklearn.datasets import load_digits
from sklearn.feature_selection import SelectPercentile, chi2
X, y = load_digits(return_X_y=True)
X.shape

X_new = SelectPercentile(score_func =chi2, percentile=10)
X_new.fit(X, y)
selected = X_new.transform(X)
X_new.get_support()
```

1.4.2) Generic Univariate Select

و فيها يتم اختيار عدد معين من الفيتشرز بناء علي احد الادوات

الصيغة :

```
#Import Libraries
from sklearn.feature_selection import GenericUnivariateSelect
from sklearn.feature_selection import chi2 , f_classif
#-------------------------------------------------
#Feature Selection by Generic
#print('Original X Shape is ' , X.shape)
FeatureSelection = GenericUnivariateSelect(score_func= chi2, mode= 'k_best', param=3) # score_func can =
f_classif : mode can = percentile,fpr,fdr,fwe
X = FeatureSelection.fit_transform(X, y)
#showing X Dimension
#print('X Shape is ' , X.shape)
#print('Selected Features are : ' , FeatureSelection.get_support())
```

وهنا يتم تحديد عدد 5 فيتشرز

```
from sklearn.datasets import load_breast_cancer
from sklearn.feature_selection import GenericUnivariateSelect, chi2
X, y = load_breast_cancer(return_X_y=True)
X.shape

transformer = GenericUnivariateSelect(chi2, 'k_best', param=5)
X_new = transformer.fit_transform(X, y)

X_new.shape

transformer.get_support()
```

1.4.3) Select KBest

يقوم كذلك باختيار عدد معين من الـ features باسلوب رياضي مختلف

الصيغة :

```
#Import Libraries
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2 , f_classif
#-----------------------------------------------
#Feature Selection by KBest
#print('Original X Shape is ' , X.shape)
FeatureSelection = SelectKBest(score_func= chi2 ,k=3) # score_func can = f_classif
X = FeatureSelection.fit_transform(X, y)

#showing X Dimension
#print('X Shape is ' , X.shape)
#print('Selected Features are : ' , FeatureSelection.get_support())
```

مثال

```
from sklearn.datasets import load_digits
from sklearn.feature_selection import SelectKBest, chi2
X, y = load_digits(return_X_y=True)
X.shape

X_new = SelectKBest(chi2, k=30).fit_transform(X, y)

X_new.shape
```

1.4.4) Select From Model

يتم اختيار الفيتشر بناء علي موديل معين بحيث الموديل نفسه يشوف انهي فيشترز مهمة , وده بامر selectfrommodel

الصيغة :

```
#Import Libraries
from sklearn.feature_selection import SelectFromModel
#-----------------------------------------------

#Feature Selection by KBest
#print('Original X Shape is ' , X.shape)

'''
from sklearn.linear_model import LinearRegression
thismodel = LinearRegression()
'''
```

```
FeatureSelection = SelectFromModel(estimator = thismodel, max_features = None) # make sure that thismodel is well-defined
X = FeatureSelection.fit_transform(X, y)

#showing X Dimension
#print('X Shape is ' , X.shape)
#print('Selected Features are : ' , FeatureSelection.get_support())
```

مثال

```
from sklearn.datasets import load_breast_cancer
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectFromModel
data = load_breast_cancer()
X = data.data
y = data.target
```

تم تحديد موديل الغابة العشوائية , برقم 20 لها , وعلي اساسها هيتم اختيار اقوي فيتشرز

```
sel = SelectFromModel(RandomForestClassifier(n_estimators = 20))
```

```
sel.fit(X,y)
selected_features = sel.transform(X)
sel.get_support()
```

لاحظ ان مش لازم يتم اختيار نفس الموديل في الترين , ممكن موديل تاني , فاختيار الفيتشرز عادي من موديل مختلف , كمان متنساش ان ممكن يتم عمل خطوات ورا بعص , يعني مثلا بولينوميال عشان اعمل فيتشرز كتير جدا , بعدها اجيب موديل يختار منهم