



Pelatihan Microcredential CERTIFICATION untuk Associate Data Scientist

1 November - 10 Desember 2021

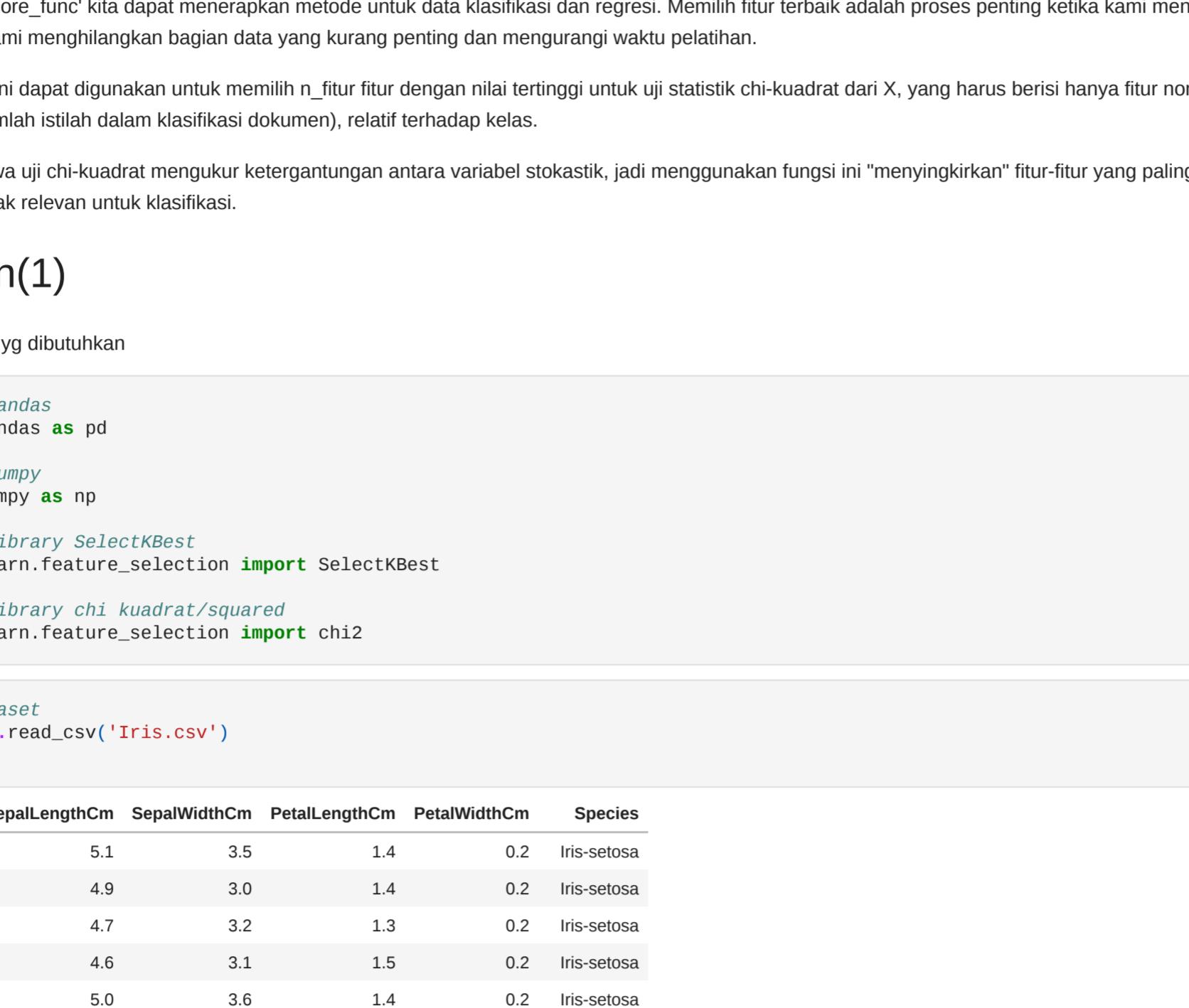


Hands-On

Hands-On ini digunakan pada kegiatan Microcredential Associate Data Scientist 2021

Pertemuan 7

Pertemuan 7 (tujuh) pada Microcredential Associate Data Scientist 2021 menyampaikan materi mengenai Menentukan Objek atau Memilih Data atau Seleksi Fitur



Info dataset: Dataset ini berisi ukuran/measures 3 spesies iris

Seleksi Univariat

Uji statistik dapat digunakan untuk memilih fitur-fitur tsb yang memiliki relasi paling kuat dengan variabel output

Scikit-learn API menyediakan kelas `SelectKBest` untuk mengekstrak fitur terbaik dari dataset yang diberikan. Metode `SelectKBest` memilih fitur sesuai dengan skor tertinggi. Dengan mengubah parameter `'score_func'` kita dapat menerapkan metode untuk data klasifikasi dan regresi. Memilih fitur terbaik adalah proses penting ketika kami siapkan kumpulan data besar untuk pelatihan. Ini membantu kami menghilangkan bagian data yang kurang penting dan mengurangi waktu pelatihan.

chi-kuadrat ini dapat digunakan untuk memilih n_fitur pertama dengan nilai tertinggi untuk uji statistik chi-kuadrat dari X, yang harus berisi hanya fitur non-negatif seperti boolean atau frekuensi (misalkan, jumlah istilah dalam klasifikasi dokumen), relatif terhadap kelas.

Ingin tahu bahwa uji chi-kuadrat mengukur ketergantungan antara variabel stokastik, jadi menggunakan fungsi ini "menyengkirkan" fitur-fitur yang paling mungkin tidak bergantung pada kelas dan oleh karena itu tidak relevan untuk klasifikasi.

Latihan(1)

import library yg dibutuhkan

```
In [2]: #import pandas
import pandas as pd

#import numpy
import numpy as np

#Import Library SelectKBest
from sklearn.feature_selection import SelectKBest

#import Library chi kuadrat squared
from sklearn.feature_selection import chi2
```

```
In [3]: #load dataset
data = pd.read_csv('Iris.csv')
data
```

```
Out[3]:
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 5 columns

Latihan(2)

buat dataframe tanpa kolom 'Id' yang ditampung dalam variabel bernama df1, lalu tampilkan

```
In [6]: #menghilangkan kolom Id
df1 = data.drop(['Id'], axis=1)
#lalu tampilkan
df1
```

```
Out[6]:
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 5 columns

Latihan(3)

- Buat variabel independent columns dan target kedalam variabel X dan y

```
In [10]: # independent columns --> SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
X = df1.drop(['Species'], axis=1)
# target columns --> species
y = df1['Species']
```

Latihan(4)

- Aplikasikan library `SelectKBest` untuk mengekstrak fitur terbaik dari dataset

```
In [11]: # menerapkan SelectKBest untuk melakukan ekstraksi
bestfeature = SelectKBest(score_func=chi2, k=4)
fit = bestfeature.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)

#mergerngkan 2 dataframe tersebut untuk visualisasi yang lebih bagus
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Field', 'Score']
print(featureScores.nlargest(10,'Score'))
```

Latihan(5)

- lihat hasil seleksi feature

```
In [12]: # menggabungkan 2 dataframe tersebut untuk visualisasi yang lebih bagus
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Field', 'Score']
print(featureScores.nlargest(10,'Score'))
```

Feature Importance (FT)

FT berfungsi memberi skor untuk setiap fitur data, semakin tinggi skor semakin penting atau relevan fitur tersebut terhadap variabel output

FT merupakan kelas inbuilt yang dilengkapi dengan Pengklasifikasi Berbasis Pohon (Tree Based Classifier), kita akan menggunakan Pengklasifikasi Pohon Ekstra untuk mengekstrak 10 fitur teratas untuk kumpulan data

Latihan(6)

buat dataframe tanpa kolom 'Id' yang ditampung dalam variabel bernama df2, lalu tampilkan

```
In [13]:
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 5 columns

Latihan(7)

- Buat variabel independent columns dan target kedalam variabel A dan b

```
In [14]: # independent columns --> SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
A = df2.drop(['Species'], axis=1)
# target columns --> species
b = df2['Species']
```

Latihan(8)

Tujuan dari `ExtraTreesClassifier` adalah untuk menyesuaikan jumlah pohon keputusan acak ke data, dan dalam hal ini adalah dari pembelajaran ensemble. Khususnya, pemisahan acak dari semua pengamatan dilakukan untuk memastikan bahwa model tidak terlalu cocok dengan data.

- Aplikasikan library `ExtraTreesClassifier` untuk mengekstrak fitur terbaik dari dataset

```
In [15]: # import library ExtraTreesClassifier
from sklearn.ensemble import ExtraTreesClassifier
# import library matplotlib
import matplotlib.pyplot as plt

# fit model ExtraTreesClassifier
model = ExtraTreesClassifier()
model.fit(A,b)

# plot heatmap
h = plt.figure(figsize=(10,10))
```

```
Out[15]: ExtraTreesClassifier()
```

Latihan(9)

- visualisasikan hasil dari model `ExtraTreesClassifier`

```
In [16]: # melakukan plot dari feature importances
print(model.feature_importances_)
feat_importance = pd.Series(model.feature_importances_, index=A.columns)
feat_importance.nlargest(10).plot(kind='bar')
plt.show()
```

[0.08734316 0.063241 0.40115786 0.44825797]

Feature Importance (FT)

FT berfungsi memberi skor untuk setiap fitur data, semakin tinggi skor semakin penting atau relevan fitur tersebut terhadap variabel output

FT merupakan kelas inbuilt yang dilengkapi dengan Pengklasifikasi Berbasis Pohon (Tree Based Classifier), kita akan menggunakan Pengklasifikasi Pohon Ekstra untuk mengekstrak 10 fitur teratas untuk kumpulan data

Latihan(6)

buat dataframe tanpa kolom 'Id' yang ditampung dalam variabel bernama df2, lalu tampilkan

```
In [13]:
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 5 columns

Latihan(7)

- Buat variabel independent columns dan target kedalam variabel A dan b

```
In [14]: # independent columns --> SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
A = df2.drop(['Species'], axis=1)
# target columns --> species
b = df2['Species']
```

Latihan(8)

visualisasikan hasil dari model `ExtraTreesClassifier`

```
In [16]: # melakukan plot dari feature importances
print(model.feature_importances_)
feat_importance = pd.Series(model.feature_importances_, index=A.columns)
feat_importance.nlargest(10).plot(kind='bar')
plt.show()
```

[0.08734316 0.063241 0.40115786 0.44825797]

Feature Importance (FT)

FT berfungsi memberi skor untuk setiap fitur data, semakin tinggi skor semakin penting atau relevan fitur tersebut terhadap variabel output

FT merupakan kelas inbuilt yang dilengkapi dengan Pengklasifikasi Berbasis Pohon (Tree Based Classifier), kita akan menggunakan Pengklasifikasi Pohon Ekstra untuk mengekstrak 10 fitur teratas untuk kumpulan data

Latihan(6)

buat dataframe tanpa kolom 'Id' yang ditampung dalam variabel bernama df2, lalu tampilkan

```
In [13]:
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.				