

LAPORAN
PRAKTIKUM BIG DATA ANALYTIC
Pertemuan Ke - 8



Dosen :
Sri Redjeki, S.Si., M.Kom.

Disusun oleh :
RAHADIYAN BONDAN PERMADI
215411119

Universitas Teknologi Digital Indonesia
UTDI
YOGYAKARTA
2022

MODELLING COLLECTED DATA

Dasar Teori

Data besar tersebut membutuhkan beban komputasi yang tinggi. Semakin banyak jumlah data, jumlah atribut (fitur) maka semakin besar pula beban komputer. Solusinya adalah melalui reduksi data sehingga jumlah data semakin kecil.

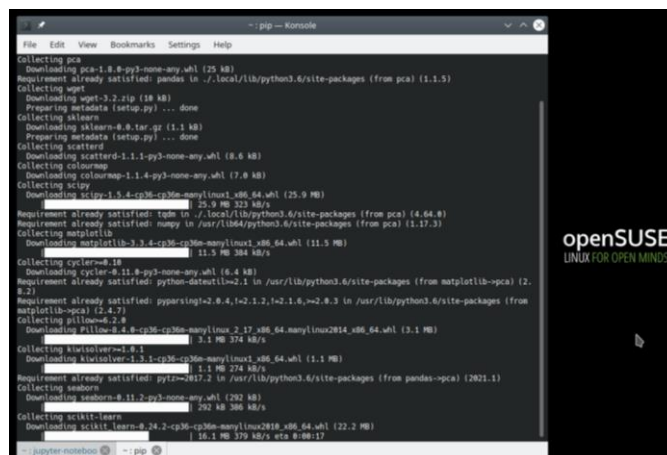
Principal Component Analysis (PCA) adalah sebuah metode bagaimana mereduksi dimensi data dengan menggunakan beberapa garis/bidang yang disebut dengan Principle Components (PCs). PCA dapat digunakan untuk visualisasi data sehingga diharapkan dapat membantu kita untuk menginterpretasikan data dan melihat pembagian data ke dalam beberapa cluster (meskipun bukan tujuan utama). Sekilas, PCA mirip dengan Teknik clustering misalnya seperti K-Means. Karena PCA berada di domain Machine Learning (bukan termasuk domain deep learning), maka PCA juga bisa digunakan untuk meningkatkan kecepatan algoritma machine learning.

Kebutuhan Alat

1. Python
2. Jupyter Notebook

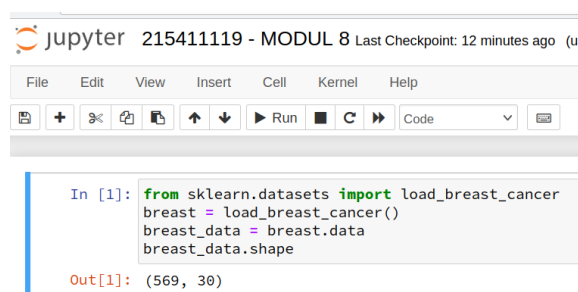
PRAKTIK

1. **INSTALL PCA “pip install pca”**



```
-- pip -- Konsole
File Edit View Bookmarks Settings Help
Collecting pca
  Downloading pca-1.8.0-py3-none-any.whl (25 kB)
Requirement already satisfied: pandas in ./local/lib/python3.6/site-packages (from pca) (1.1.5)
Collecting wget
  Downloading wget-3.2.zip (18 kB)
  Preparing metadata (setup.py) ... done
Collecting sklearn
  Downloading sklearn-0.9.tar.gz (1.1 kB)
  Preparing metadata (setup.py) ... done
Collecting scatter
  Downloading scatter-1.1.1-py3-none-any.whl (8.6 kB)
Collecting colourmap
  Downloading colourmap-1.1.4-py3-none-any.whl (7.0 kB)
Collecting numpy
  Downloading numpy-1.5.4-cp36-cp36m-manylinux1_x86_64.whl (25.9 MB)
Requirement already satisfied: tqdm in ./local/lib/python3.6/site-packages (from pca) (4.64.0)
Requirement already satisfied: numpy in /usr/lib64/python3.6/site-packages (from pca) (1.17.3)
Collecting matplotlib
  Downloading matplotlib-3.3.4-cp36-cp36m-manylinux1_x86_64.whl (11.5 MB)
Collecting cytoolz
  Downloading cytoolz-0.11.0-py3-none-any.whl (16.4 kB)
Requirement already satisfied: python-dateutil<2.1 in /usr/lib/python3.6/site-packages (from matplotlib-pca) (2.8.2)
Requirement already satisfied: pyparsing<2.8.4, >2.1.2, <2.1.6, >2.0.3 in /usr/lib/python3.6/site-packages (from matplotlib-pca) (2.4.7)
Collecting pillow<6.2.0
  Downloading pillow-6.0.0-cp36-cp36m-manylinux_2_17_x86_64_muslmanylinux2014_x86_64.whl (3.1 MB)
Collecting kiwisolver<1.0.1
  Downloading kiwisolver-1.3.1-cp36-cp36m-manylinux1_x86_64.whl (1.1 MB)
Requirement already satisfied: pyparsing<2.8.4, >2.1.2, <2.1.6, >2.0.3 in /usr/lib/python3.6/site-packages (from pandas-pca) (2.021.1)
Collecting seaborn
  Downloading seaborn-0.11.2-py3-none-any.whl (262 kB)
Collecting scikit-learn
  Downloading scikit-learn-0.24.2-cp36-cp36m-manylinux2018_x86_64.whl (22.2 MB)
Out[1]: (569, 30)
```

2. **LOAD DATA IMAGE DATA CANCER**



```
Jupyter 215411119 - MODUL 8 Last Checkpoint: 12 minutes ago (ui)
File Edit View Insert Cell Kernel Help
+ - - - - - Run - - - - - Code
In [1]: from sklearn.datasets import load_breast_cancer
breast = load_breast_cancer()
breast_data = breast.data
breast_data.shape
Out[1]: (569, 30)
```

Terdapat data sebanyak 569 dengan 30 atribut

3. Menambah satu atribut sebagai target baru dengan koding dibawah ini :

Sebelumnya kita harus load data label dengan syntax

```
breast_labels = breast.target  
breast_labels.shape  
sebanyak 569 label
```

```
jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)  
File Edit View Insert Cell Kernel Help  
+ - - - - -  
In [3]: from sklearn.datasets import load_breast_cancer  
breast = load_breast_cancer()  
breast_data = breast.data  
breast_data.shape  
Out[3]: (569, 30)  
In [10]: breast_labels = breast.target  
breast_labels.shape  
Out[10]: (569,)  
In [11]: import numpy as np  
labels = np.reshape(breast_labels,(569,1))  
final_breast_data = np.concatenate([breast_data,labels],axis=1)  
final_breast_data.shape  
Out[11]: (569, 31)
```

4. Menampilkan fitur dataset :

```
jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)  
File Edit View Insert Cell Kernel Help  
+ - - - - -  
Out[11]: (569, 31)  
In [12]: import pandas as pd  
breast_dataset = pd.DataFrame(final_breast_data)  
features = breast.feature_names  
features  
Out[12]: array(['mean radius', 'mean texture', 'mean perimeter', 'mean area',  
'mean smoothness', 'mean compactness', 'mean concavity',  
'mean concave points', 'mean symmetry', 'mean fractal dimension',  
'radius error', 'texture error', 'perimeter error', 'area error',  
'smoothness error', 'compactness error', 'concavity error',  
'concave points error', 'symmetry error',  
'fractal dimension error', 'worst radius', 'worst texture',  
'worst perimeter', 'worst area', 'worst smoothness',  
'worst compactness', 'worst concavity', 'worst concave points',  
'worst symmetry', 'worst fractal dimension'], dtype='<U23')
```

5. Menampilkan isi dataset dgn 31 feature Feature tambahan diberi nama label

```
jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)  
File Edit View Insert Cell Kernel Help  
+ - - - - -  
In [13]: features_labels = np.append(features,'label')  
breast_dataset.columns = features_labels  
breast_dataset.head()  
Out[13]:  
   mean radius  mean texture  mean perimeter  mean area  mean smoothness  mean compactness  mean concavity  mean concave points  mean symmetry  mean fractal dimension  ...  worst texture  worst perimeter  worst area  worst smoothness  worst compactness  
0    17.99      10.38      122.80     1001.0      0.11840      0.27760      0.3001      0.14710      0.2419      0.07871  ...      17.33      184.60     2019.0      0.1622  
1    20.57      17.77      132.90     1326.0      0.08474      0.07864      0.0869      0.07017      0.1812      0.05667  ...      23.41     158.80     1956.0      0.1238  
2    19.69      21.25      130.00     1203.0      0.10960      0.15990      0.1974      0.12790      0.2069      0.05999  ...      25.53     152.50     1709.0      0.1444  
3    11.42      20.38       77.58      386.1      0.14250      0.28390      0.2414      0.10520      0.2597      0.09744  ...      26.50      98.87      567.7      0.2098  
4    20.29      14.34      135.10     1297.0      0.10030      0.13280      0.1980      0.10430      0.1809      0.05883  ...      16.67     152.20     1575.0      0.1374  
5 rows x 31 columns
```

6. Mengganti isi label (replacement)

Dataset yang ada menunjukkan bahwa isi dari kolom label adalah "0" dan "1" (lihat dengan menggeser output yang ada). Apabila angka 0 dan 1 akan di ganti dengan class lain maka kodingnya

jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted Python 3

5 rows x 31 columns

```
In [14]: breast_dataset['label'].replace(0, 'Benign', inplace=True)
breast_dataset['label'].replace(1, 'Malignant', inplace=True)
breast_dataset.tail()
```

```
Out[14]:
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	com
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0	0.14100	
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660	
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390	
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500	
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996	

5 rows x 31 columns

7. Normalisasi dataset

jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Help Trusted Python 3

```
Out[14]:
```

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst texture	worst perimeter	worst area	worst smoothness	com
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	26.40	166.10	2027.0	0.14100	
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	38.25	155.00	1731.0	0.11660	
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	34.12	126.70	1124.0	0.11390	
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	39.42	184.60	1821.0	0.16500	
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	30.37	59.16	268.6	0.08996	

5 rows x 31 columns

```
In [15]: from sklearn.preprocessing import StandardScaler
x = breast_dataset.loc[:, features].values
x = StandardScaler().fit_transform(x) # normalizing the features
x.shape
```

```
Out[15]: (569, 30)
```

Mengecek nilai rata2 dan deviasi standar

```
In [16]: np.mean(x), np.std(x)
```

```
Out[16]: (-6.826538293184326e-17, 1.0)
```

Melakukan normalisasi dataset

```
In [17]: feat_cols = ['feature'+str(i) for i in range(x.shape[1])]
normalised_breast = pd.DataFrame(x, columns=feat_cols)
normalised_breast.tail()
```

```
Out[17]:
```

	feature0	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	...	feature20	feature21	feature22	feature23	fr
564	2.110995	0.721473	2.060786	2.343856	1.041842	0.219060	1.947285	2.320965	-0.312589	-0.931027	...	1.901185	0.117700	1.752563	2.015301	0
565	1.704854	2.085134	1.615931	1.723842	0.102458	-0.017833	0.893043	1.263669	-0.217664	-1.058611	...	1.536720	2.047399	1.421940	1.494959	0
566	0.702284	2.045574	0.672676	0.577953	-0.840484	-0.038680	0.046588	0.105777	-0.809117	-0.895587	...	0.561361	1.374854	0.579001	0.427906	0
567	1.838341	2.336457	1.982524	1.735218	1.525767	3.272144	3.296944	2.658866	2.137194	1.043695	...	1.961239	2.237926	2.303601	1.653171	1
568	-1.808401	1.221792	-1.814389	-1.347789	-3.112085	-1.150752	-1.114873	-1.261820	-0.820070	-0.561032	...	-1.410893	0.764190	-1.432735	-1.075813	0

5 rows x 30 columns

8. Mereduksi fitur dataset

Dari 30 fitur/atribut/dimensi akan di jadikan 2 feature menggunakan metode PCA. Ini untuk memudahkan visualisasi dataset :

```

jupyter 215411119 - MODUL 8 Last Checkpoint: an hour ago (unsaved changes)
File Edit View Insert Cell Kernel Help
In [18]: from sklearn.decomposition import PCA
pca_breast = PCA(n_components=2)
principalComponents_breast = pca_breast.fit_transform(x)
principal_breast_Df = pd.DataFrame(data = principalComponents_breast
, columns = ['principal component 1', 'principal component 2'])
principal_breast_Df.tail(20)

Out[18]:
principal component 1  principal component 2
549 -2.551440 0.228330
550 -4.694923 -0.767478
551 -2.025037 1.261242
552 -2.895948 -1.451436
553 -3.502201 1.800832
554 -2.153904 -0.830069
555 -2.055084 1.816459
556 -3.877290 1.084255
557 -4.063862 0.122168
558 -0.088667 -0.213560
559 -1.089376 1.292848
560 -0.481771 -0.178020
561 -4.870310 -2.131106
562 5.917613 3.482637
563 8.741338 -0.873855
564 6.439315 -3.576817
565 3.793382 -3.584048
566 1.256179 -1.902797
567 10.374794 1.672010
568 -5.475243 -0.670837

```

Ploting hasil PCA dibawah ini :



Demikian laporan Pertemuan Ke-Delapan yang dapat saya rangkum dan saya kerjakan, saya Mampu memahami dan mengimplementasikan Teknik reduksi data menggunakan Principal Component Analysis serta menampilkannya dalam Ploting

=====TerimaKasih=====