Assignment7

December 22, 2021

[11]:
```
import pandas as pd
import numpy as np
%matplotlib inline
from sklearn.datasets import load_iris
from sklearn import tree
from sklearn import metrics, model_selection, preprocessing
from sklearn.tree import export_text
from IPython.display import Image,display
import matplotlib.pyplot as plt
import sklearn.cluster as cluster
import seaborn as sns
```

[5]:
```
#import data via csv file
train_data = pd.read_csv("C:/Users/jdrex/OneDrive/Documents/DSC540/Assignment7/
↪Train.csv")
test_data = pd.read_csv("C:/Users/jdrex/OneDrive/Documents/DSC540/Assignment7/
↪Test.csv")
```

[5]:
```
     feature_1  feature_2  feature_3  feature_4  feature_5  feature_6  \
0         60.0      468.0     7.8000        1.0          0          0
1        108.0      179.0     1.6574        1.0          0          0
2          1.0        1.0     2.0000        0.0          0          0
3         60.0      468.0     7.8000        1.0          0          0
4         60.0      120.0     2.0000        1.0          0          0
..         …          …          …          …          …          …
751        1.0        1.0     2.0000        0.0          0          0
752        1.0        1.0     2.0000        0.0          0          0
753       12.0      101.0     8.4166        1.0          0          0
754       31.0       88.0     2.8387        1.0          0          0
755       49.0       87.0     1.7755        0.0          0          0

     feature_7  feature_8  feature_9  feature_10  …  feature_1549  \
0            0          0          0           0  …             0
1            0          0          0           0  …             0
2            0          0          0           0  …             0
3            0          0          0           0  …             0
4            0          0          0           0  …             0
..         …          …          …          …  …           …
```

```
751          0          0          0          0  …             0
752          0          0          0          0  …             0
753          0          0          0          0  …             0
754          0          0          0          0  …             0
755          0          0          0          0  …             0

     feature_1550  feature_1551  feature_1552  feature_1553  feature_1554  \
0               0             0             0             0             0
1               0             0             0             0             0
2               0             0             0             0             0
3               0             0             0             0             0
4               0             0             0             0             0
..            …             …             …             …             …
751             0             0             0             0             0
752             0             0             0             0             0
753             0             0             0             0             0
754             0             0             0             0             0
755             0             0             0             0             0

     feature_1555  feature_1556  feature_1557  feature_1558
0               0             0             0             0
1               0             0             0             0
2               0             0             0             0
3               0             0             0             0
4               0             0             0             0
..            …             …             …             …
751             0             0             0             0
752             0             0             0             0
753             0             0             0             0
754             0             0             0             0
755             0             0             0             0

[756 rows x 1558 columns]
```

1 3 Questions about the data

1- by grouping these wafers into groups based off their features, will it identify certain groups that all have large feature x and y, that results in an annomolie in feature z for instance. In a wafer manufacturing fab- this is helpful to know to identify if a certain process mixed with another process might be causing a certain type of defect.

2- does grouping these wafers based off their features clearly define perhaps certain product groups that are made differently and therefore have different features? This is possibly important to know, because it will tell me know different my wafers are in the fab, and how many different products I may have

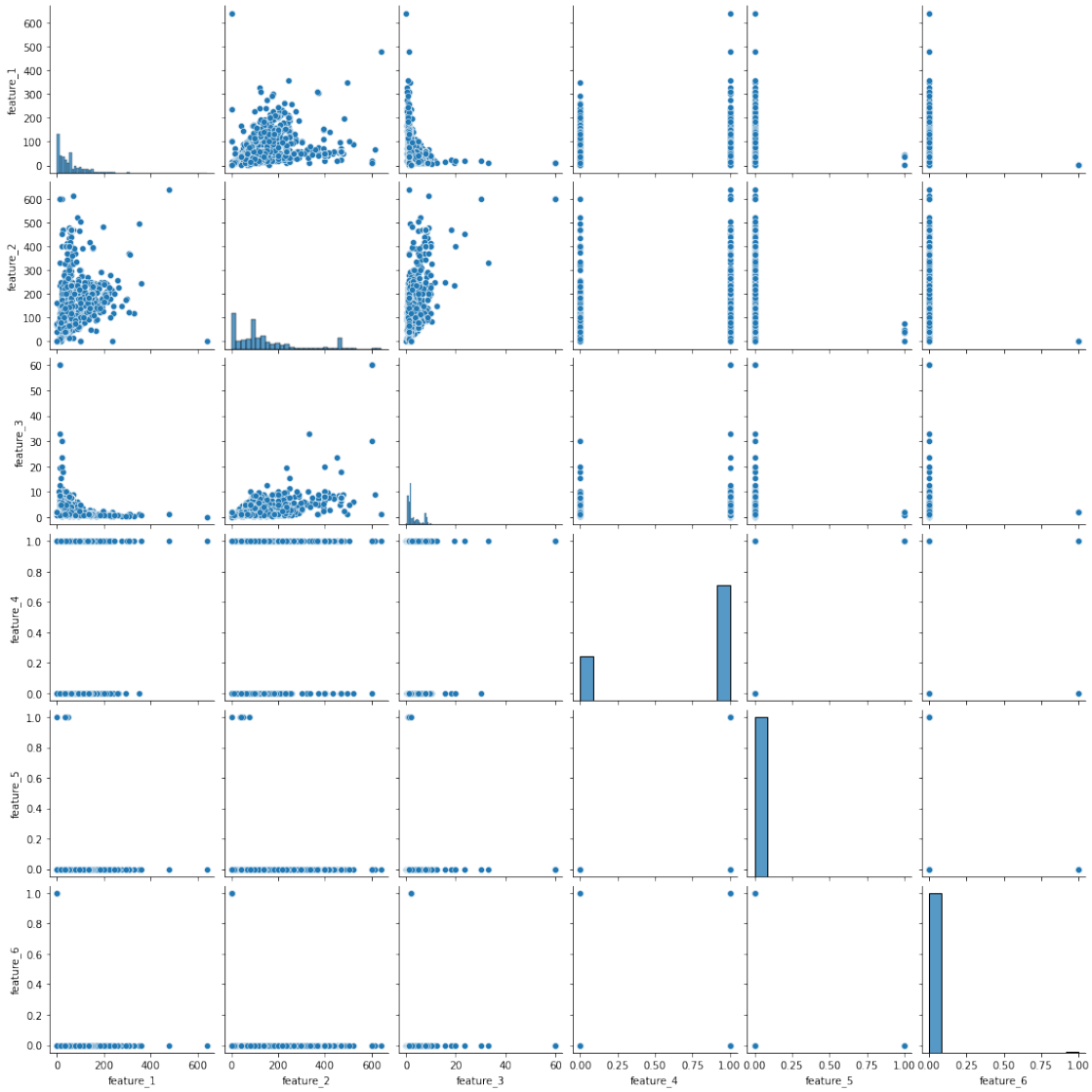
3- can grouping these wafers show me where I have a certain group of possible outliers, that had a rogue experiment or process that made perform a certain way, that is just a little different from

the rest of the group. This is important to know in a fab to uderstand perhaps if there is a certain process of engineer that is putting some far off experiments on certain lots.

2 Plot 1

[36]:
```
sns.
↪pairplot(train_data[['feature_1','feature_2','feature_3','feature_4','feature_5','feature_6
```

[36]:
```
<seaborn.axisgrid.PairGrid at 0x1ca81b0fd30>
```



[25]:
```
kmeans= cluster.KMeans(n_clusters=3, init="k-means++")
kmeans = kmeans.fit(train_data[['feature_1','feature_2']])
```

[26]:
```
kmeans.cluster_centers_
```

[26]:
```
array([[ 65.55491329, 457.12716763],
       [ 24.59753788,  51.14204545],
       [105.41011236, 168.69662921]])
```

[27]:
```
train_data['Clusters']=kmeans.labels_
```

[28]:
```
train_data.head()
```

[28]:

	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	\
0	100	160	1.6000	0	0	0	
1	20	83	4.1500	1	0	0	
2	99	150	1.5151	1	0	0	
3	40	40	1.0000	0	0	0	
4	12	234	19.5000	1	0	0	

	feature_7	feature_8	feature_9	feature_10	…	feature_1551	\
0	0	0	0	0	…	0	
1	0	0	0	1	…	0	
2	0	0	0	0	…	0	
3	0	0	0	0	…	0	
4	0	0	0	0	…	0	

	feature_1552	feature_1553	feature_1554	feature_1555	feature_1556	\
0	0	0	0	0	0	
1	0	0	0	1	0	
2	0	0	0	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	

	feature_1557	feature_1558	Class	Clusters
0	0	0	0	2
1	0	0	0	1
2	0	0	0	2
3	0	0	0	1
4	0	0	0	2

[5 rows x 1560 columns]

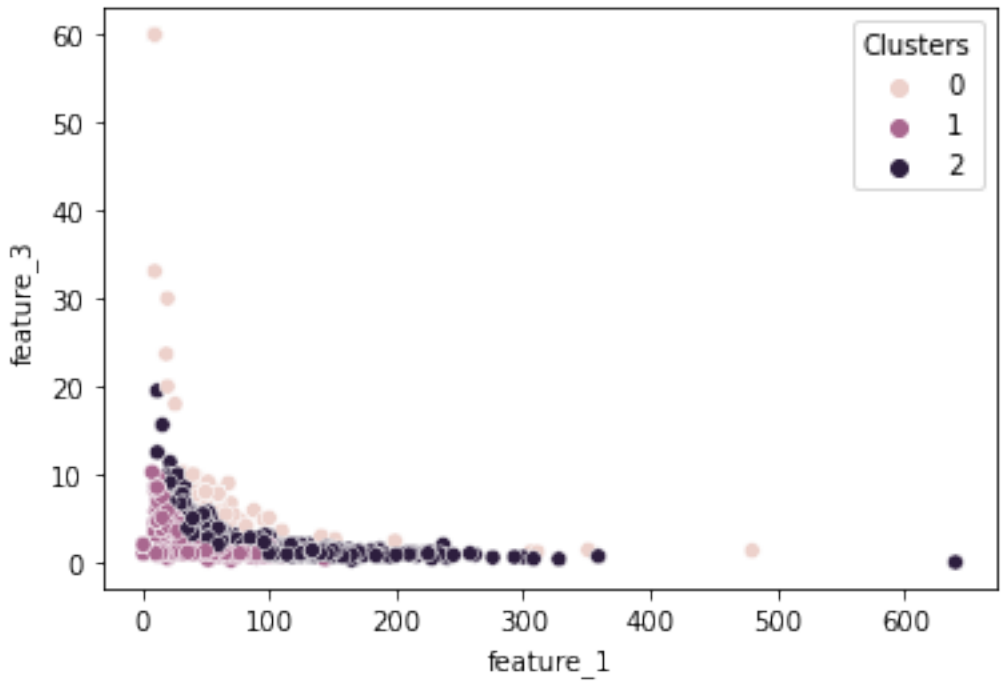
[29]:
```
train_data['Clusters'].value_counts()
```

[29]:
```
1    1054
2     536
0     173
Name: Clusters, dtype: int64
```

3 Plot 2

[45]:
```
sns.scatterplot(x="feature_1",y="feature_3",hue='Clusters',data=train_data)
```

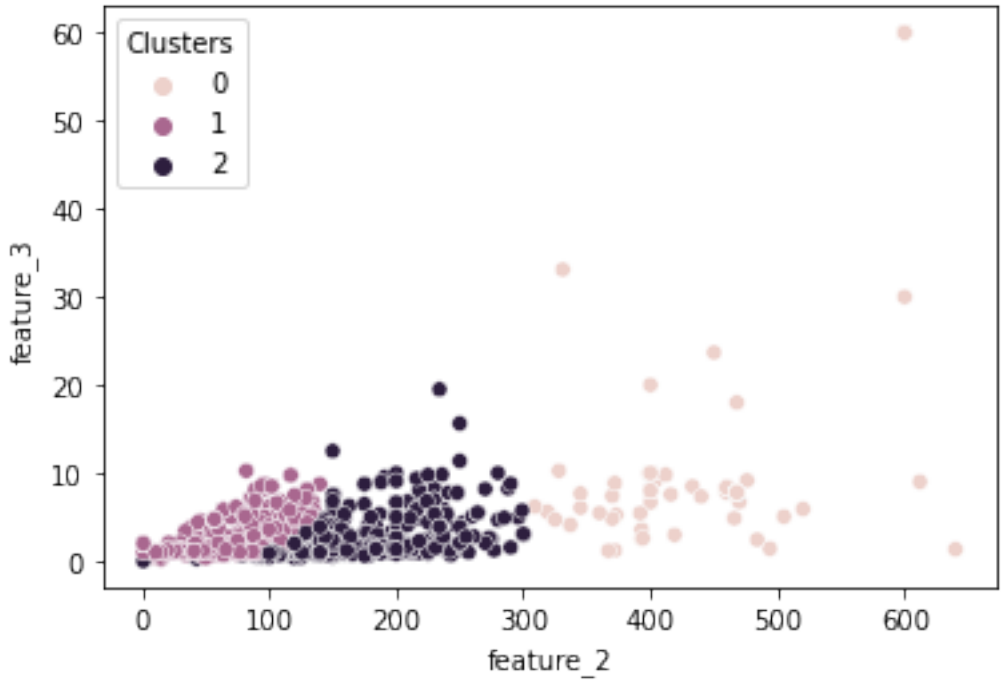
[45]:
```
<AxesSubplot:xlabel='feature_1', ylabel='feature_3'>
```



4 Plot 3

[40]:
```
sns.scatterplot(x="feature_2",y="feature_3",hue='Clusters',data=train_data)
```

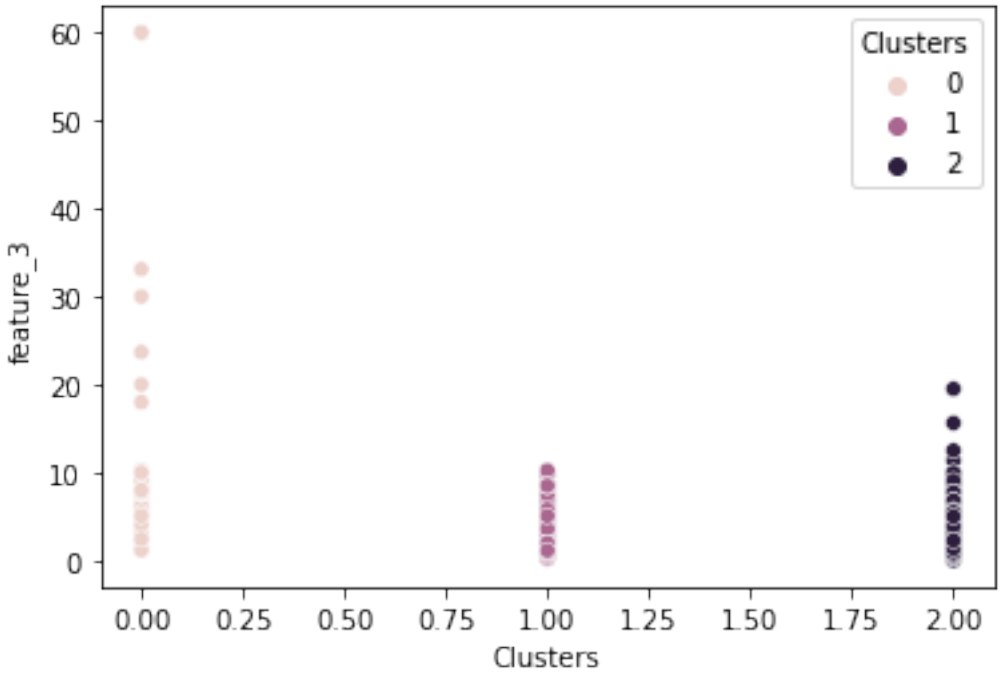
[40]:
```
<AxesSubplot:xlabel='feature_2', ylabel='feature_3'>
```



5 Plot 4

[49]:
```
sns.scatterplot(x="Clusters",y="feature_3",hue='Clusters',data=train_data)
```

[49]:
```
<AxesSubplot:xlabel='Clusters', ylabel='feature_3'>
```



6 Analysis

Looking at the results of our clusters, I started with 5, and then updated to 3 total cluster groups as it seemed to fit better. Another option is always to run a prediction with the test data, and then compare the results to the actual results, and find the mean squared error. But for simplicity, just left it at the three clusters.

Before answering the below questions, because the columns aren't labeled, I would like to make some assumptions about what each of the features could be, to make it seem more realistic, and easier to relate back to the question..

-asusming my feature 3 was an output such as a feature that indicates how well the wafer performed at an inline read- the lower the number being it perfromed well, and the higher the number being it performed poorly.
-assuming feature 1 and 2 are two different types of material, and the higher the number, the more qty of material was added to the wafer of that material.

7 Answering my questions:

1- (Question) by grouping these wafers into groups based off their features, will it identify certain groups that all have large feature x and y, that results in an annomolie in feature z for instance. In a wafer manufacturing fab- this is helpful to know to identify if a certain process mixed with another process might be causing a certain type of defect.

1- (Answer) The goal here was to understand if is a certain feature could have a strong correlation

with another that could be leading to the reason that the secondary feature is as high or low as it is.. so using feature 1,2, and 3 for example,comparing both my feature 1 and feature 2 to feature 3, there isn't a strong relation between 1 and 3, but there is a potential relation between 2 and 3, and specifically, you can see that cluster 2 is where things really start to have a larger feature 2 value, and also a larger feature 3 value.. so you could conclude that perhaps cluster group 2 was getting more of a certain type of or feature 2 material, and that results in worst performance (high faults, aka higher feature 3 value).

2- (Question) does grouping these wafers based off their features clearly define perhaps certain product groups that are made differently and therefore have different features? This is possibly important to know, because it will tell me know different my wafers are in the fab, and how many different products I may have

2- (Answer) referring to plot 3, I could see a definite difference in cluster 2, so then understanding this, went deeper in plot 4 to understand feature 3 (how well they performed at an inline checkpoint) to our cluster 2, and there is a definite increase in feature 3 value (more defects, so performed worst) , so the cluster 2 group must have had a bad process somewhere. Also this tells me that this is potentially a different "part" or grouping of wafers, compared to the others.

3- (Question) can grouping these wafers show me where I have a certain group of possible outliers, that had a rogue experiment or process that made perform a certain way, that is just a little different from the rest of the group. This is important to know in a fab to uderstand perhaps if there is a certain process of engineer that is putting some far off experiments on certain lots.

3- (Answer) referring to plot one, it is clear that both feature 1, and 2 have outliers compared to the rest of their grouping. So one option would be to further explore additional features, and see if there is a certain feature potentially causing these outliers. But in general- I think this is helpful to see, where the wafers land on a scale of each feature, and where there are ones far off. If I were an engineer manager, I might go work with those engineers to understand why their experiments were so different from the status quo.

8 Etchial Aspects

Understanding where the data comes from, how the "story" being told from the data could impact others, etc. is so important. Not only does it give the data scientist an understanding of why and what they are looking for, but it allows them to provide caution in how, where, and when thigns are explained. For instance, in this example, if I were a data scientist investigating this data, and found that there was a certain feature that appeared to be causing a bad read inline, and that all lots from a certain group were getting them, and then because of this, they decided to change that process, and turns out it was actually a different feature, this could lead to time, money, and resources wasted.

Or, if I shared that there was a certain feature causing a really good read, I would also need to be catious to not share this potentially with a customer, or competitor, that could use this same material to their advantage.

9 References:

Naik, K. Sklearn K-means Python Example- Interpreting Clustering Results. Youtube. Retrieved from: https://www.youtube.com/watch?v=3Spa10-mwsw

Ask9. Detecting Anomalies in Wafer Manufacturing. Kaggle. Retrieved from: https://www.kaggle.com/arbazkhan971/anomaly-detection