CMPE 255 - Data Mining
Wine Data Clustering and Recommendation

Team 1

SAN JOSÉ STATE
UNIVERSITY

Submitted to
Dr. Carlos Rojas
On
12/15/2020
By

Wei He
Binwang Luo
Gabrielle

Chapter 1: Introduction

1.1 Motivation

Italian wine is produced in every region of Italy, home to some of the oldest wine-producing regions in the world. People can choose hundreds of wine grape varieties from Italy. As is known to us, each kind of grape wine has different chemical composition, which decide the taste, smell and nutritional value. For this reason, we were intrigued to work on a Wine dataset and apply our knowledge of data mining.

1.2 Objectives

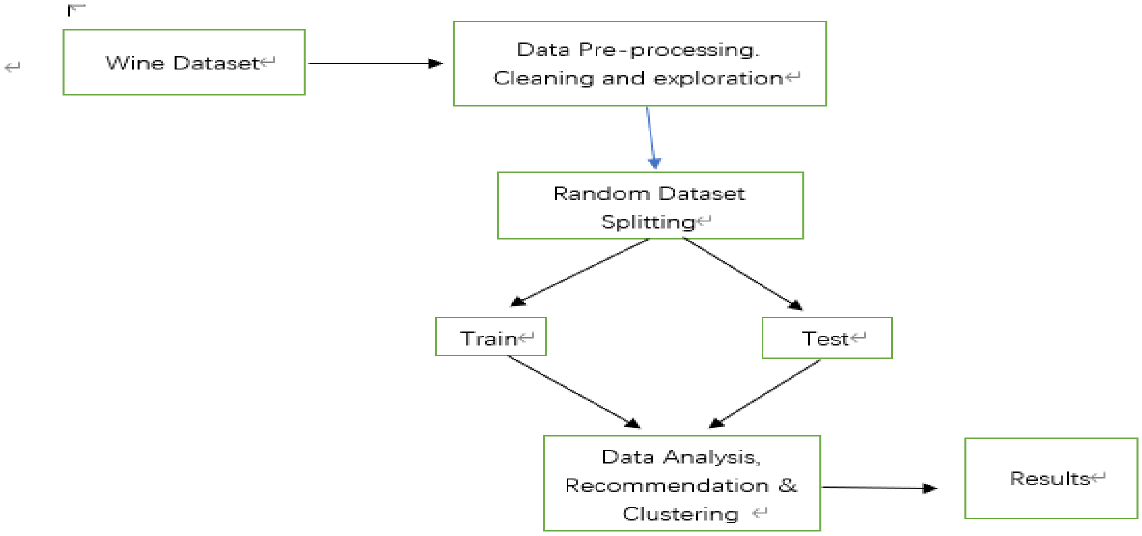
Our project objectives are focused on the classification and recommendation of the wine based on the chemical analysis and description available in the dataset.

Below is the list of our objectives that we hope to achieve as part of our project implementation:

- Cluster wines based on their chemical constituents.
- Content distribution of different constituents.
- Using chemical analysis determine the origin of wines.
- With a someone's preference of a wine, we can show "what he/she may also interested in" with similar composition recommendation.

Chapter 2: System Design and Architecture

System Design and Architecture



Chapter 3: Experiments

3.1 Dataset

Wine Dataset

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

3.2 Data Preprocessing

3.2.1 Identify all the attributes needed in the analysis.

The data was read and only the relevant columns were kept for further processing. Specially in this dataset that all of the data is needed to be analyzed. So we can use this dataset completely after confirming the count of attributes.

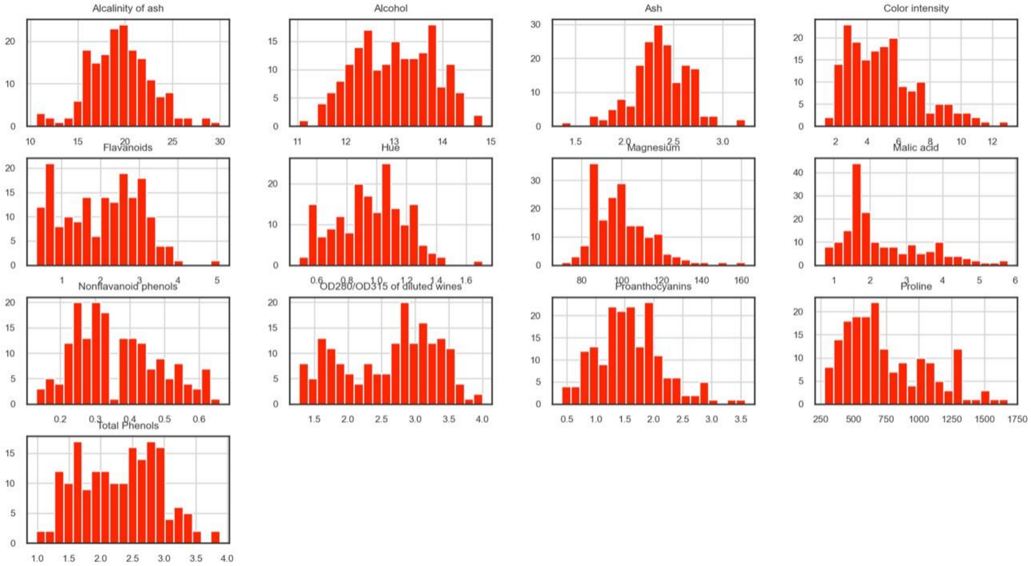
3.2.2 Cleaning data and add column names.

Specially this dataset is very clean and don’t need to do so much cleaning. But the original dataset has no attribute name for each column. In some question like clustering, I need to do add column names and check if there is missing value to fix.

3.2.3 Do the basic statistics and visualization

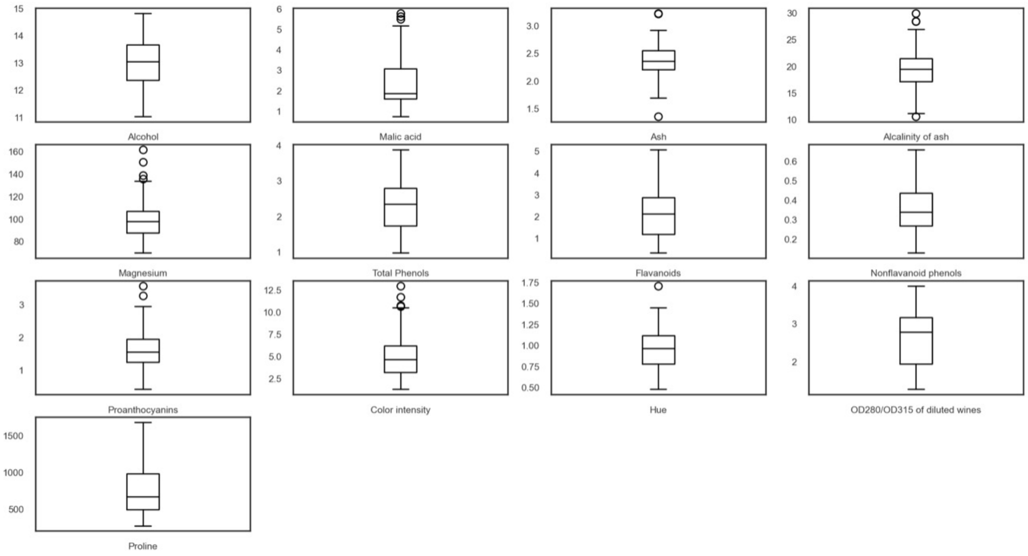
We did some basic statistic in preprocessing.

1.First we analyzed and dataset and find that there is the distribution's count of chemistry of wines. As we can see there are 13 types of chemical in wines recorded in this data set, which means we only analyze these chemicals in this project. Proline is the highest content of some types of wine. And for most of wines, the content of Nonflavanoid phenols is the lowest.

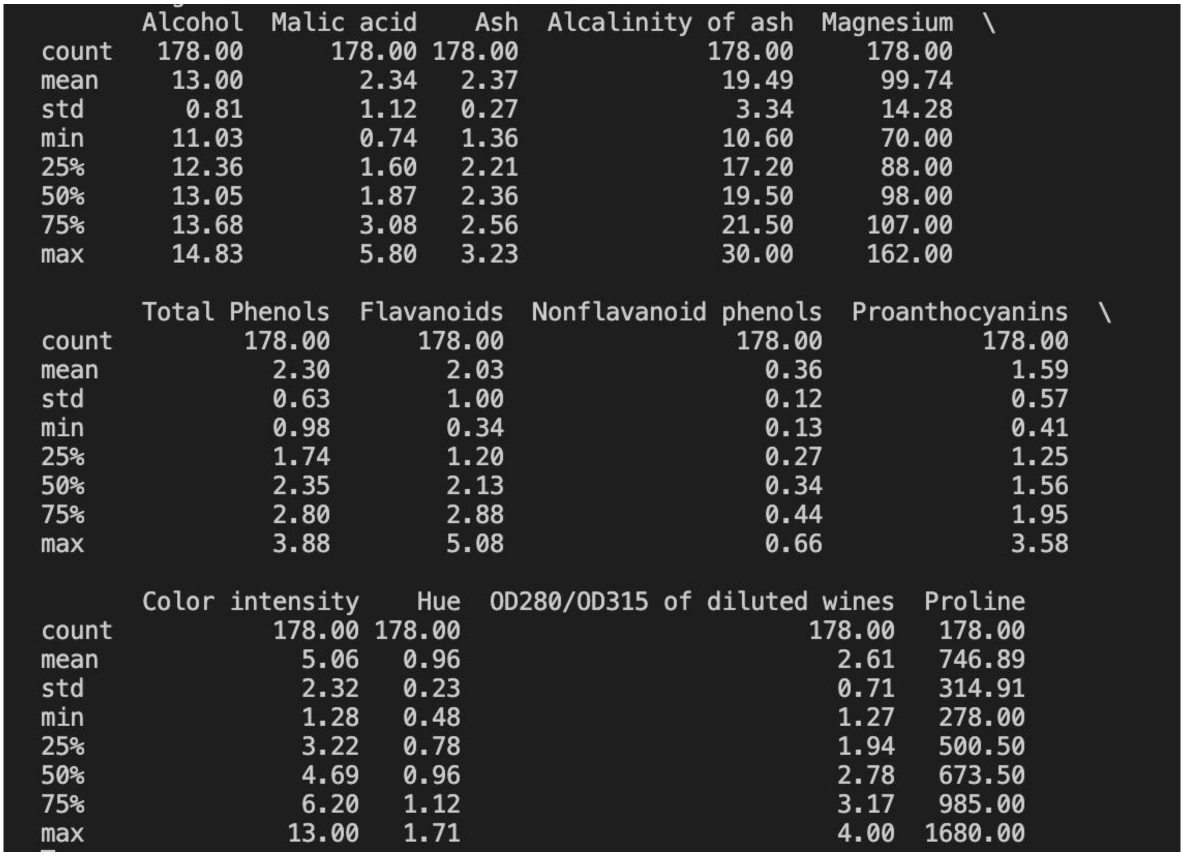


2. Second we analyzed the dataset and visualized it in another way. And then we get the

diagram below. We can see Alcohol is a most common chemistry for most wines. And some types of wines have special high or low content of some chemistry such as Magesium or makic acid(especially high in some wines) and Ash (especially low in some wines). These wines can be used to determine the origin of wines.

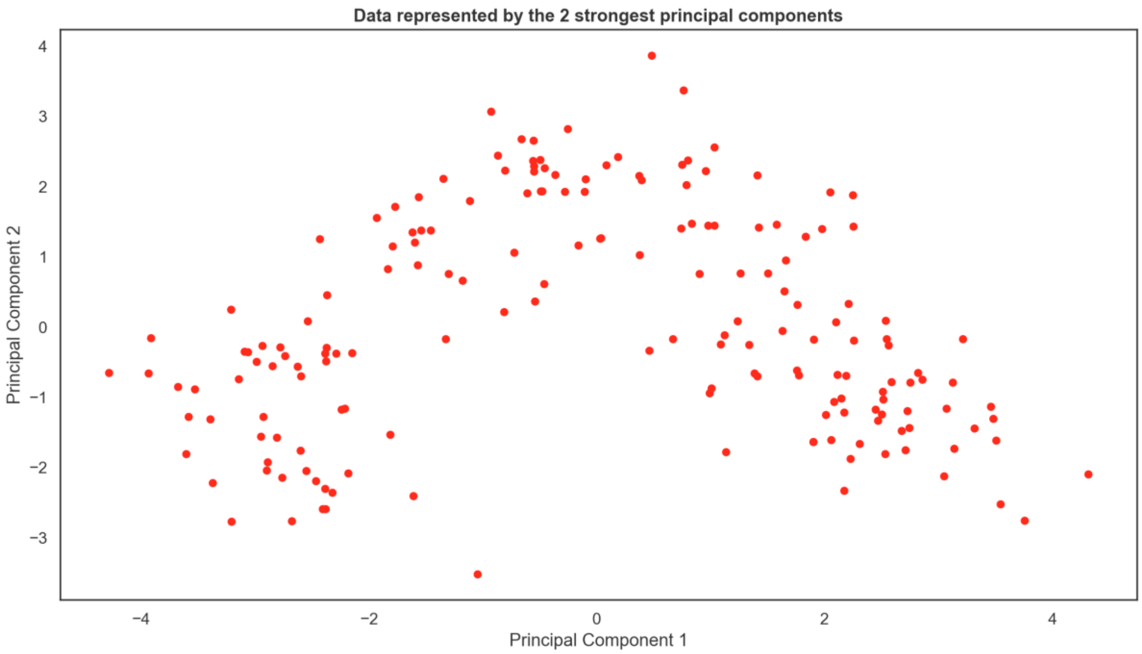


3. After that we did the basic statistics. We can see the specific quantity of all types of wines. We can see that all wines have these 13 types of chemistry.



```
        Alcohol  Malic acid    Ash  Alcalinity of ash  Magnesium  \
count    178.00      178.00 178.00             178.00     178.00
mean      13.00        2.34   2.37              19.49      99.74
std        0.81        1.12   0.27               3.34      14.28
min       11.03        0.74   1.36              10.60      70.00
25%       12.36        1.60   2.21              17.20      88.00
50%       13.05        1.87   2.36              19.50      98.00
75%       13.68        3.08   2.56              21.50     107.00
max       14.83        5.80   3.23              30.00     162.00

        Total Phenols  Flavanoids  Nonflavanoid phenols  Proanthocyanins  \
count          178.00      178.00                178.00           178.00
mean             2.30        2.03                  0.36             1.59
std              0.63        1.00                  0.12             0.57
min              0.98        0.34                  0.13             0.41
25%              1.74        1.20                  0.27             1.25
50%              2.35        2.13                  0.34             1.56
75%              2.80        2.88                  0.44             1.95
max              3.88        5.08                  0.66             3.58

        Color intensity    Hue  OD280/OD315 of diluted wines  Proline
count            178.00 178.00                        178.00   178.00
mean               5.06   0.96                          2.61   746.89
std                2.32   0.23                          0.71   314.91
min                1.28   0.48                          1.27   278.00
25%                3.22   0.78                          1.94   500.50
50%                4.69   0.96                          2.78   673.50
75%                6.20   1.12                          3.17   985.00
max               13.00   1.71                          4.00  1680.00
```

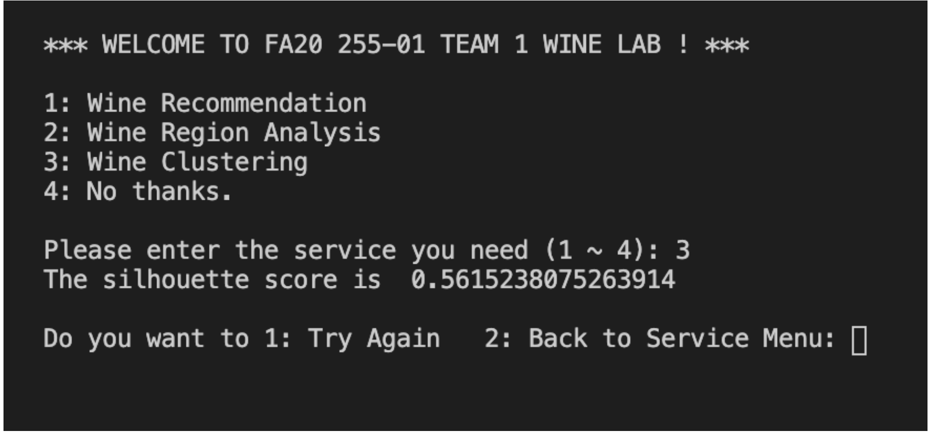
4.Finially we did the basic clustering. We can see that there should be 3 clusters in corresponding to three types of wines from three different cultivars.



3.3 Main code

3.3.1 Service menu

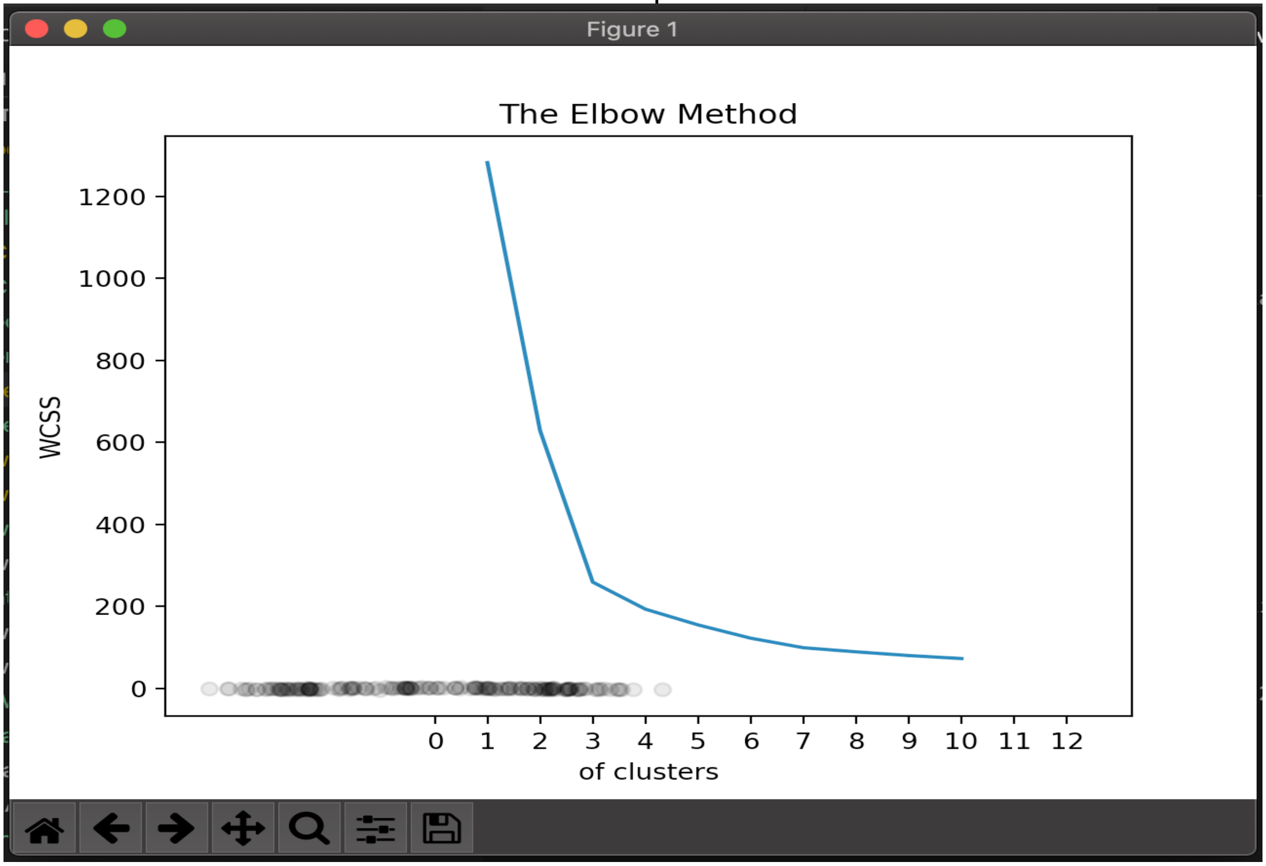
We built this project as a back-end of a wine recommendation system. In the system you can choose the service you want in the menu page like below. The code of this page is service.py. After choosing the service you want with tying in the service number, the system will link to the service page you choose.



```
*** WELCOME TO FA20 255-01 TEAM 1 WINE LAB ! ***

1: Wine Recommendation
2: Wine Region Analysis
3: Wine Clustering
4: No thanks.

Please enter the service you need (1 ~ 4): 3
The silhouette score is  0.5615238075263914

Do you want to 1: Try Again   2: Back to Service Menu: 
```

3.3.2 Wine Clustering.

As describe of the data set, the region of these wines should be divided to 3 parts. We did the clustering for these wines to the follows steps. The code of this part in in Clustering of wine dataset, which will help to recommendation and finding the region of a wine. We used K-means and clustered the wines based on their chemical constituents. We standardized the data to have a zero mean and unit variance. We used PCA and reduced the dimensionality of data. To determine ideal number of clusters i applied elbow method which has given me 3 cluster. All the metrics indicates that 3 is the best cluster number. The silhouette score obtained is 0.56. The output as below:



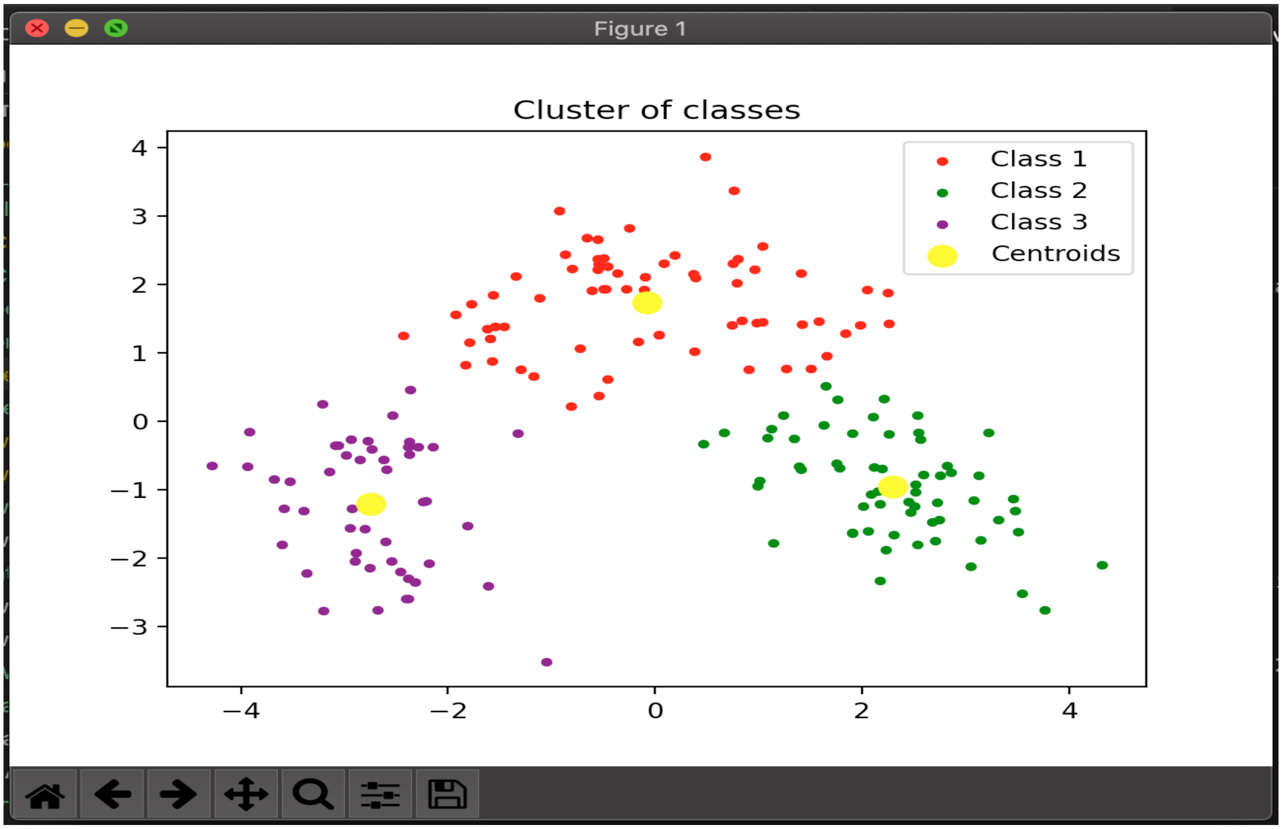
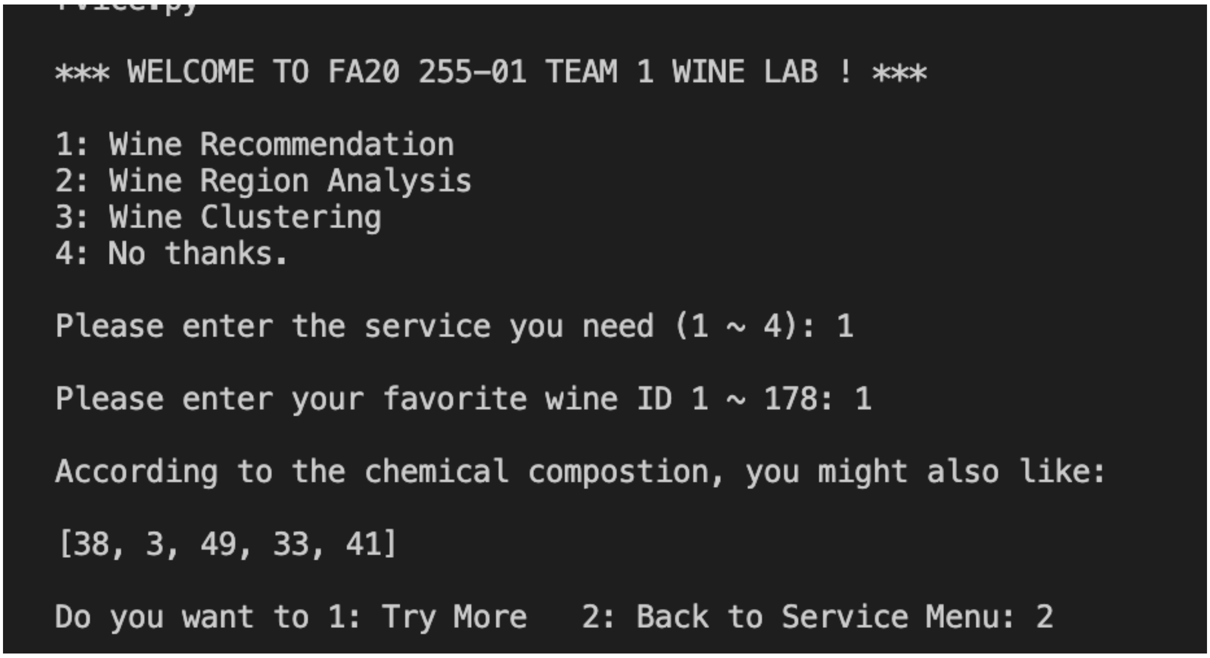


Figure 1
Cluster of classes
4
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
● Class 1
● Class 2
● Class 3
● Centroids

3.3.3 Recommendation system

This is the core part of our project. It based on the result of clustering. We are going to use a K-means algorithm, as it uses the distance as the principal metric to locate the data in your respective clusters we need to be careful with scale, because we can give more "relevance" to large scale features and despite the low scale ones. This service allows you to type in the id of wine you are interested in like the picture shows below. After I choosing the service 1, the system will ask you to type in the ID of wine you are interested in. Here I choose the wine which’s ID is 1 as the example.



```
rvice.py

*** WELCOME TO FA20 255-01 TEAM 1 WINE LAB ! ***

1: Wine Recommendation
2: Wine Region Analysis
3: Wine Clustering
4: No thanks.

Please enter the service you need (1 ~ 4): 1

Please enter your favorite wine ID 1 ~ 178: 1

According to the chemical compostion, you might also like:

[38, 3, 49, 33, 41]

Do you want to 1: Try More   2: Back to Service Menu: 2
```

Then you can see the system recommend you the other ID of wines you might be interested in.

3.3.4 Use chemical analysis to determine the origin of wines

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. After you choose the service 3, you can type in the as much chemical of a wine as you know. Then the system will tell you the region of the wine you want to know.

Chapter 4: Conclusions

4.1. Difficulties Faced

- The dataset that the team selected is not big enough so we should set more function and algorithm to make the accuracy of our conclusions.

- The team also faced difficulties in implementing algorithms and obtain the best possible results considering all the factors such as the complexity of data, the volume of data, a large number of features, etc.

4.2 Things that work well

- We got our clustering and recommendation system perfectly.
- Selection of data cleaning and preprocessing tasks worked really well as a result of which we got the best out of our algorithms.

4.3Things that didn't work well

- The service 3 still can't work well.
- Team size is small and everyone should work more than other team.

4.4 Conclusion

- While implementing this project, we learned about a wide array of techniques, algorithms, and different preprocessing tasks involved in data analysis and prediction and how they affect the performance of the algorithms as a whole.
- We learned how to handle a imbalance dataset.
- We learned how to apply various algorithms and use recommendation models.
- We learned that the output of a model varies based on the requirement of the predictions.