Data Engineering HWK1 Explanatory Report

For Problem 1&2, please view the well-commented code for Food Web python file, the file name is Yichun_Zhou_DE_hwk_1_food_web.py.
For Problem 3&4, please view the well-commented code, named yichun_zhou_DE_HWK1_stats_plots.py, and the explanation below.

Data is retrieved from UCI Machine Learning Repository, link is provided in the reference. These data are the results of a chemical analysis of wines. The analysis determined the quantities of 13 constituents found in each of the three types of wines. All attributes are continuous, they are Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines and Proline. Below are some steps I have done.

1. Used pandas package to read and cleaned the dataset, the data has 178 rows, and 14 columns. Partial First five rows of the data are shown below.

	Label	Alcohol	Malic Acid	...	Hue	OD280/OD315 of diluted wines	Proline
0	1	14.23	1.71	...	1.04	3.92	1065
1	1	13.20	1.78	...	1.05	3.40	1050
2	1	13.16	2.36	...	1.03	3.17	1185
3	1	14.37	1.95	...	0.86	3.45	1480
4	1	13.24	2.59	...	1.04	2.93	735

2. Used describe function to get some simple statistics of the dataset. Please view code.

3. Seaborn Pairplot helped to show the relationship between each attribute and the classification Label. Figure 1 below is a quick scan of the pairplot, for a clear view, please view the wine_pairplot.png submitted with this report.

figure 1

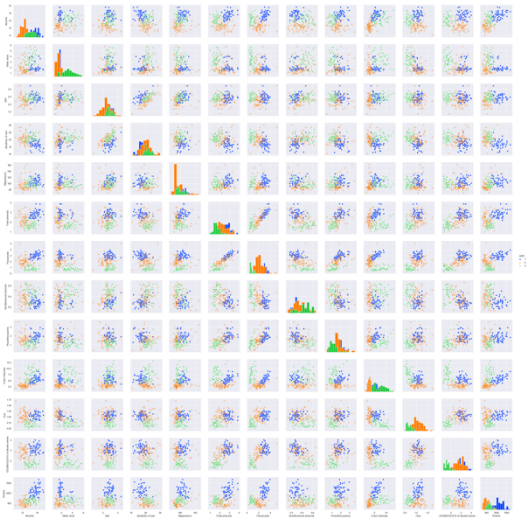

figure 2

4. Histograms to diagonals of Seaborn pairplot help better visualize the relationship. Figure 2 above is a quick scan of the pairplot woth histgram, for a clear view, please view the wine_hist_pairplot.png submitted with this report.

5. Next, Created a bivariate scatter with Seaborn. Used the attributes Alcohol and Flavanoids, set the hue with Label. View the scatter plot below (figure 3).

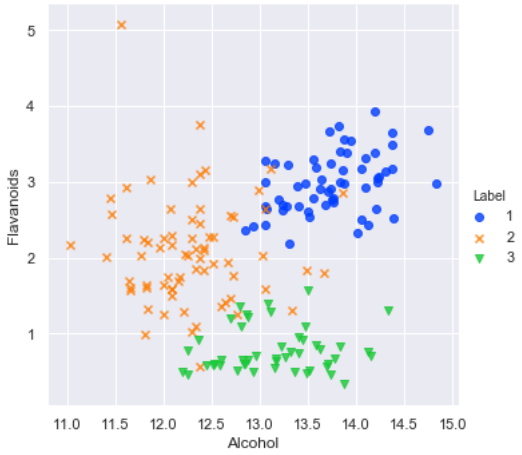

figure 3

6. Plot a violin graph, set x-axis with 'Label', and y-axis with 'Alcohol', View the violin plot below (figure 4).

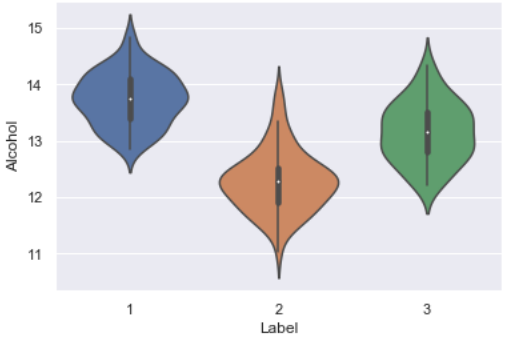

figure 4

7. Reduced dimensions with PCA with Pandas DataFrame, set 2 dimensions wanted. First five rows of the PCA DataFrame is shown below. Figure 5 shows how the transformed data looks.

	pca1	pca2	Label
0	318.562979	21.492131	1
1	303.097420	-5.364718	1
2	438.061133	-6.537309	1
3	733.240139	0.192729	1
4	-11.571428	18.489995	1

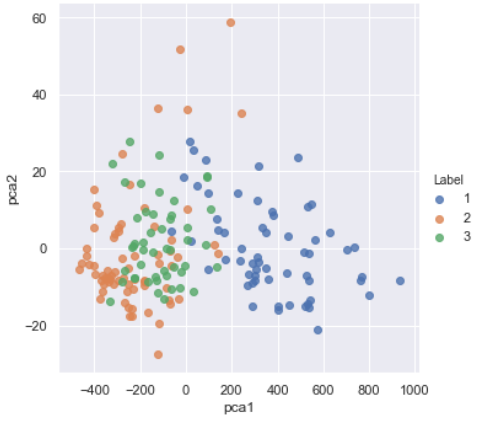

figure 5

8. Reduced dimensions with LDA method, set 2 dimensions wanted. First five rows of the LDA DataFrame is shown below. Figure 6 shows how the transformed data looks.

	lda1	lda2	Label
0	-4.700244	1.979138	1
1	-4.301958	1.170413	1
2	-3.420720	1.429101	1
3	-4.205754	4.002871	1
4	-1.509982	0.451224	1

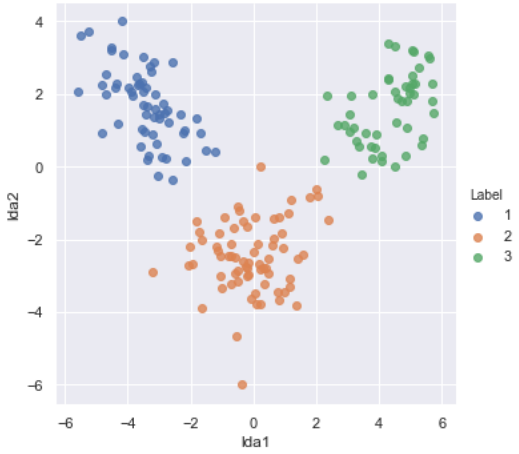

figure 6

9. Compared PCA vs. LDA methods by plotting the violin graphs. Figure 7 is the violin plot of pca1 and Label columns, Figure 8 is the violin plot of lda1 and Label columns.

figure 7

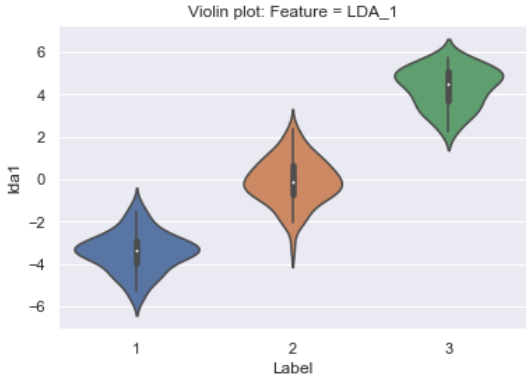

figure 8

10. Clustered With k-means and checked silhouette score of PCA and LDA, get PCA silhouette score 0.5722554756855064 and LDA silhouette score 0.663170.
Split the dataset into train/validation/test set, set test size equals 0.3, First five rows of the train DataFrame is shown below.

	lda1	lda2	Label
138	3.549934	0.915963	3
104	-1.129012	-2.326852	2
78	-0.707096	-2.123044	2
36	-2.758085	1.569704	1
93	-0.591029	-2.938454	2

Then, classified with SVM, set the penalty term C equals 0.8, get the f1 score for SVM classifier equals 0.864228. Change the C to 1, get f1 score for SVM classifier equals 0.885519.
Finally used RandomForestClassifier to tune the data, get the f1 score for SVM classifier equlas 0.981493, which is a good result.

References

http://archive.ics.uci.edu/ml/datasets/wine

http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data