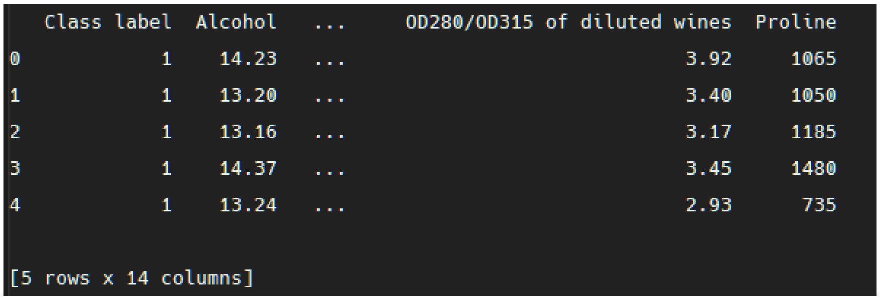
Joseph Loss (loss2)

IE598 MLF F18

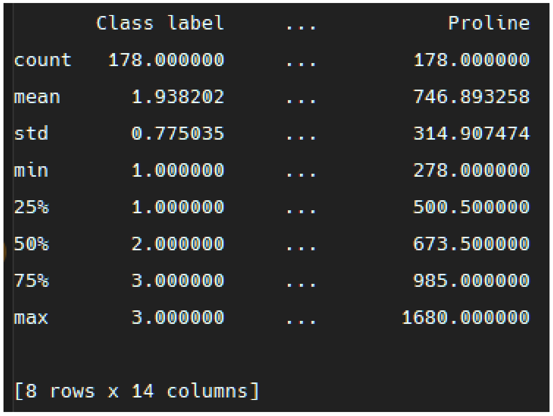
Module 5 Homework (Dimensionality Reduction)

Part 1: Exploratory Data Analysis



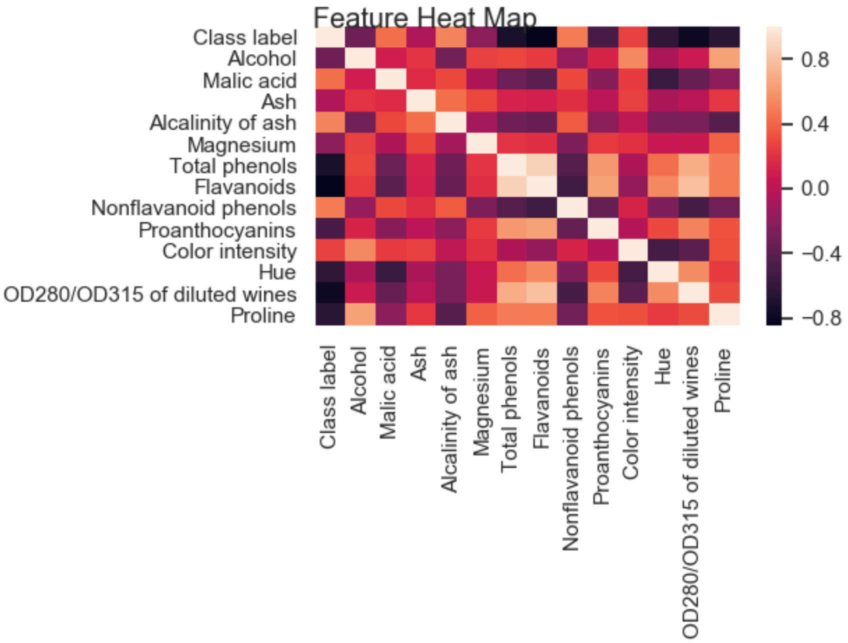
```
   Class label  Alcohol    ...     OD280/OD315 of diluted wines  Proline
0            1    14.23    ...                             3.92     1065
1            1    13.20    ...                             3.40     1050
2            1    13.16    ...                             3.17     1185
3            1    14.37    ...                             3.45     1480
4            1    13.24    ...                             2.93      735

[5 rows x 14 columns]
```

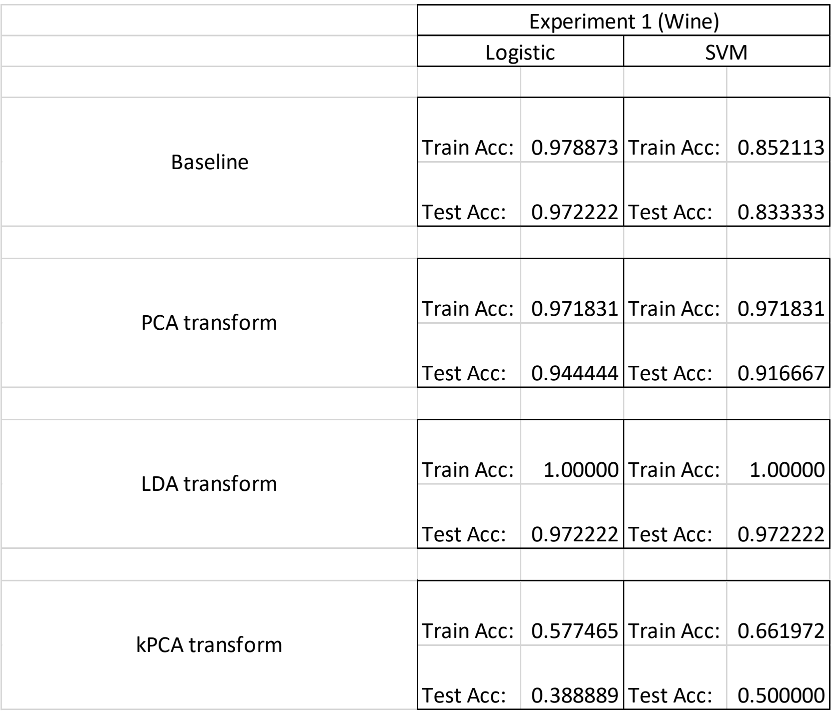


```
       Class label      ...          Proline
count   178.000000      ...       178.000000
mean      1.938202      ...       746.893258
std       0.775035      ...       314.907474
min       1.000000      ...       278.000000
25%       1.000000      ...       500.500000
50%       2.000000      ...       673.500000
75%       3.000000      ...       985.000000
max       3.000000      ...      1680.000000

[8 rows x 14 columns]
```



Part 2 - 5: Logistic regression classifier v. SVM classifier



	Experiment 1 (Wine)			
	Logistic		SVM	
Baseline	Train Acc:	0.978873	Train Acc:	0.852113
	Test Acc:	0.972222	Test Acc:	0.833333
PCA transform	Train Acc:	0.971831	Train Acc:	0.971831
	Test Acc:	0.944444	Test Acc:	0.916667
LDA transform	Train Acc:	1.00000	Train Acc:	1.00000
	Test Acc:	0.972222	Test Acc:	0.972222
kPCA transform	Train Acc:	0.577465	Train Acc:	0.661972
	Test Acc:	0.388889	Test Acc:	0.500000

Part 6: Conclusions

On the untransformed data, the logistic regression model performed better than the SVM model. In regards to overall performance increase, it would seem that the LDA transformation performed better than all other models. However, one might notice that the training accuracy score is 100% and the testing accuracy score is 97.22% for both the LogReg and the SVM models, which is an alarming sign of overfitting the data.

If I had to decide which transformation to use, I would probably choose the kPCA ‘rbf-kernel’ transformation, as this seemed to walk the line in terms of accuracy while also refraining from over/under-fitting the data. Note that I tested several different inputs for Gamma, and it seemed that Gamma=1.0 led to the best individual score for each model. As Gamma increased (towards 15.0, as instructed), the scores for each model both converged on 45.75% for the training set and 38.39% for the testing set. This seemed to underfit the data slightly, so I decided to use Gamma=1.0 for reporting my results.

Note: In the above case and for the PCA/LDA models as well, it may be apparent that the scores for both the LogReg and SVM models are extremely similar, if not identical. I believe this is because I used the LinearSVC function for the SVM model, which is a very similar model to Logistic Regression. As a result, one might expect these models to return very similar results (as they in fact did).

Part 7: Appendix

https://github.com/chicago-joe/IE598_F18_HW5