Jingxia Zhu (jingxia6)

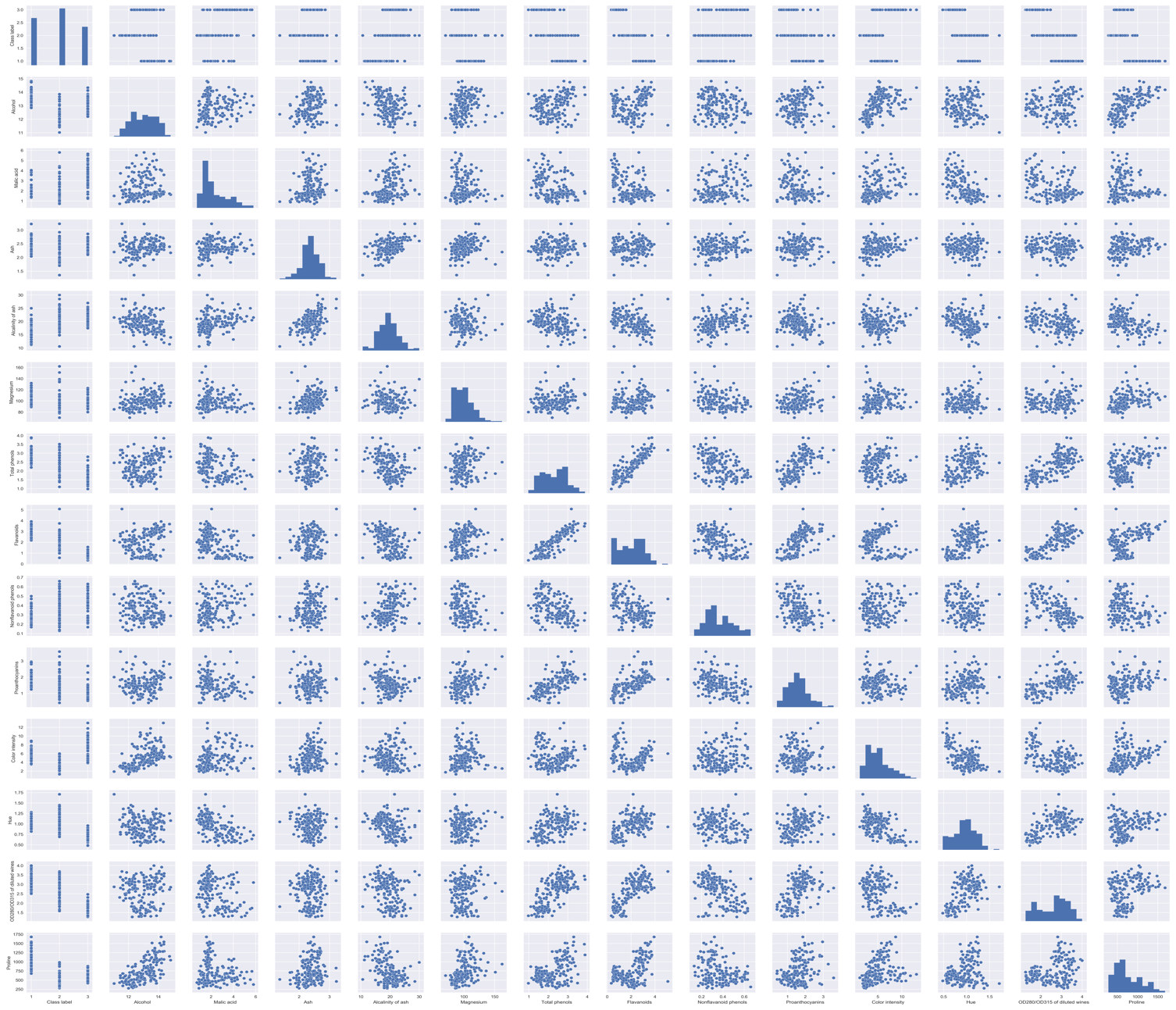
IE598 MLF F18

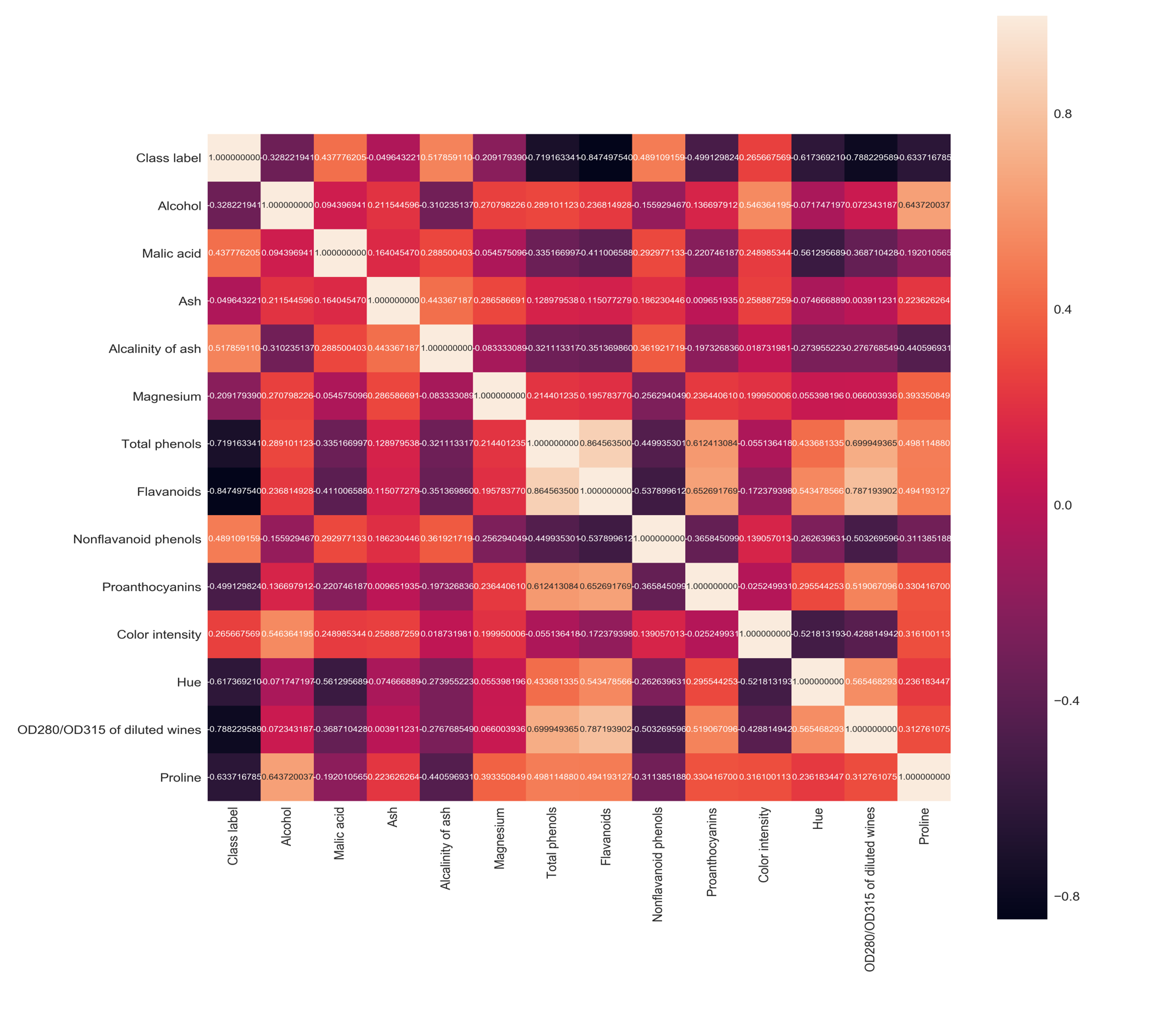
Module 5 Homework (Dimensionality Reduction)

**Part 1: Exploratory Data Analysis**

Describe the data sets sufficiently using the methods and visualizations that we used previously. Include any output, graphs, tables, heatmaps, box plots, etc. that you think is necessary to represent the data. Label your figures and axes. DO NOT INCLUDE CODE, only output figures!

Split data into training and test sets. Use random\_state = 42. Use 80% of the data for the training set. Use the same split for all experiments.

------------------------------------------------------------------------------------------------------------------------------------------

****

**Part 2: Logistic regression classifier v. SVM classifier - baseline**

Fit a logistic classifier model to both datasets using SKlearn. Calculate its accuracy score for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

Fit an SVM classifier model to both datasets using SKlearn. Calculate its accuracy score for both in sample and out of sample (train and test sets). (You may use CV accuracy score if you wish).

------------------------------------------------------------------------------------------------------------------------------------------

﻿Logistic regression train/test accuracies 1.000/1.000

﻿SVM train/test accuracies 1.000/0.944

**Part 3: Perform a PCA**

Refit both a logistic and SVM classifier on the PCA transformed datasets. You may choose to use only 2 components, or select a higher appropriate intrinsic dimension. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

------------------------------------------------------------------------------------------------------------------------------------------

﻿PCA\_LR train/test accuracies 0.972/0.944

PCA\_SVM train/test accuracies 0.972/0.944

**Part 4: Perform and LDA**

Refit both a logistic and SVM classifier on the LDA transformed datasets. You may choose to use only 2 discriminants, or select a higher appropriate number. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

------------------------------------------------------------------------------------------------------------------------------------------

﻿﻿LDA\_LR train/test accuracies 1.000/0.972

LDA\_SVM train/test accuracies 1.000/0.944

**Part 5: Perform a kPCA**

Refit both a logistic and SVM classifier on the kPCA transformed datasets. Use the rbf kernel. Test several different values for Gamma. Calculate accuracy scores for both in sample and out of sample (train and test sets) on both datasets.

------------------------------------------------------------------------------------------------------------------------------------------

﻿﻿﻿Gamma: 0.001

KPCA\_LR train/test accuracies 0.478873/0.555556

KPCA\_SVM train/test accuracies 0.873239/0.861111

Gamma: 0.01

KPCA\_LR train/test accuracies 0.971831/0.972222

KPCA\_SVM train/test accuracies 0.978873/0.944444

Gamma: 0.1

KPCA\_LR train/test accuracies 0.978873/1.000000

KPCA\_SVM train/test accuracies 0.971831/0.972222

Gamma: 0.5

KPCA\_LR train/test accuracies 0.464789/0.444444

KPCA\_SVM train/test accuracies 0.802817/0.666667

Gamma: 1

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.500000/0.388889

Gamma: 2

KPCA\_LR train/test accuracies 0.450704/0.388889

KPCA\_SVM train/test accuracies 0.429577/0.388889

Gamma: 3

KPCA\_LR train/test accuracies 0.415493/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 4

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 5

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 6

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 7

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 8

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 9

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 10

KPCA\_LR train/test accuracies 0.401408/0.388889

KPCA\_SVM train/test accuracies 0.401408/0.388889

Gamma: 15

KPCA\_LR train/test accuracies 0.415493/0.388889

KPCA\_SVM train/test accuracies 0.429577/0.388889

Gamma: 20

KPCA\_LR train/test accuracies 0.429577/0.388889

KPCA\_SVM train/test accuracies 0.429577/0.388889





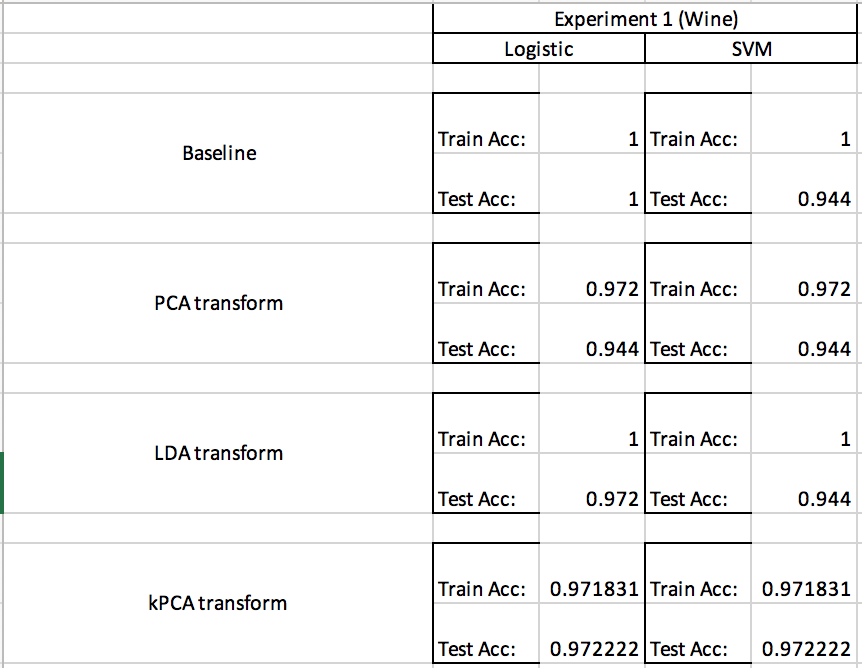
**Part 6: Conclusions**

Write a short paragraph summarizing your findings. Which model performs best on the untransformed data? Which transformation leads to the best performance increases? Report your results using the Results worksheet format. Embed the completed table in your report.

------------------------------------------------------------------------------------------------------------------------------------------

The wine data set is pretty linear itself originally. So, using transformations will not make the model perform better on this data set. After comparing different gammas on kPCA, the best one I would choose is 0.1 as it gives the highest accuracy score.

Logistic Regression performs best on the untransformed data. Transformations, in this case, do not leads to better performances, or great performance increases. The following sheet includes my results of accuracy scores:



**Part 7: Appendix**

Link to github repo: https://github.com/jzhuuhzj/IE598\_F18\_HW5.git