Stanly Gomes
Student ID: 801118166

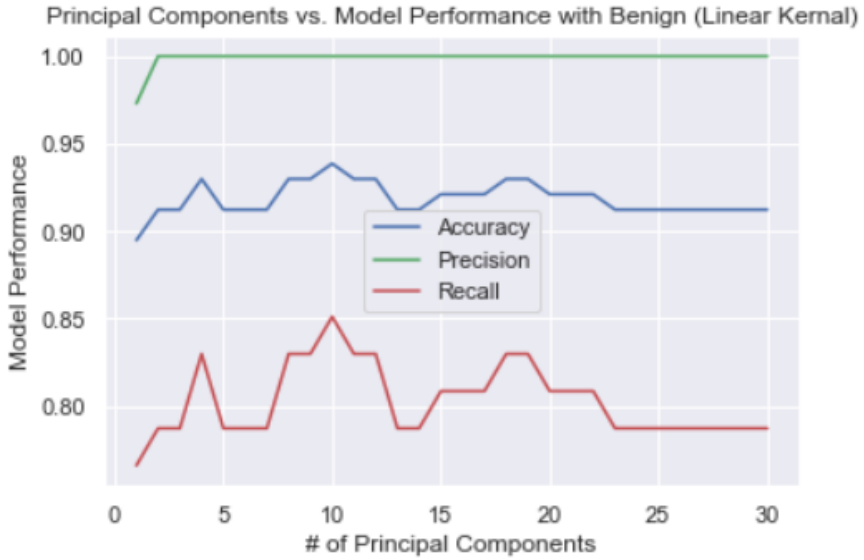
Homework 4

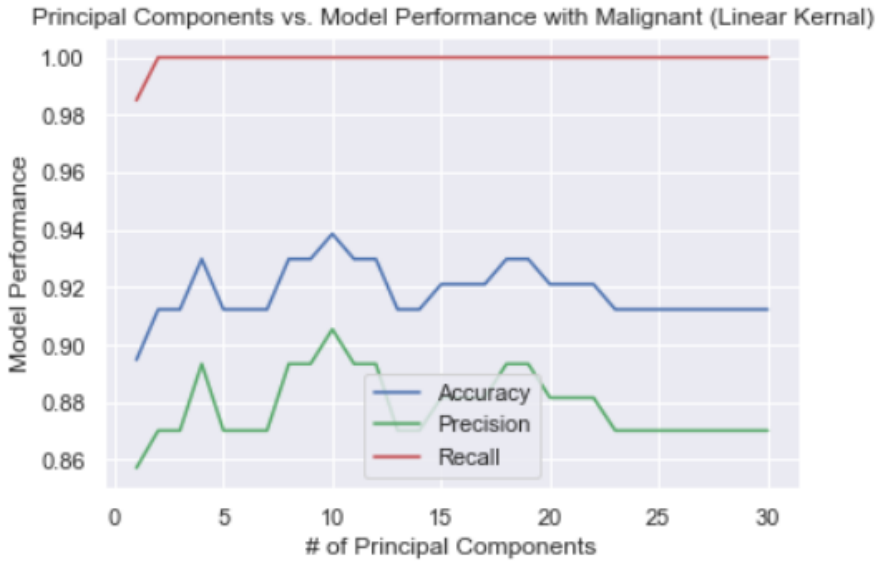
GitHub Repository Link: https://github.com/NaraPvP/IntroToML

For all the problems, an 80/20 training split was done on the dataset. Along with this, the features were scaled using standardization.

Problem 1:

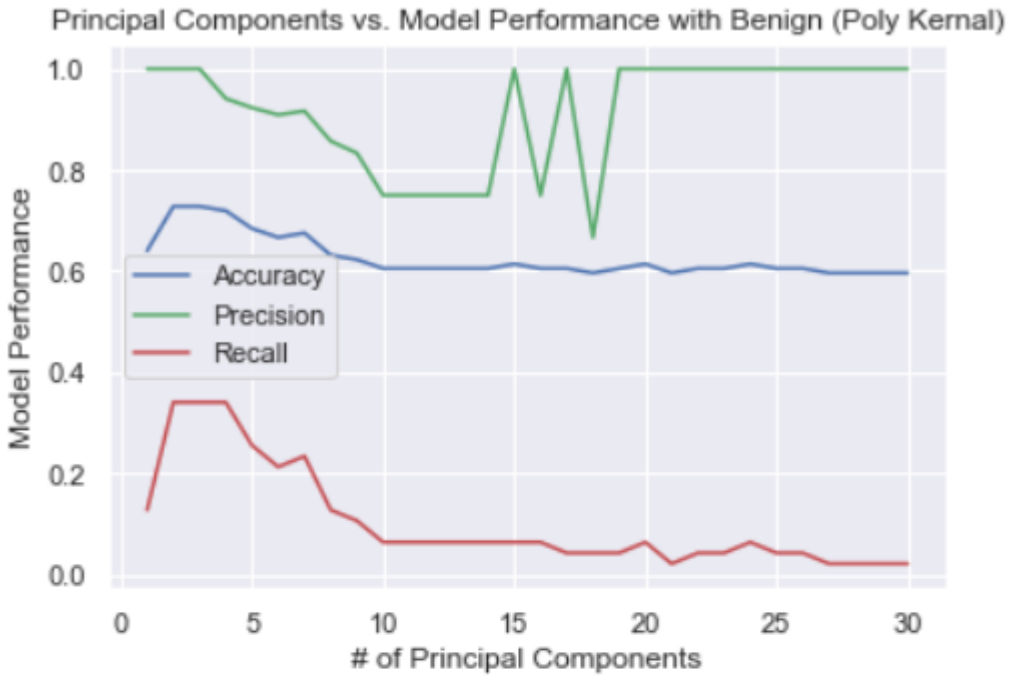
For the SVM classifier using a linear kernel, the optimal number of principal components was 10 with an accuracy of 93.86%.

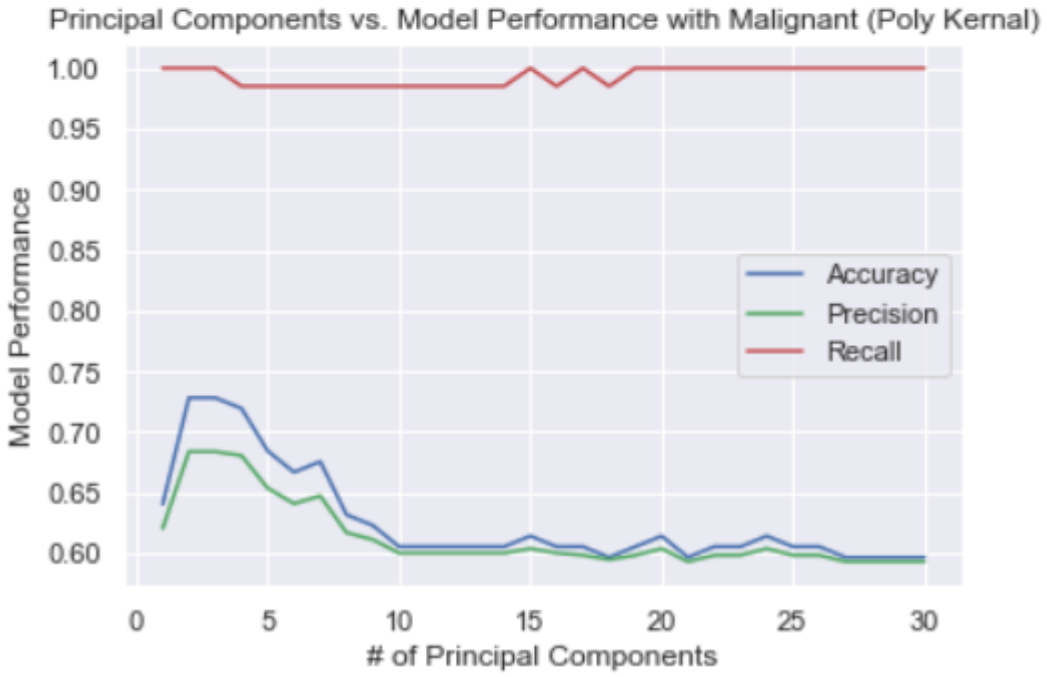




To better capture non-linearities in the data, the polynomial and radial basis function kernel were tested to determine if higher accuracies could be yielded. For the polynomial kernel, the optimal number of principal components was 2 with an accuracy of 72.81%.

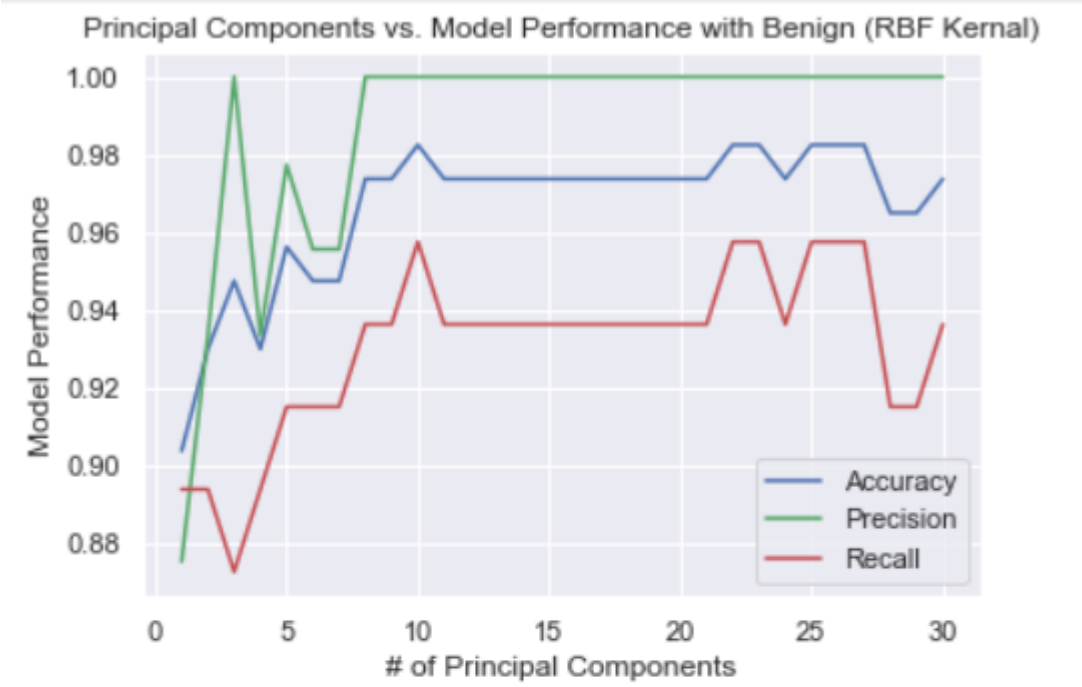
Stanly Gomes
Student ID: 801118166

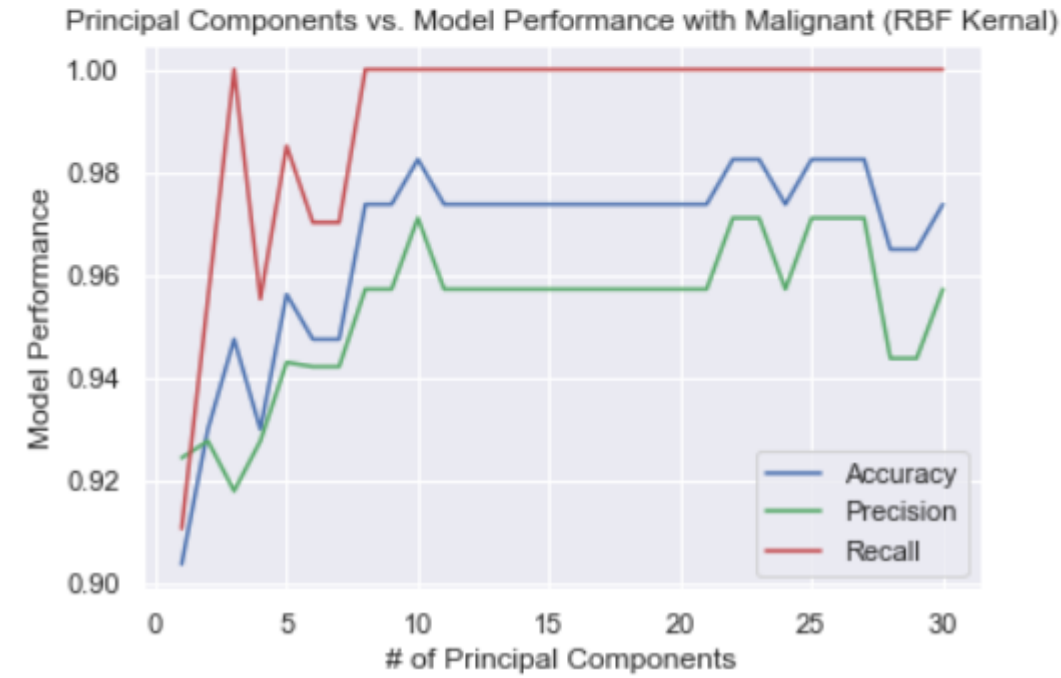




For the radial basis function kernel, the optimal number of principal components was 10 with an accuracy of 98.25%.

Stanly Gomes
Student ID: 801118166





From all of the kernels explored, the radial basis function kernel outperformed the others with the cancer dataset used.

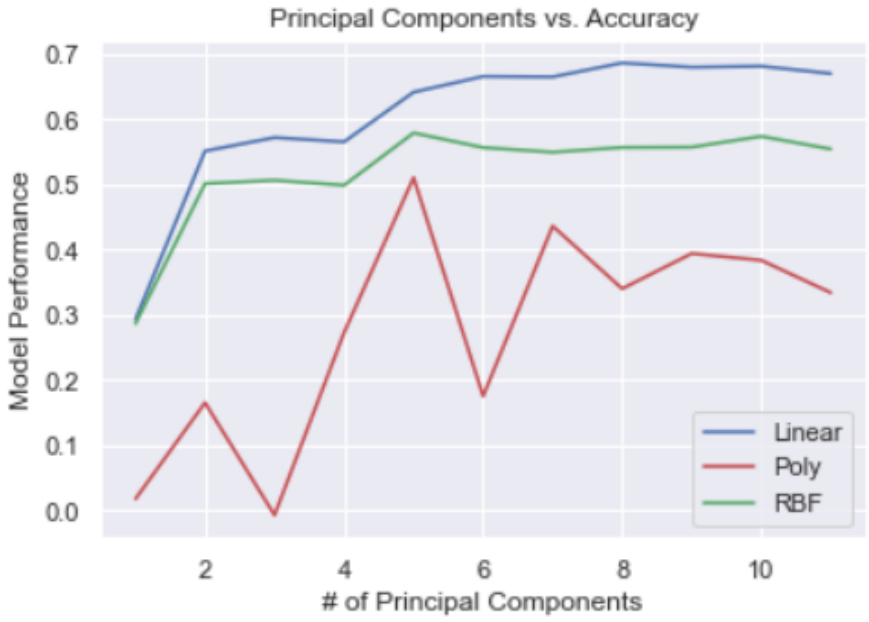
Compared to logistic regression performed in Homework 3, they perform very similarly when the RBF kernel is used. Based on both models' most optimal principal components, logistic regression slightly performs better than SVM.

Problem 2:

Stanly Gomes
Student ID: 801118166

Due to the increased dimensionality of the dataset compared to the one used in the Canvas example, the plot will just show the predictions made by each SVR kernel:

Afterwards, PCA feature extraction was used to determine the optimal number of principal components that will yield the highest regression accuracy. For the linear kernel, 8 principal components provided the highest accuracy of 68.69%. For the polynomial kernel, 5 principal components yielded the highest accuracy of 51.09%. For the RBF kernel, 5 principal components gave the highest accuracy of 57.95%. Compared to linear regression that was used in a previous homework, the SVR model performs drastically worse than linear regression. This plot shows the different kernel's accuracy in the SVR model compared to the number of principal components:



Stanly Gomes
Student ID: 801118166

Contrasting to the previous problem, the linear kernel performed the best out of the tested kernels. This may be due to the housing dataset being more linear than the cancer dataset.