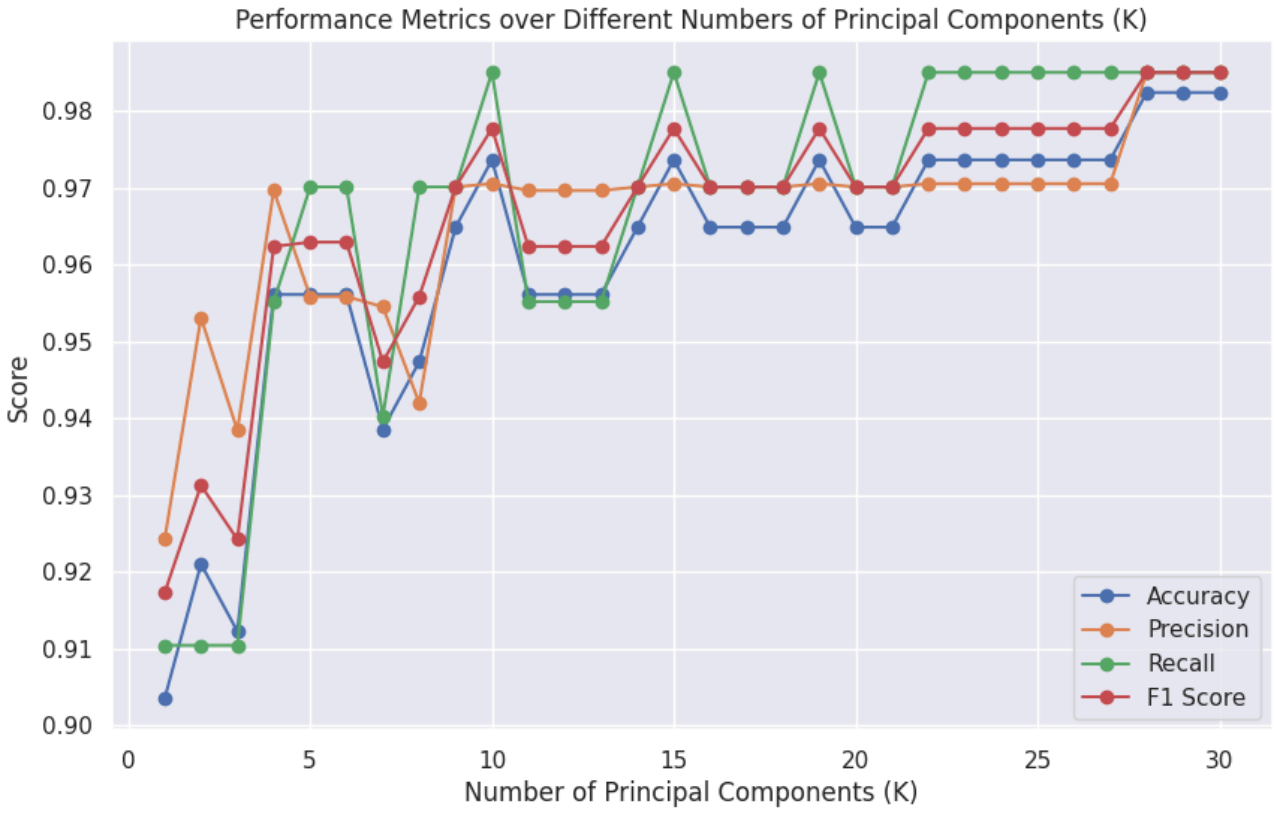
Nathon Tadeo
Student ID: 801265462
11/1/24
Homework 4
Github:https://github.com/nathon-tadeo/Intro-to-ML/blob/main/homework_4_intro_to_ml.ipynb

Problem 1 (50pts):
Use the cancer dataset to build an SVM classifier to classify the type of cancer (Malignant vs. benign). Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K).

1. Identify the optimum number of K, principal components that achieve the highest classification accuracy.

Since we used the Cancer set using logistic regression for Homework 3, setting up the dataset was facilitated. All the scaling and plotting were mostly the same for this dataset. The PCA feature extraction was also the same, but we used the SVM classifier instead of a logistic regression model. With all the changes, the code had the best accuracy with a 'k' of 28. This equated to an accuracy of 0.9824.

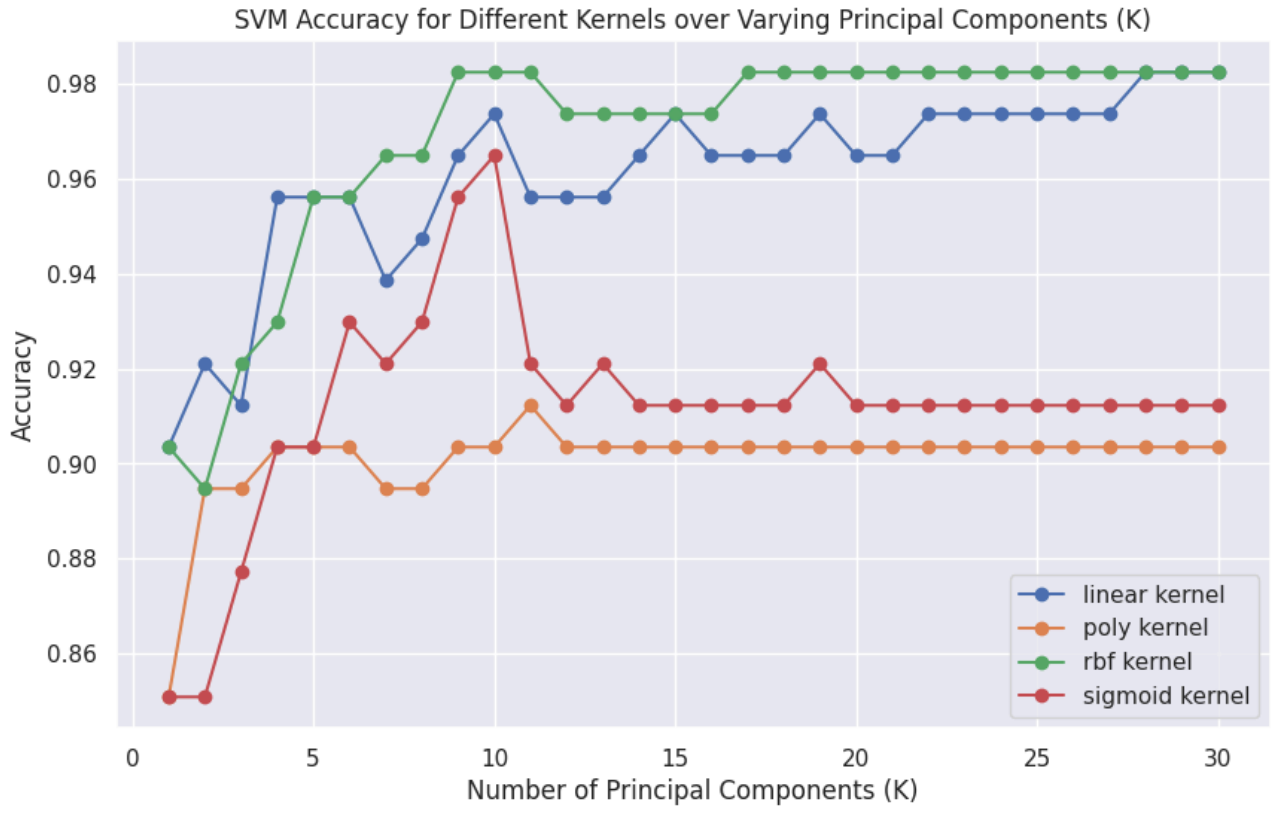
2. Plot your classification accuracy, precision, and recall over a different number of Ks.



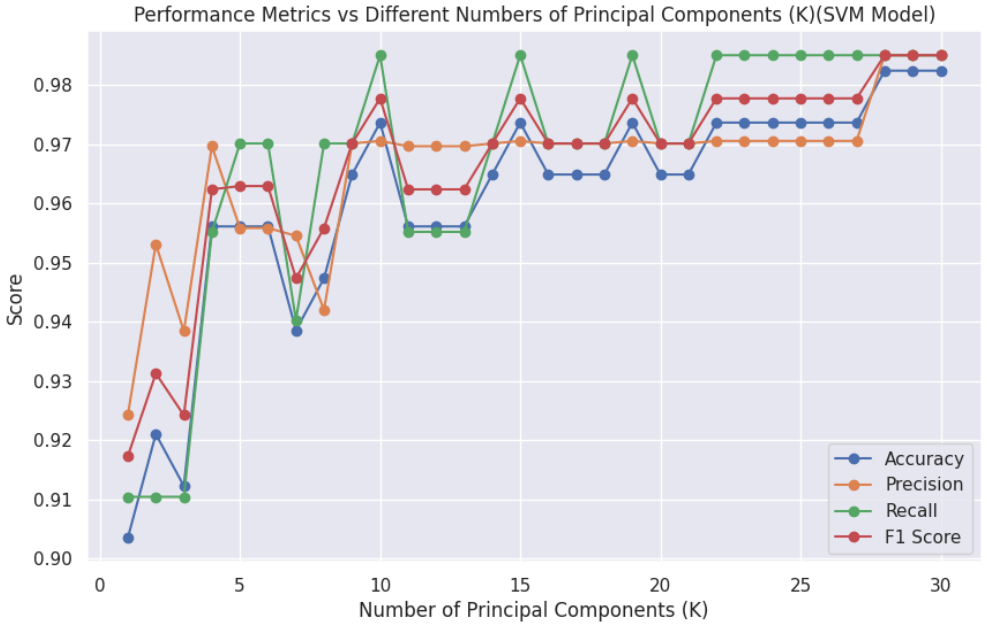
3. Explore different kernel tricks to capture non-linearities within your data. Plot the results and compare the accuracies for different kernels.

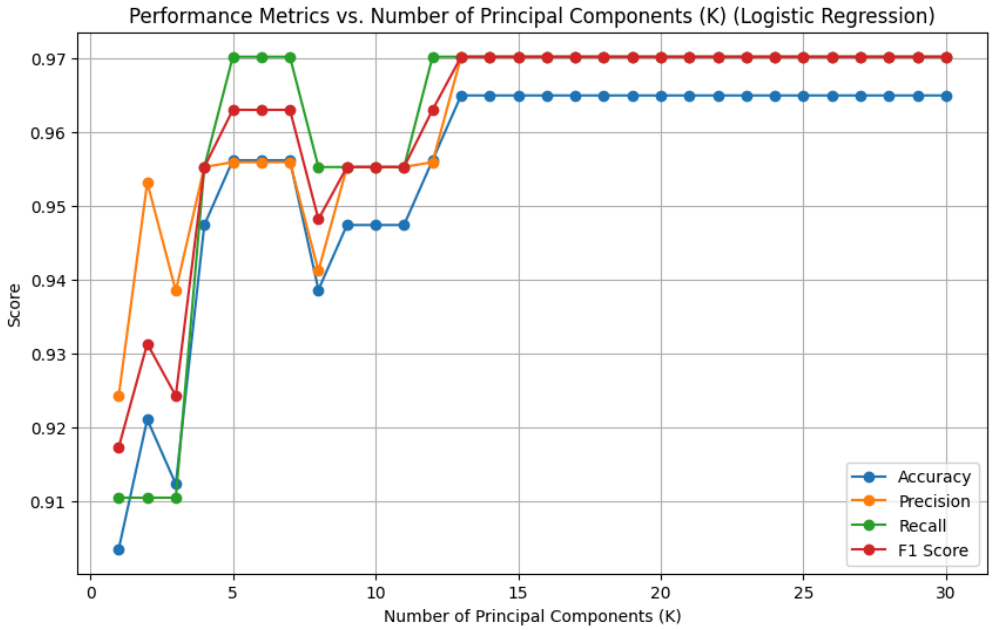
For this problem, different kernel types were explored to capture non-linearities in the data. This analysis involved applying PCA to reduce dimensionality and subsequently training the SVM using various kernel functions. The kernels evaluated included linear, polynomial (poly), radial basis function (rbf), and sigmoid. Both the linear and RBF kernels achieved the highest accuracy, meaning they are effective for this dataset. The RBF kernel utilized a principal component count (K) of 9, while the linear model achieved its best accuracy with K set to 28.

Kernel: linear
Best K: 28
Accuracy for best K: 0.9825
Kernel: poly
Best K: 11
Accuracy for best K: 0.9122
Kernel: rbf
Best K: 9
Accuracy for best K: 0.9825
Kernel: sigmoid
Best K: 10
Accuracy for best K: 0.9649



4. Compare your results against the logistic regression that you have done in homework 3. When comparing the Support Vector Machine (SVM) model to the logitistc regression model, the logistics regression model had a accuracy score ranging from 0.90 to 0.97 while the SVM model with PCA ranged from 0.90 to 0.99. Both graphs have a steep upwards trend as the principal components approached five. The logistic regression model metrics plateaued out after about 13 principal components, while the SVM model continues to fluctuate across the entire range of k





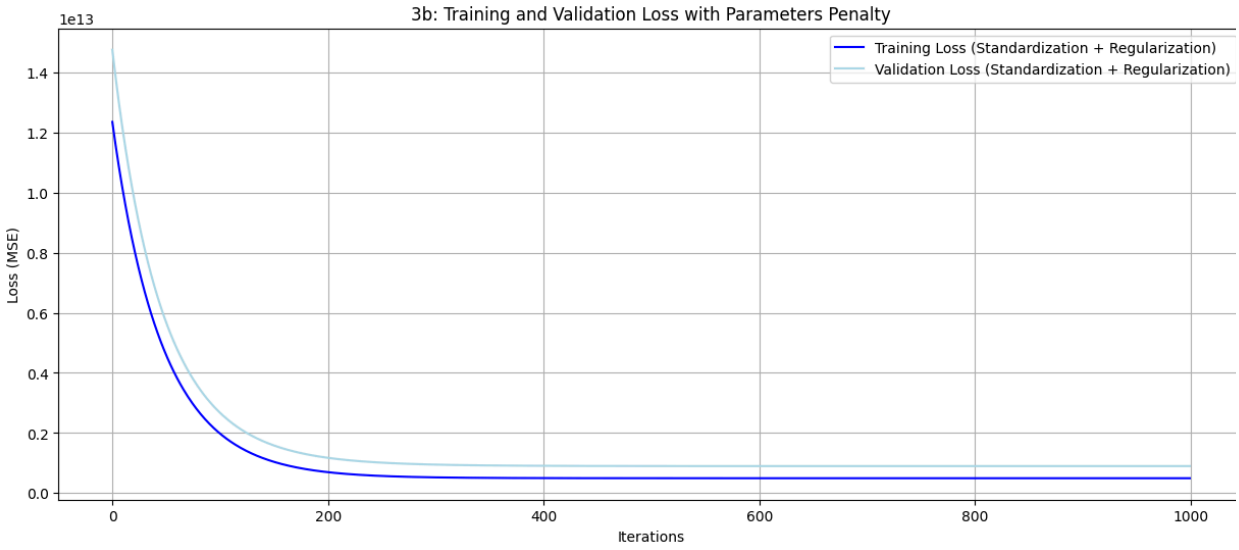
Problem 2 (50pts):
Develop a SVR regression model that predicts housing price based on the following input variables: Area, bedrooms, bathrooms, stories, mainroad, guestroom, basement, hotwaterheating, airconditioning, parking, prefarea.

1. Plot your regression model for SVR similar to the sample code provided on Canvas.

After importing the house data set with all the necessary training and processing, the four kernels were set up to be displayed on the support vector regression model. Using the model from canvas for the SVR, the dataset was used to fill in the model. Most of the models did not have any success with the housing data. No trends could be extrapolated from the plot. This could be due to the mass amounts of data contained in the housing dataset. Further analysis or data preprocessing might be required to enhance the model's effectiveness.

2. Compare your results against linear regression with regularization loss that you already did in homework 2.

While the linear regression with regularization provided stable and interpretable results, the SVR model showcased non-linear patterns, but the plotted SVR doesn't have interpreteable results.



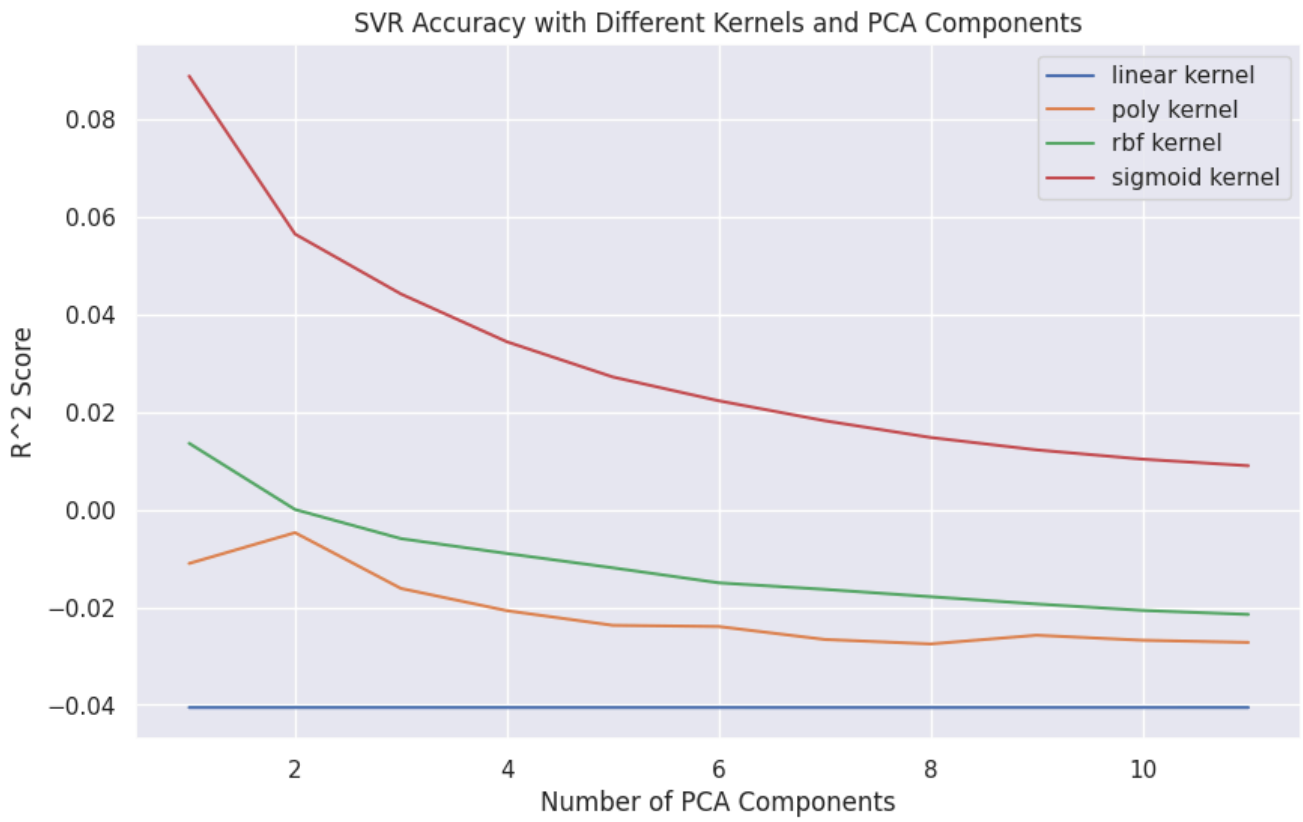
3. Use the PCA feature extraction for your training. Perform N number of independent training (N=1, …, K). Identify the optimum number of K, principal components that achieve the highest regression accuracy.

Similar to the first problem, the housing dataset went through the process of the PCA and SVR regression model. The model achieved the best accuracy when k was '1'. This equated to an accuracy of -0.041.

4. Explore different kernel tricks to capture non-linearities within your data. Plot the results and compare the accuracies for different kernels.

When exploring the different kernels to capture non-linearities, the k values ranged from 1 to 2 for the best accuracies. The rbf and the sigmoid model resulted in the best accuracies being 0.014, and 0.089 respectively. Unfortunately, the linear and poly gave negative values. For accuracy measurement, the R^2 score was used to evaluate how well each model explained the variance in the data.

linear | Best K: 1 | Accuracy: -0.041
poly | Best K: 2 | Accuracy: -0.005
rbf | Best K: 1 | Accuracy: 0.014
sigmoid | Best K: 1 | Accuracy: 0.089


SVR Accuracy with Different Kernels and PCA Components
linear kernel
poly kernel
rbf kernel
sigmoid kernel
R^2 Score
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
2
4
6
8
10
Number of PCA Components