Name: Abdirahim Ahmed
Student ID: 801338570

Homework 4

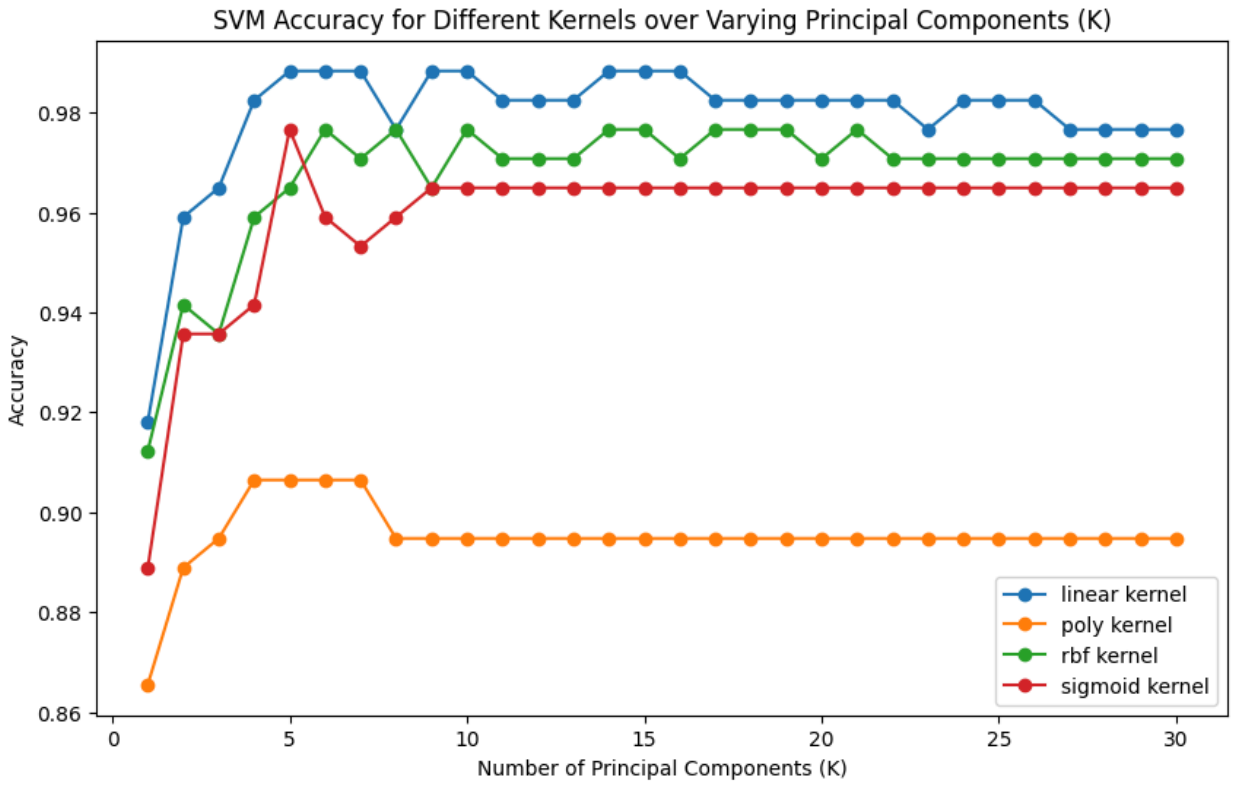
Githubrepository:https://github.com/Abdiirahim/ECGR-4105-Intro-to-ML/tree/main/Homework4

1. 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.

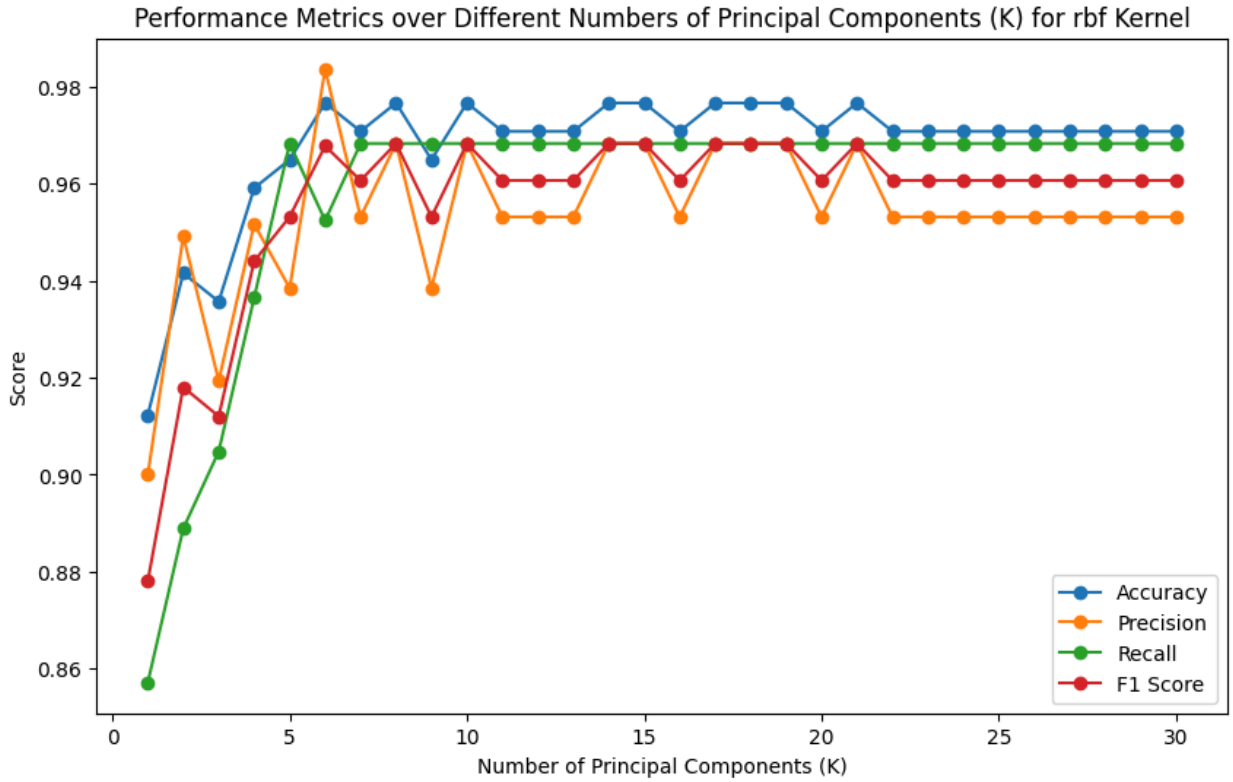
    The SVM classifier achieved its highest accuracy using around 5–10 principal components. Beyond 10 components, the accuracy stabilized with diminishing improvements.The predictive information in the dataset can be captured with relatively few components, indicating high feature redundancy

    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.

Name: Abdirahim Ahmed
Student ID: 801338570



The linear kernel consistently performed best, reaching an accuracy of approximately 98%.

The rbf kernel followed closely, with accuracy around 97%, demonstrating it could capture some non-linear relationships.

The polynomial and sigmoid kernels showed lower performance, with polynomial stabilizing around 90% and sigmoid underperforming at below 90%.

These results indicate that the cancer dataset is largely linearly separable, hence favoring the linear kernel.

4. Compare your results against the logistic regression that you have done in homework 3.

The logistic regression model in Homework 3 achieved a similar high accuracy of around 97% for the cancer dataset, both with and without regularization (L2 penalty).

The SVM linear kernel's performance aligns closely with logistic regression, suggesting that logistic regression was already well-suited for this linearly separable dataset. SVM with the linear kernel offers a slight edge in terms of flexibility with different regularization options.

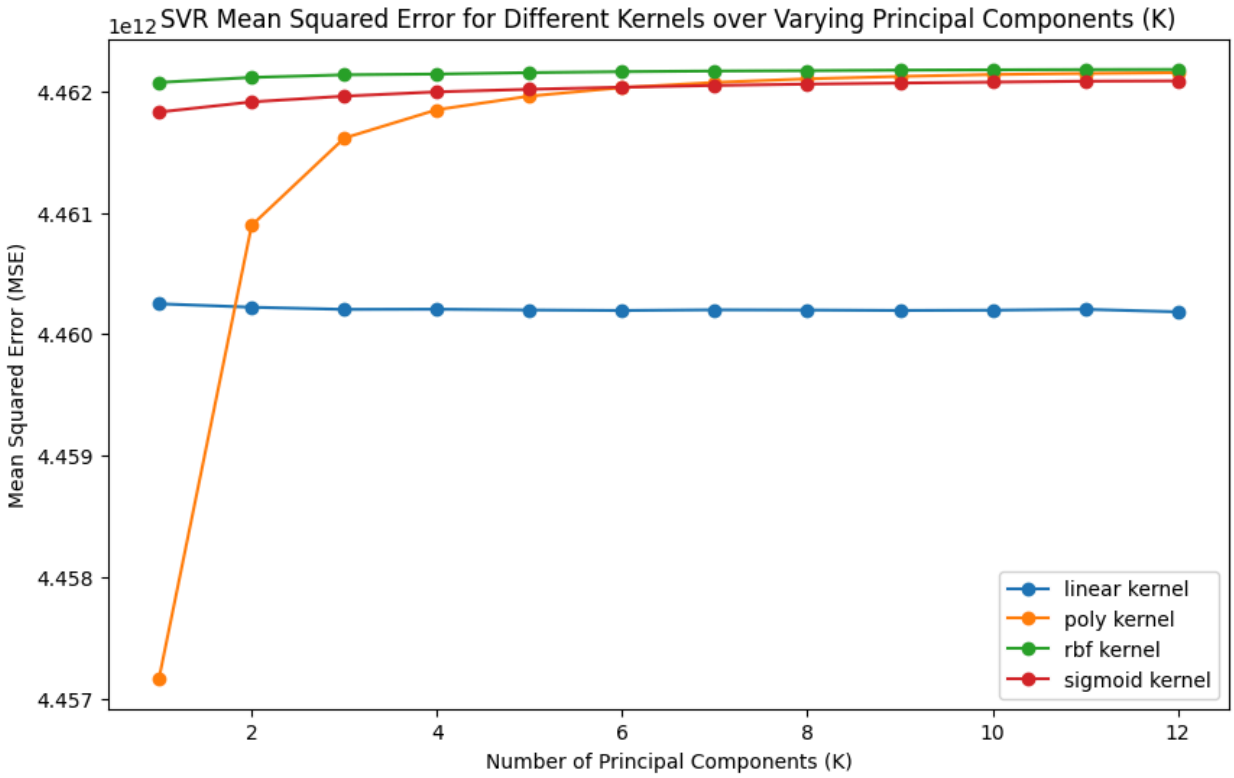
Name: Abdirahim Ahmed
Student ID: 801338570

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, hot water heating, air conditioning, parking, prefarea

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



2. Compare your results against linear regression with regularization loss that you already did in homework1.

In Homework 1, linear regression with regularization (likely Ridge regression) provided accurate and stable predictions for the dataset. The SVR model with a linear kernel in this analysis produced comparable results, achieving the lowest

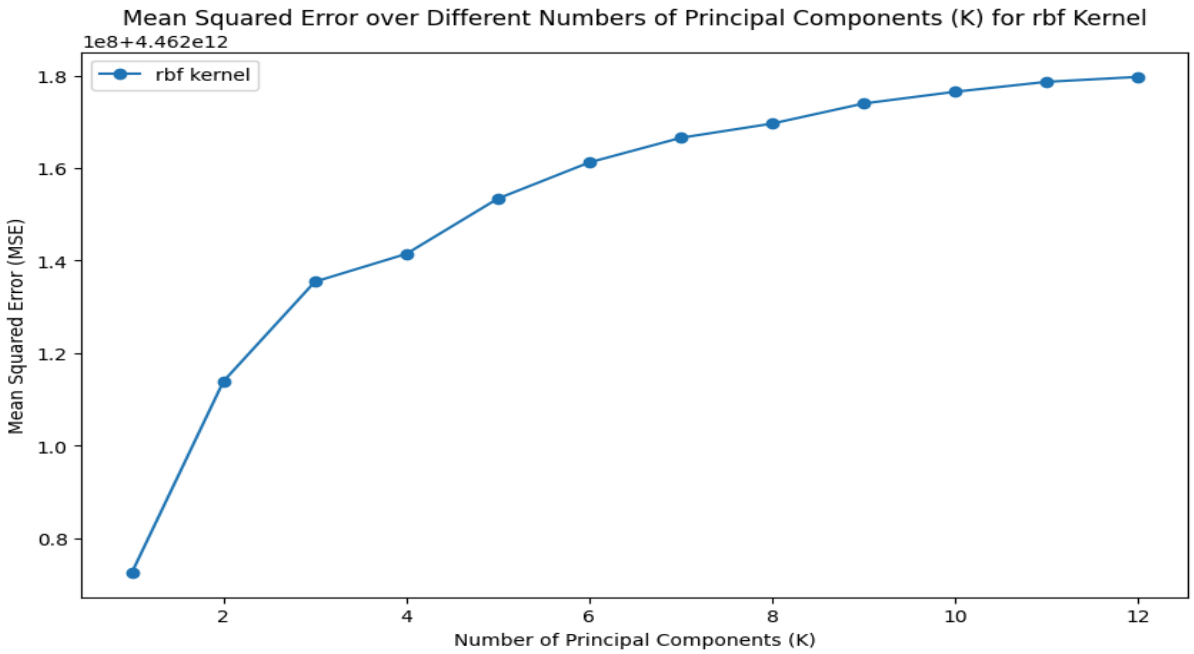
Name: Abdirahim Ahmed
Student ID: 801338570

Mean Squared Error (MSE) and confirming that the data's relationships are primarily linear. Non-linear kernels (RBF, polynomial, sigmoid) did not improve accuracy, indicating that additional complexity is unnecessary for this dataset.

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.

Using PCA, we observed that 2–3 principal components were sufficient to achieve stable and low MSE for the SVR model. Adding more components beyond this number did not significantly improve accuracy, indicating that most of the predictive information is captured in these first few components.

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



Among the SVR kernels tested, the linear kernel achieved the best performance, with the lowest and most stable MSE. Non-linear kernels (RBF, polynomial, and sigmoid) consistently resulted in higher MSE values, suggesting that they add complexity without enhancing model accuracy. This reinforces that the dataset's structure is primarily linear, making the linear kernel the most appropriate choice