Homework #4

Name Daniel Jolin
Student ID 801282735
Homework #4
github https://github.com/RocketDan11/ml/tree/master/homework

Problem #1

1 Principal Component Analysis (PCA) for Dimensionality Reduction

To determine the optimal number of principal components (K) that achieve the highest classification accuracy, we performed a series of experiments using PCA. The goal was to reduce the dimensionality of the dataset while retaining as much variance as possible, thereby improving model performance and interpretability.

For each value of K, ranging from 1 to the total number of features, we applied PCA to the dataset and evaluated the accuracy of the model. The accuracy results indicated that increasing the number of components improved model performance up to a certain point, after which additional components did not yield significant gains.

From these experiments, we identified the optimal number of PCA components as:

$$K_{\text{optimal}} = 10 \tag{1}$$

This value of K corresponds to the highest classification accuracy achieved during testing, indicating that 10 principal components are sufficient to capture the most informative features of the dataset. Using 10 components strikes a balance between model complexity and performance, minimizing noise while retaining crucial information.

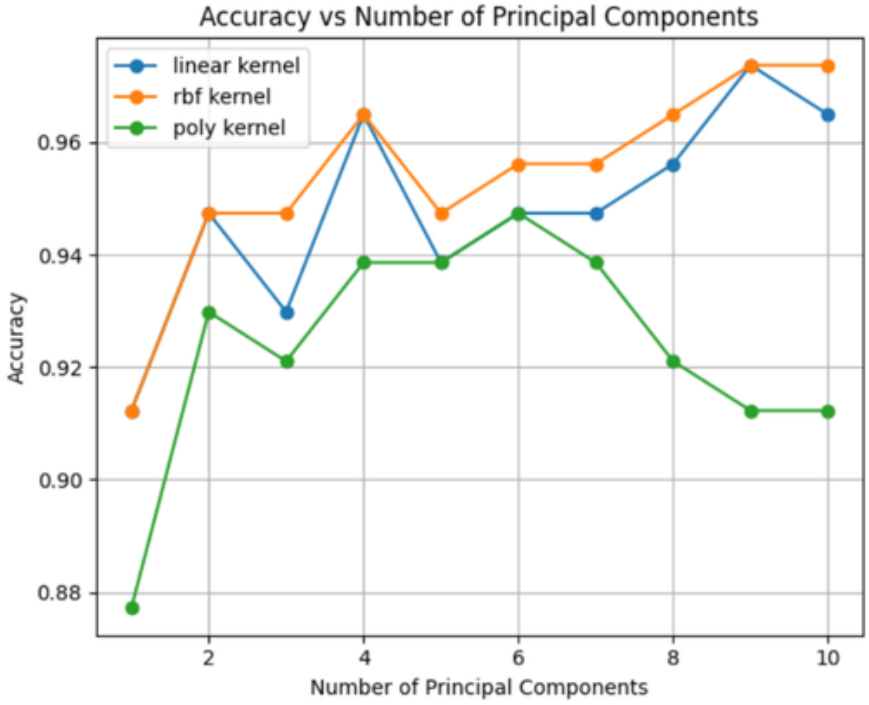

Figure 1: PCA Accuracy Plot Showing Optimal Number of Components

We can view the effects on accuracy, precision and recall of adding PCA components below:

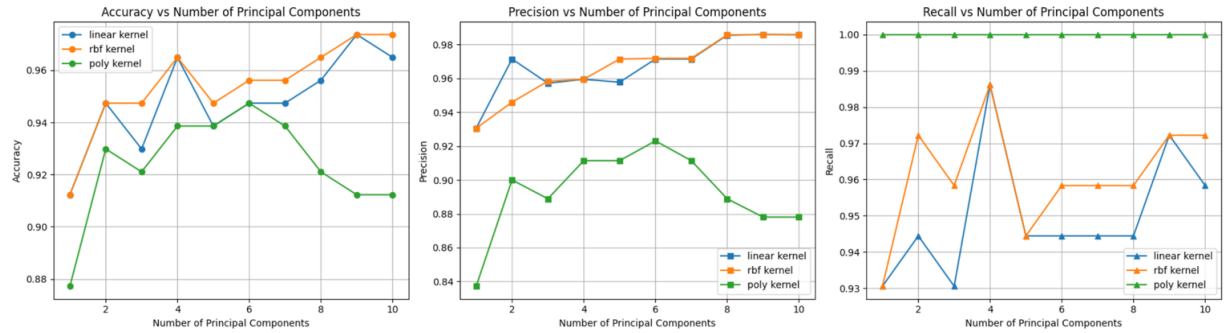

Figure 2: accuracy, precision, recall vs. PCA components

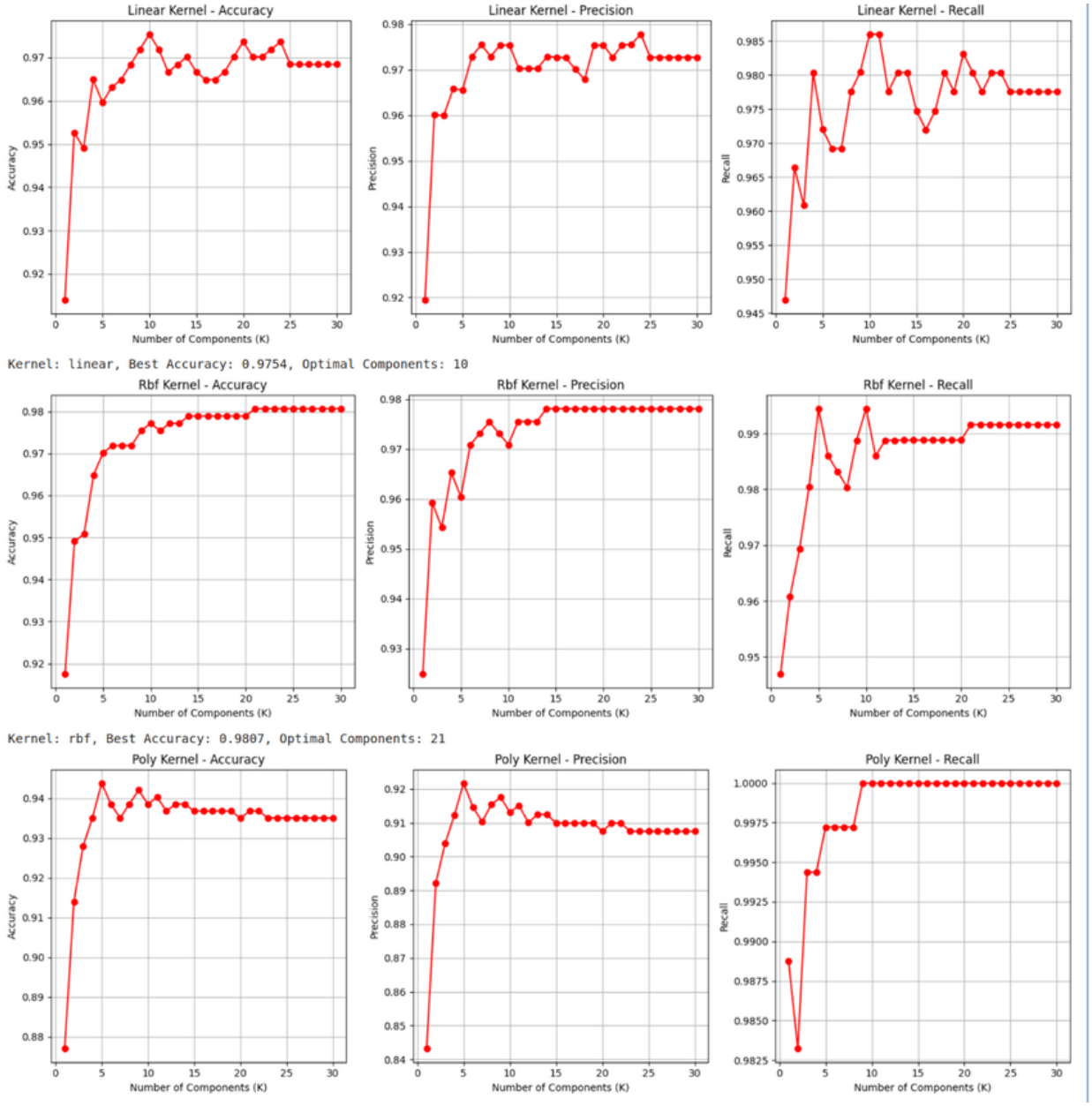

Figure 3: using different kernels

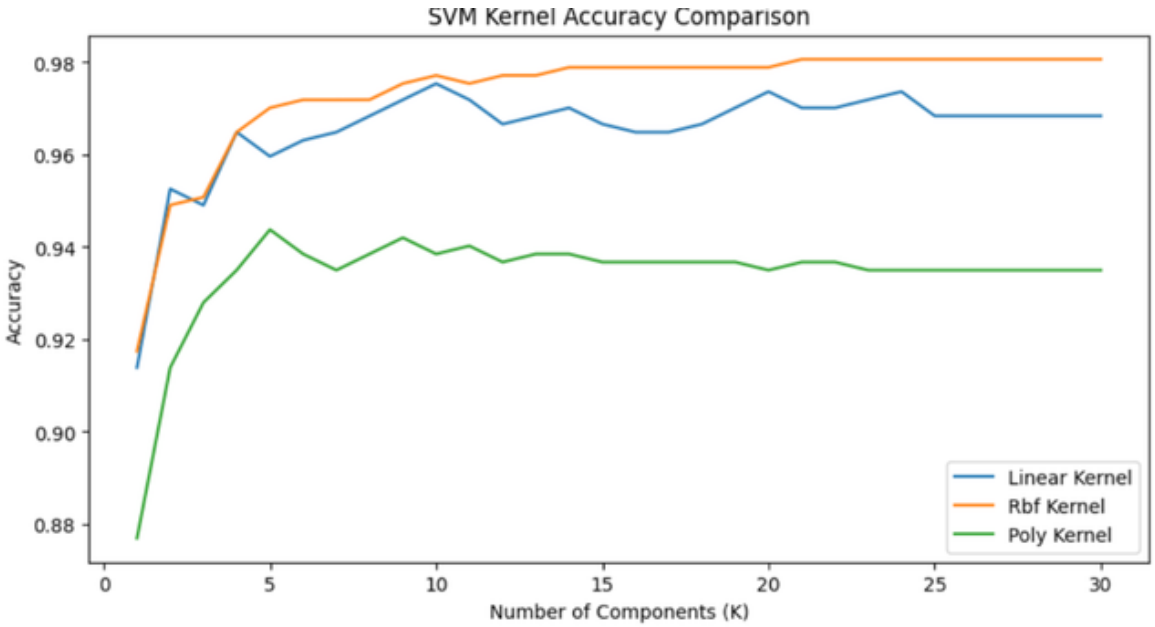

Figure 4: comparing all 3

By analyzing the comparison above we see we can actually achieve a higher accuracy by using a polynomial kernel and less PCA components (5)
Compared to the results from homework 3.. Logistic Regression (HW3) - Accuracy: 0.98, Precision: 0.99, Recall: 0.98 The accuracy from homework 3 was very good.. but we can match it.

Problem #2

Now developing an SVR regression model to predict housing prices given housing dataset, selectiong features "Area, bedrooms, bathrooms, stories, mainroad, guestroom, basement, hotwaterheating, airconditioning, parking, prefarea"

note: we must discretize categorical data before calculation.

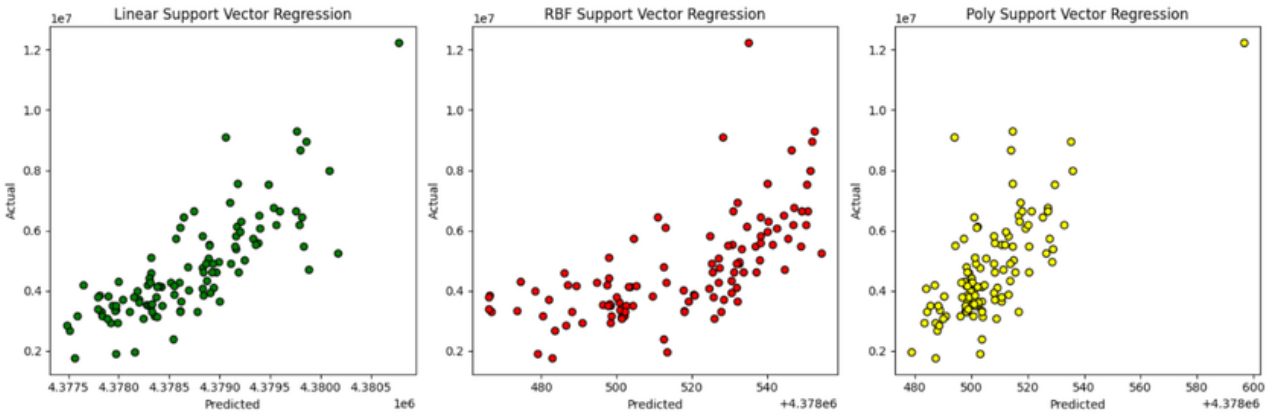

Figure 5: scatter

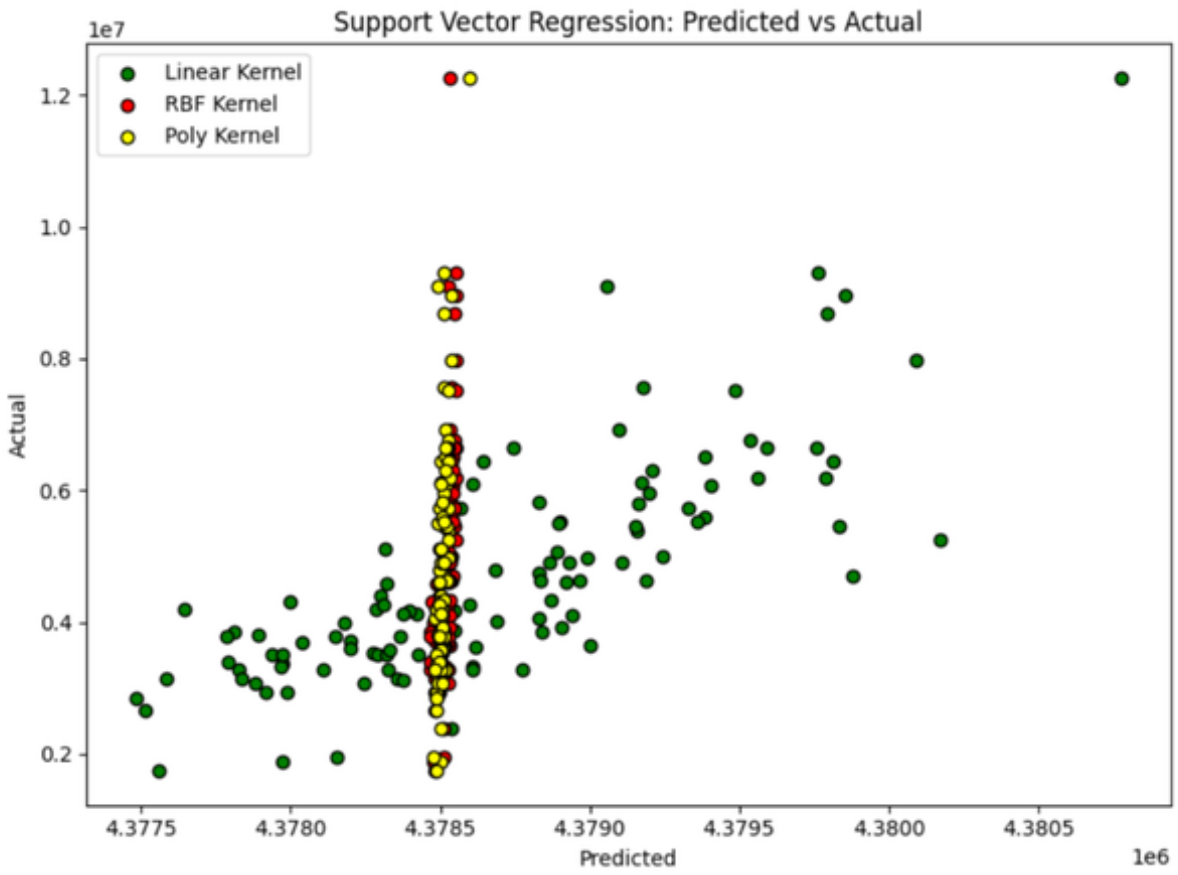

Figure 6: comparing error

plotting the SVR...

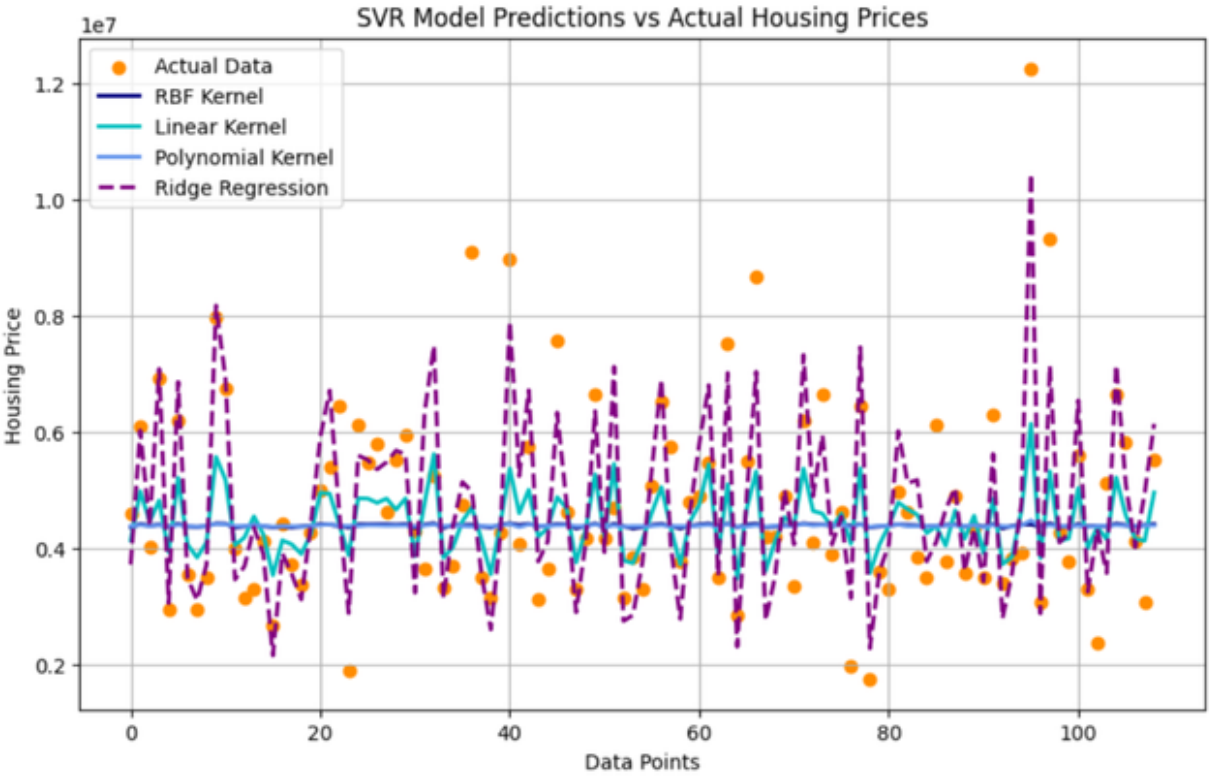

Figure 7: plotting SVR

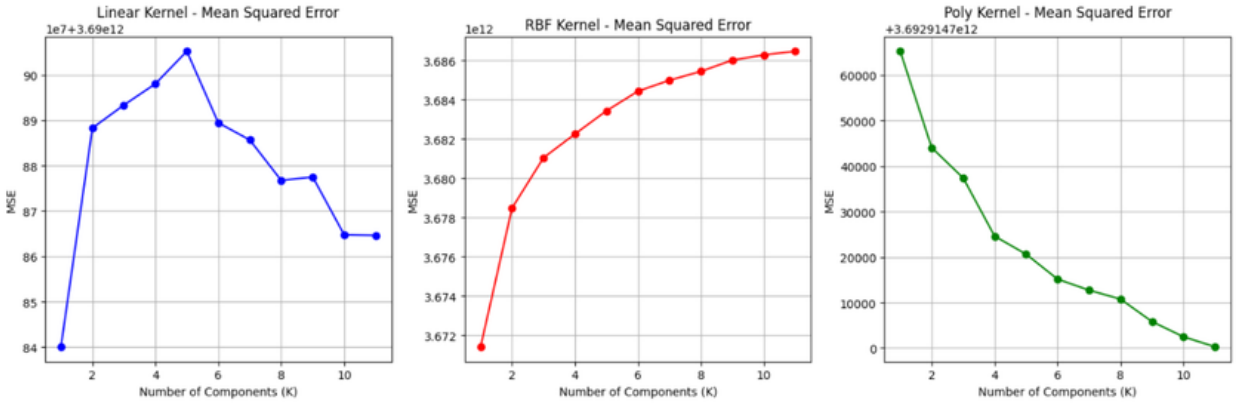

Figure 8: comparing error

From the data, it seems that the optimal number of PCA components is dependant on the type of kernel used for SVR.

Linear : 5
RBF : 10+
Polynomial : 1