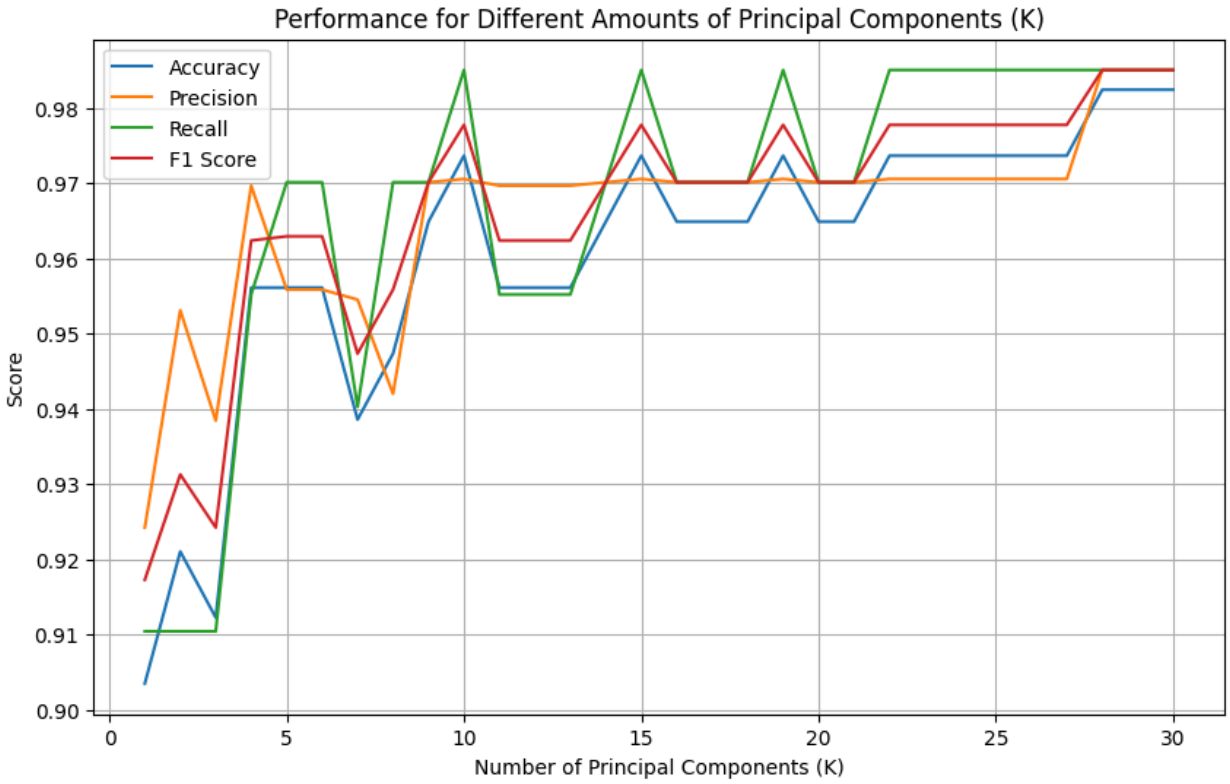
# Intro to ML

## Homework 4

Sophia Godfrey
801149485
Github: Intro-To-ML/Homework 4 at main · QueenSophiaLo/Intro-To-ML

# Problem 1

For Problem 1, a Support Vector Machine (SVM) classifier was created to distinguish between malignant and benign tumors using the Breast Cancer dataset. Similarly to Homework 3, the dataset was split 80/20 for training and testing to ensure fair evaluation. Different kernel functions (linear, polynomial, RBF, and sigmoid) were tested to capture potential non-linear patterns. Features were standardized using StandardScaler to balance their influence, as SVMs are sensitive to scale. Principal Component Analysis (PCA) was then applied to reduce dimensionality and assess how varying the number of components (K) impacts model accuracy and efficiency.




The initial experiment used a linear kernel to evaluate baseline performance across different PCA dimensions. For each PCA configuration, four main metrics were recorded:

- **Accuracy:** The proportion of correctly classified samples.
- **Precision:** The proportion of true positive predictions among all positive predictions.
- **Recall (Sensitivity):** The proportion of true positive predictions among all actual positives.
- **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.

These metrics were plotted against the number of principal components (K = 1–30) to visualize performance trends. As seen above, there is not much variability in score among the performance metrics for each of the associated k values.

| | Kernel | Best K | Accuracy for Best K |
|---|---|---|---|
| 0 | linear | 28 | 0.982456 |
| 1 | poly | 11 | 0.912281 |
| 2 | rbf | 9 | 0.982456 |
| 3 | sigmoid | 10 | 0.964912 |

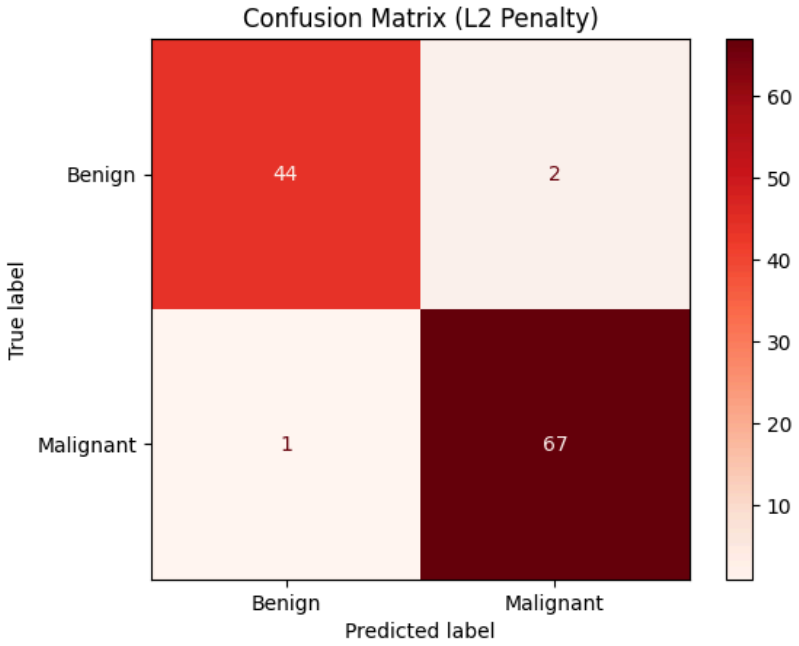
The linear SVM achieved strong results even with a small number of principal components. As the number of components increased, the model's performance stabilized, indicating that most of the predictive information is captured in the first few principal components.

Both linear (K=28) and RBF (K=9) kernels achieved the highest accuracy of 0.9825, indicating strong performance even with fewer components in the RBF model.

This suggests that dimensionality reduction can be beneficial by simplifying the feature space without significantly compromising classification performance. However, adding more components beyond the optimal K provided diminishing returns, indicating redundancy among features.

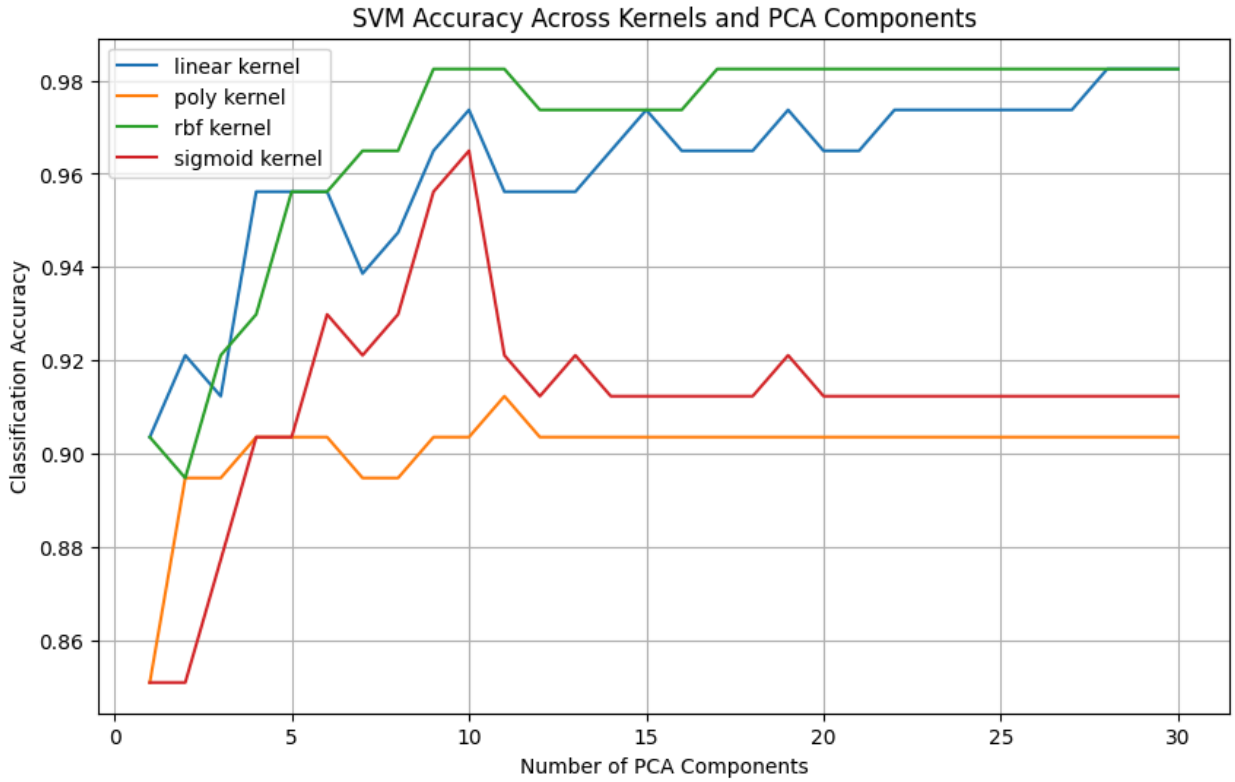
**HW3 Results:**

| Metric | Value |
|---|---|
| Accuracy | 0.9737 |
| Precision | 0.9710 |
| Recall | 0.9853 |
| F1 Score | 0.9781 |




In Homework 3, logistic regression achieved an accuracy of 0.9737, precision of 0.9710, recall of 0.9853, and an F1 score of 0.9781 on the Breast Cancer dataset. Both the unregularized and L2-regularized models performed nearly identically, indicating strong linear separability of the data.

**Problem 1 HW4 results:**




In contrast, the SVM classifier in Homework 4 achieved slightly higher accuracy (up to 0.9825) when using the linear and RBF kernels. The best-performing SVMs used K = 28 (linear) and K = 9 (RBF) principal components after PCA. These results suggest that SVM, particularly with the RBF kernel, captured subtle non-linear patterns beyond what logistic regression could model, though both methods performed very similarly overall.

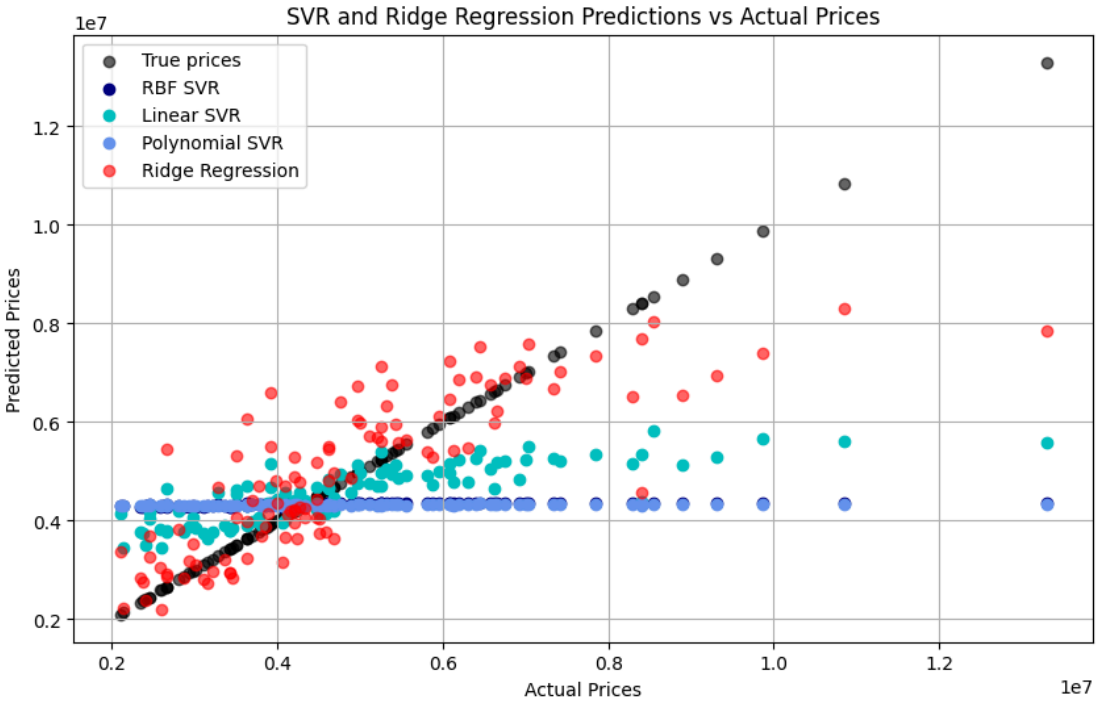
Overall, both logistic regression and SVM demonstrated good performance on this dataset. Logistic regression remains computationally simpler and interpretable, while SVM—especially with the RBF kernel—offers slightly improved generalization by handling complex feature interactions.

# Problem 2

For Problem 2, a Support Vector Regression (SVR) model was created to predict housing prices based on multiple features (area, bedrooms, bathrooms, stories, and various binary attributes). Then, the SVR results were compared with the Ridge (regularized linear) regression model from Homework 1. The same housing.csv dataset was loaded in for both HW1 and HW4, with binary categorical variables (e.g., mainroad, guestroom, basement) converted to 0/1 encoding, and standardized features converted with StandardScaler before training. The models trained included Support Vector Regression (SVR) using three different kernels—RBF, Linear, and Polynomial—as well as Ridge Regression, which served as a linear baseline with L2 regularization.

| | Model | MSE | R2 |
|---|---|---|---|
| 0 | RBF SVR | 4.066866e+12 | -0.060 |
| 1 | Linear SVR | 2.483255e+12 | 0.353 |
| 2 | Polynomial SVR | 4.115708e+12 | -0.072 |
| 3 | Ridge Regression | 1.303046e+12 | 0.661 |

Ridge Regression achieved the best performance ($R^2 = 0.661$), indicating it explains a substantial portion of the variance in housing prices. SVR models performed poorly, with RBF and Polynomial kernels giving negative $R^2$, meaning predictions were worse than simply predicting the mean. Linear SVR performed moderately better ($R^2 = 0.353$) but still underperformed compared to Ridge Regression. These results suggest that for this dataset, a linear model with regularization captures the relationship between features and price better than SVR with the chosen hyperparameters.



Predictions from SVR and Ridge Regression were plotted against the actual prices. Ridge Regression predictions aligned more closely with the true values, whereas SVR predictions showed large deviations, especially for high-priced houses. Ridge Regression is the preferred model for predicting housing prices in this dataset. SVR, despite exploring different kernel functions, did not outperform a regularized linear model. Fine-tuning SVR hyperparameters or feature engineering might improve SVR performance, but for now, Ridge Regression provides the most reliable predictions.