Spam Email Classification Report

**KNN Binary Classification**

KNN Binary predicts the target value for a new instance by averaging the values of its K nearest neighbors in the training set.

About implementing the KNN binary classifier model, I set the k-values from 1 to 20 to compare. After that, I go through all k-values of the KNN model and check for the 5-Fold Cross-Validation accuracy of each value, then store them in knn_result.  Then I have this result:

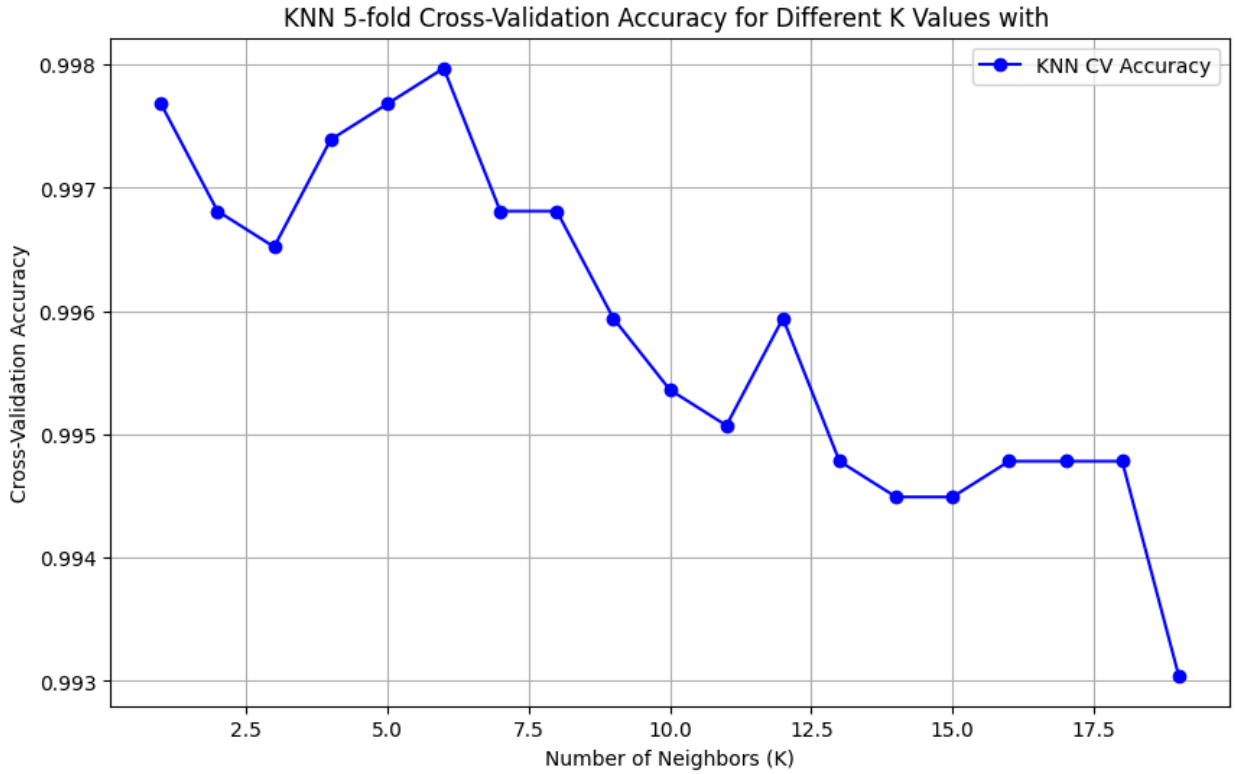



Figure 1

KNN 5-Fold Cross-Validation:

- The graph shows that all K values have a cross-validation (CV) accuracy above 0.99 prove that the dataset is straightforward for KNN to classify.
- My best cross validation accuracy is 0.9980 when K = 6.
- With small K (K = 1), which usually risks overfitting, it performs well in KNN (0.9977 accuracy score).
- With optimal K (K = 4 to 8), all of them perform pretty well around 0.9974 to 0.9980.
- With larger K (K >= 9), it decreases in accuracy, indicating potential underfitting due to excessive smoothing.

=> Best choice is K = 6 with a cross validation accuracy of 0.9980.

**Logistic regression**

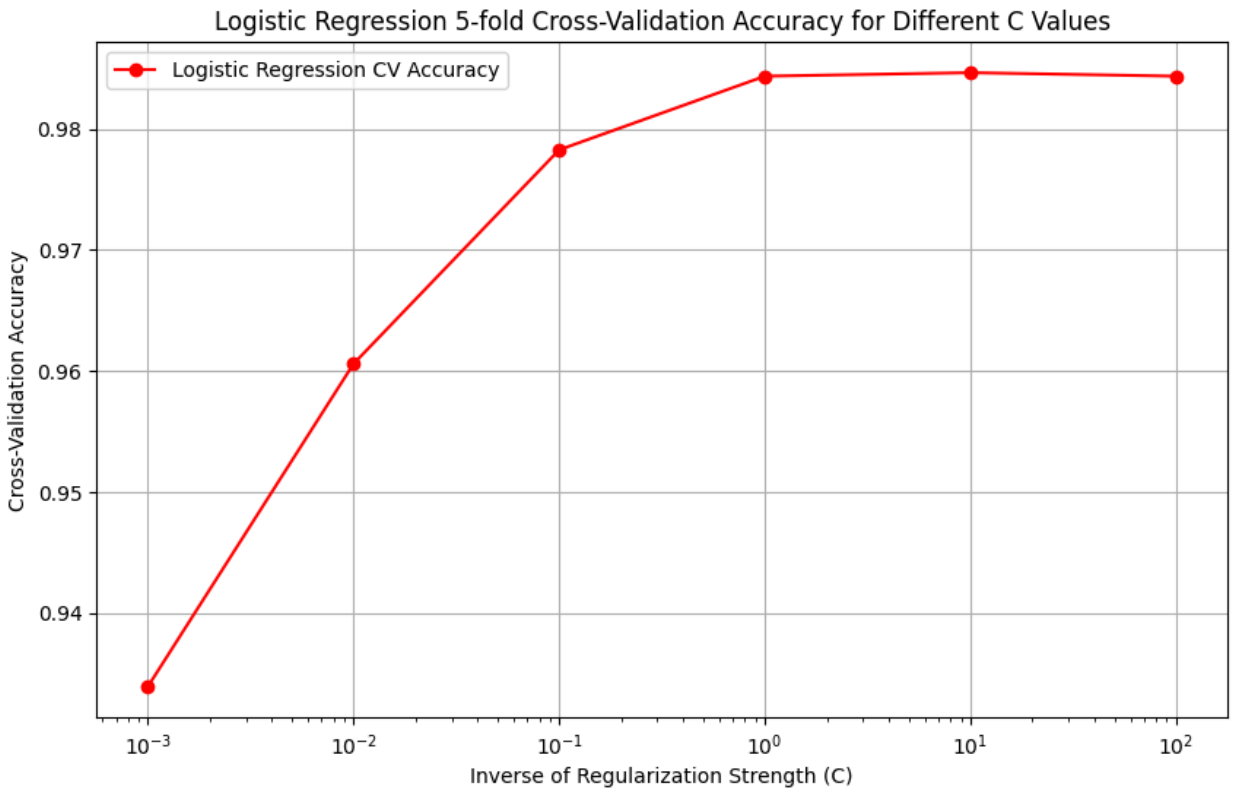
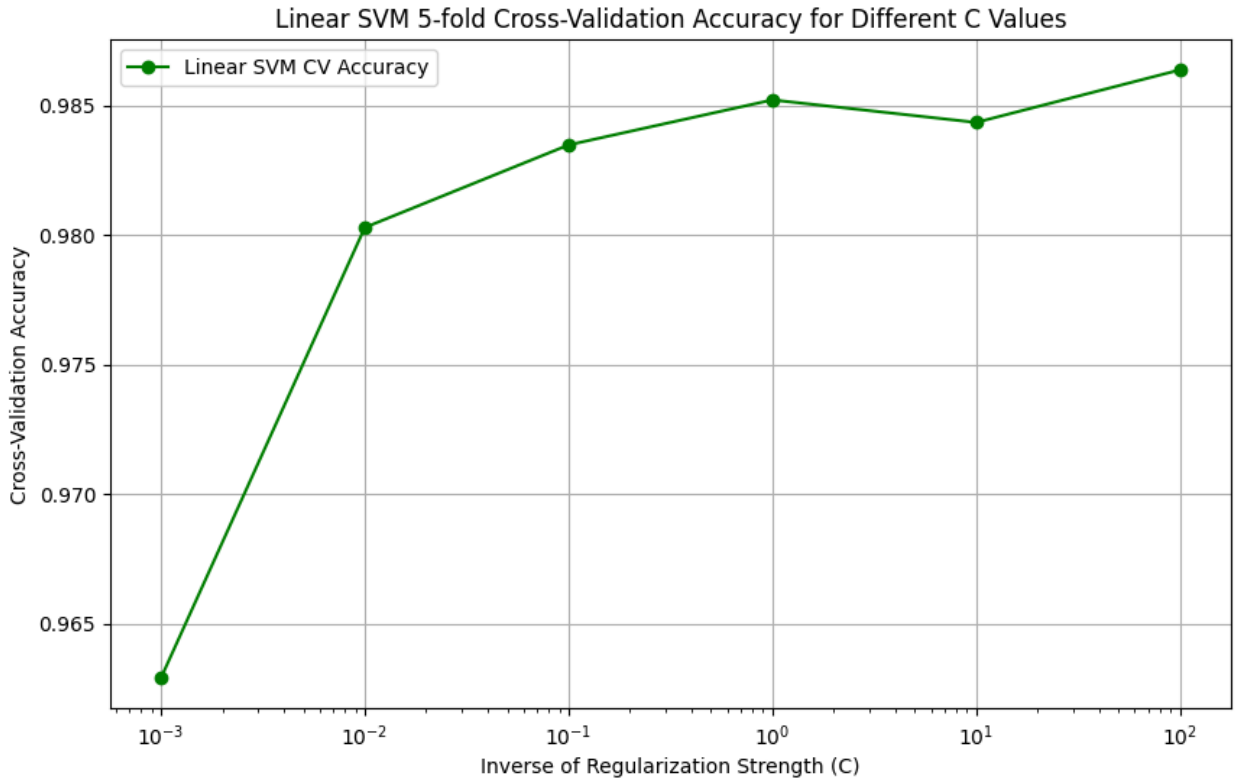



Figure 2

Logistic Regression varying C with 5-Fold Cross-Validation:

- At very low (C = 1), the model is heavily regularized, yielding a relatively low CV accuracy (93.39%), which suggests **underfitting** due to high bias.
- When C is range from 0.001 to 10, the CV accuracy steadily improves, reaching 0.9846 at C=10. We can see that reducing regularization allows the model to better capture the underlying patterns in the data.
- At high C (C = 100), the accuracy slightly drops to 0.9843, which shows that reducing regularization doesn't improve performance and start to risk overfitting, though the change is minimal.

**Linear Support Vector Machine**




Linear Support Vector varying in C with 5-Fold Cross Validation:

- At very low (C = 0.001), the model is overly constrained (high regularization), leading to a low CV accuracy 0.9629. This indicates underfitting.
- At C range from 0.01 to 10, the model rises higher with only drop at 1 to 10, which shows that the model reduced regularization helping it capture underlying data structure better.
- At C = 100, it performs the most accuracy at 0.9864, which indicates that the model benefits from minimal regularization, which allows a more flexible decision boundary.

=> The best choice for this Linear SVM is C = 100.

**Comparing testing accuracy**

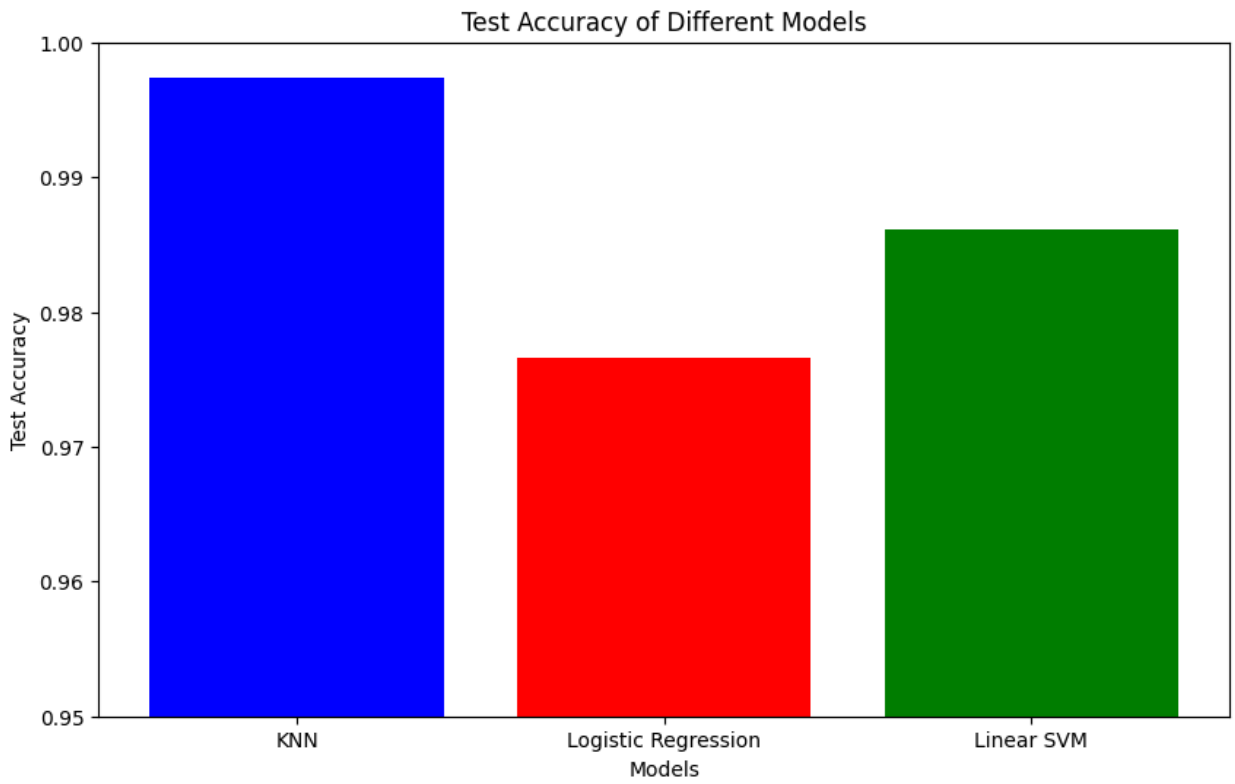



Figure 4

- As you can see, KNN achieves the best accuracy on this test set. Linear SVM ranks second, followed by Logistic Regression. In my opinion, KNN with k = 6 is the best choice with the highest accuracy on tests with the highest accuracy on 5-fold Validation.