SENG 474 Assignment 2

Zheng Yin

March 1, 2021

1. Logistic Regression Analysis

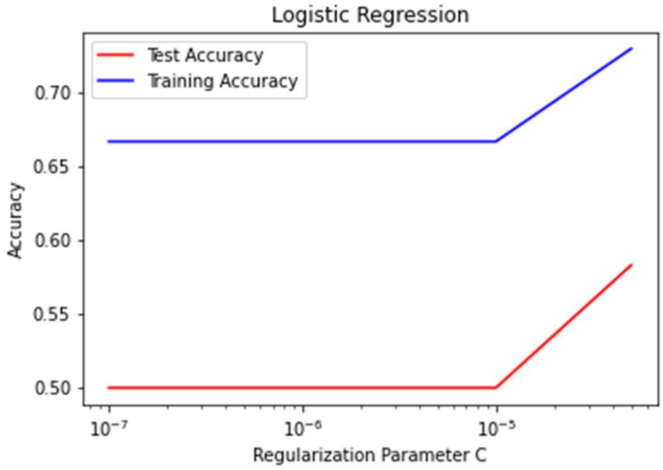
In assignment 2 of SENG 474, the first method we use on the fashion-mnist dataset is Logistic Regression. Logistic regression is a statistical model by using the logistic function to model some binary model, which the sigmoid function to map the prediction. There should be a set of probability from the prediction function which are divided to 0 or 1 as a map.

L2 regularization is used in this Logistic Regression method which is a function that helps to prevent overfitting. L2 is leading to function below which λ *is called the regularization parameter that controls two things: fit well on data and avoid overfitting by keeping the params small.*

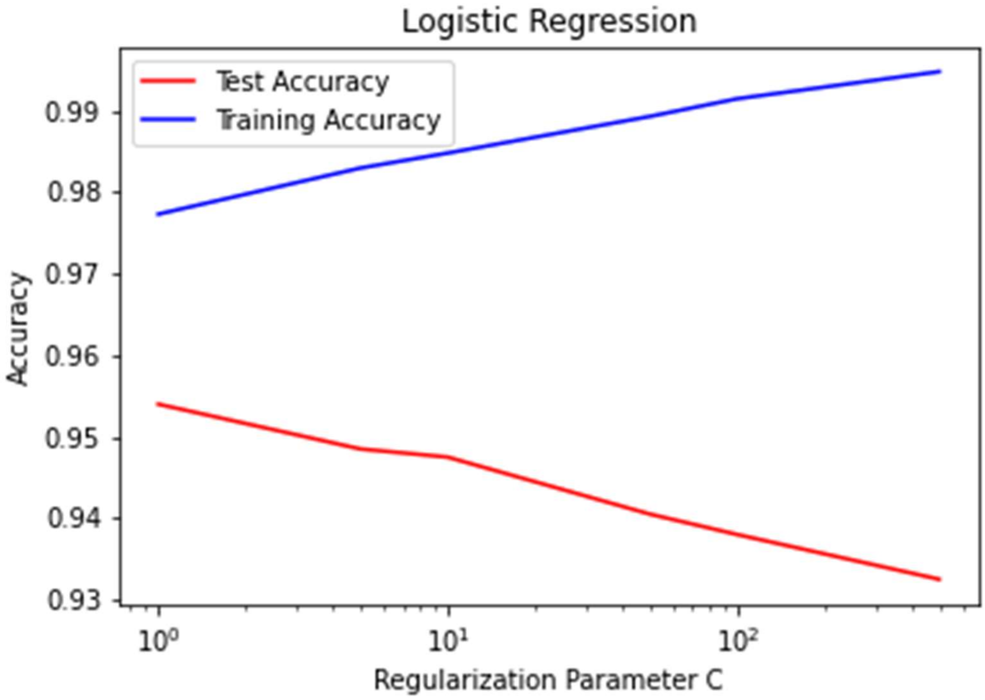
$$J(w) = \frac{1}{m}\sum_{i=1}^{m} Cost(\,h(x^{(i)}),\, y^{(i)})\, + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

("implement logistic regression with L2 regularization from scratch in Python", Tulrose Deori, Jul 26, 2020, https://towardsdatascience.com/implement-logistic-regression-with-l2-regularization-from-scratch-in-python-20bd4ee88a59)

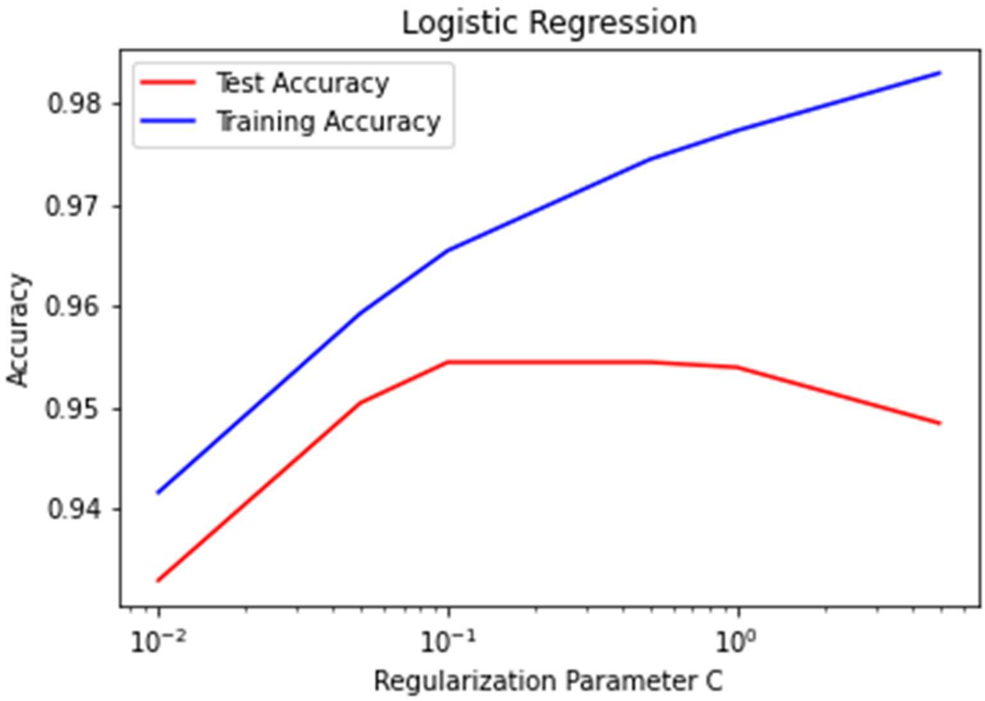
With L2 regularization, the system can find the smaller weight to make the prediction more accurate, which is the C value we input. The size of the dataset is another problem in which a large amount of data require a lot of time on our device, so we reduce the training example number to 6000 from the "Sandal" (4000) and "Sneaker" (2000) class. At the same time, there is 2000 test example from this dataset to evaluate the prediction. We test several C numbers which got different results on the chart. The first chart I got is with C around 10^-6 as shown below:



The second C is around 0 to 10^3 which is overfitting which training accuracy is increasing from about 98% to above 99%, but the test accuracy is decreasing from 95% to about 93% which means this C is likely overfitting (chart below)

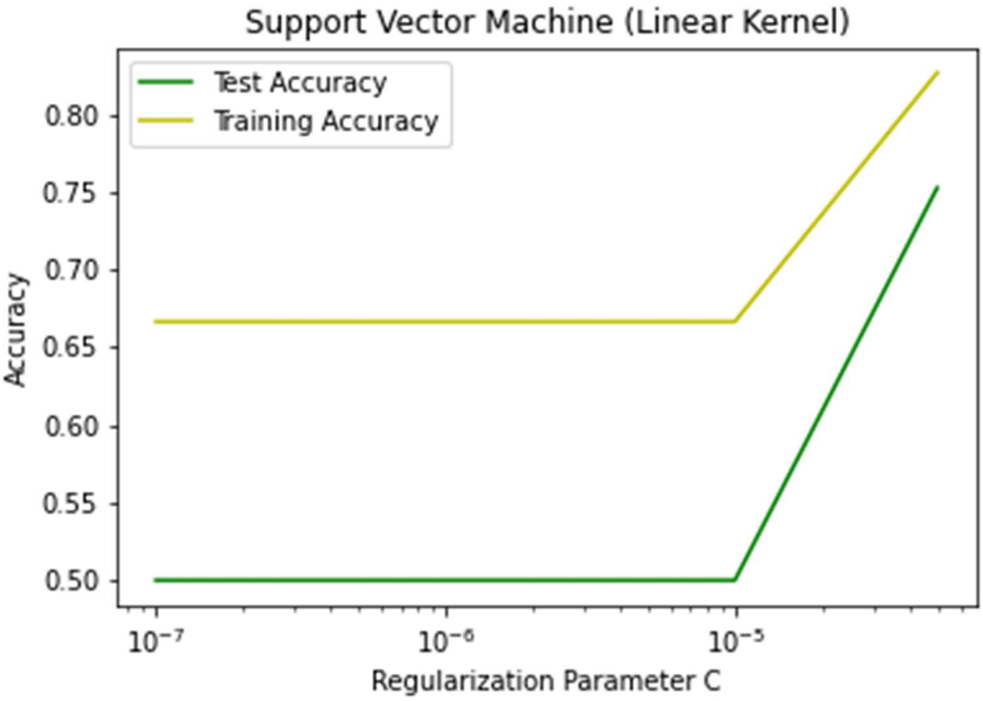


When C reaches about 10^-1, the training accuracy and the test accuracy are most close, after that, the test accuracy begins to drop. The chart below is a midrange of tested C, it shows the highest point of test accuracy on about 0.1, because of this, we can say that this chart provides the most accurate result for Logistic Regression.

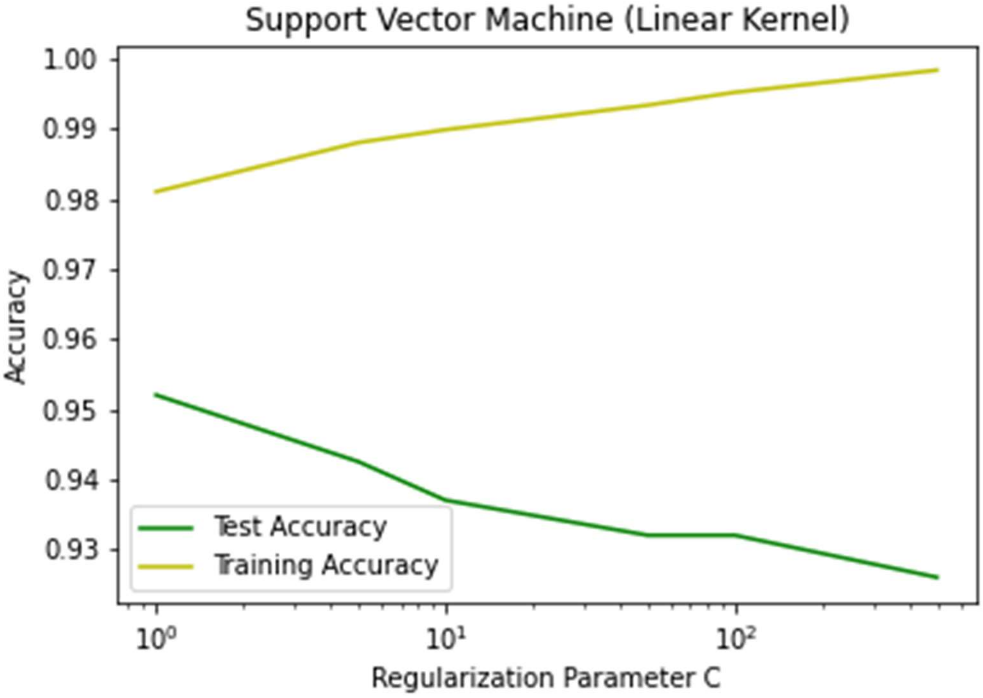


2. SVM

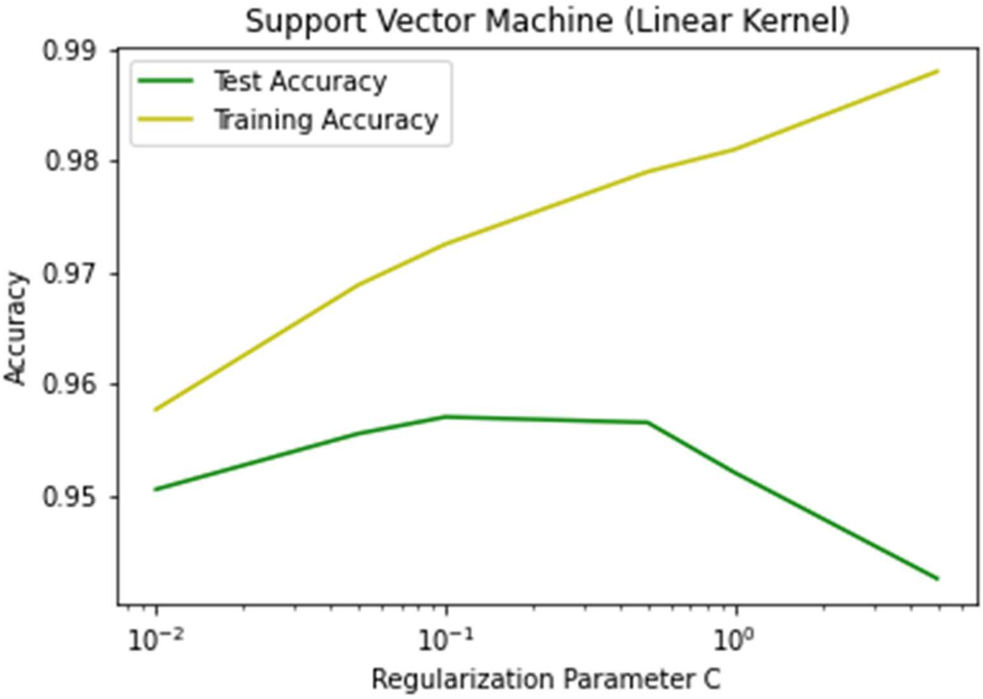
Support Vector Machine is the second method on this dataset. SVM can work well on high dimensional space which is helpful on the fashion-mnist dataset. In the first test, SVM is designed to run at linear kernel which SVM have the same C set as logistic regression which first is from 10^-7 to 10^-5:



As the chart shows, both test accuracy and training accuracy are at a low level, and both begin to climb at around 10^-5, which means C with about 10^-7 is underfitting like the logistic regression. We can consider 10^-5 is the key point that accuracy begins at 10^-5



In the range of 1 to 100, the result of SVM is the same with logistic regression, the chart above shows the sign of overfitting which test accuracy is decreasing and the training accuracy is increasing.



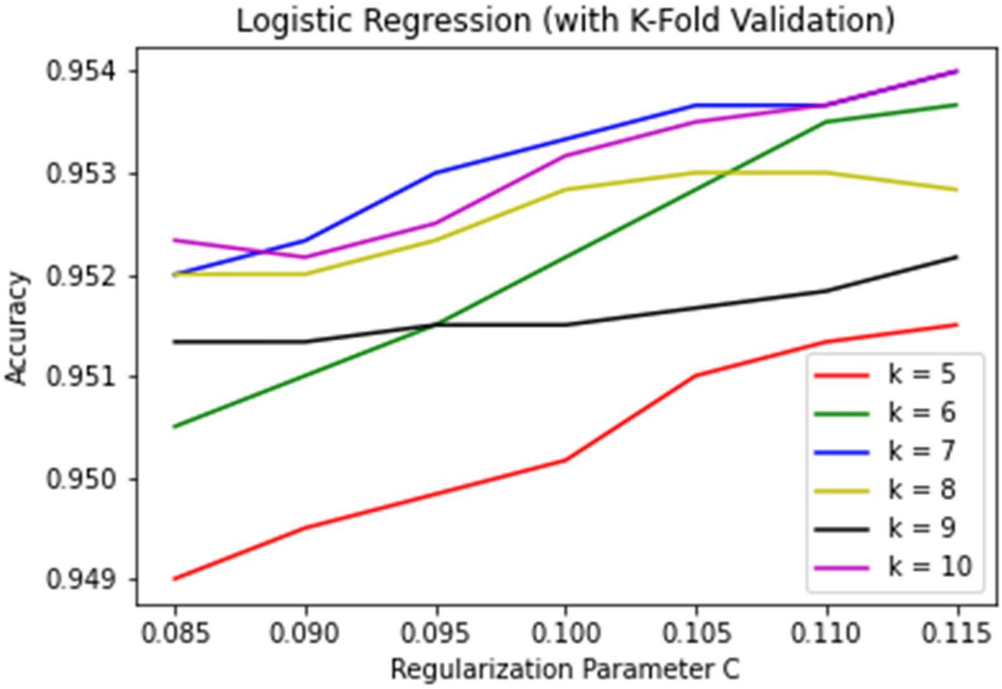
The chart above is the C value which is midrange in logistic regression, and it shows the same thing as the logistic regression which test accuracy begins to decrease at 0.1, but SVM has better test accuracy in C = 0.01 which logistic regression is below 94%, but SVM is 95%.

In a conclusion, both SVM and logistic regression show a C value which suitable for this dataset is about 0.01. More training data will increase the accuracy for tests and training, but I think there is a limit to which more training data will not change the accuracy much. In the same time, the time usage of the larger dataset is a problem that it takes SVM 5 minute to finish training on 6000 example.

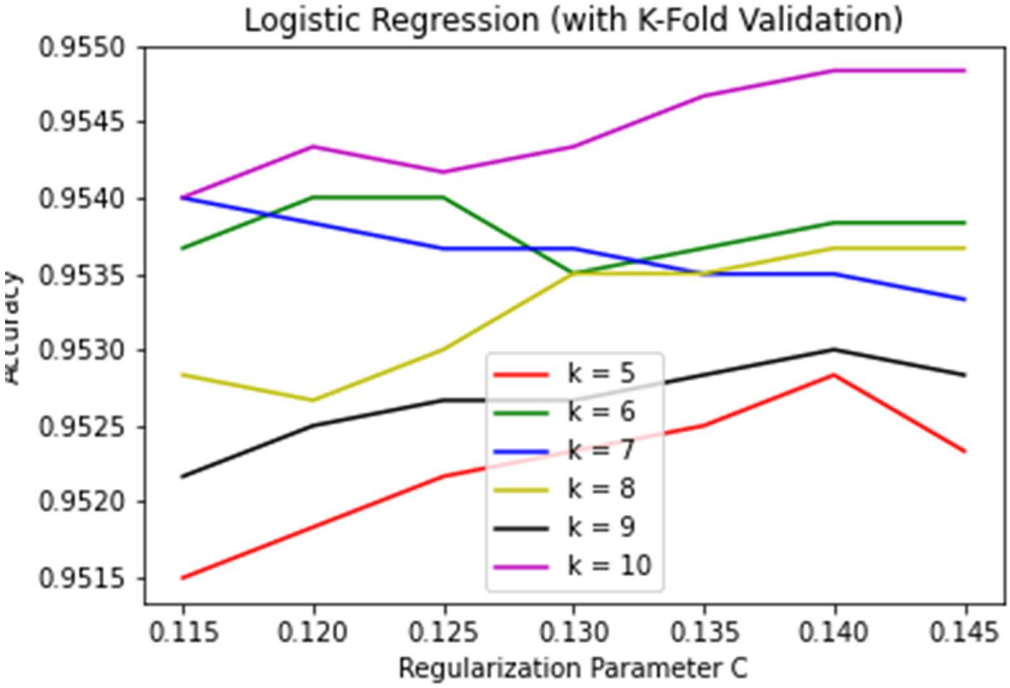
3. K-Fold Cross-Validation

As we learn from SENG 474 course, on the way of compare two methods which are SVM and logistic regression, K-fold is a choice that we need to know the C value is for this dataset. In K-Fold Cross-Validation, the training set, and test set is separate to k sets, and the result is evaluated by the average accuracy of the test set.

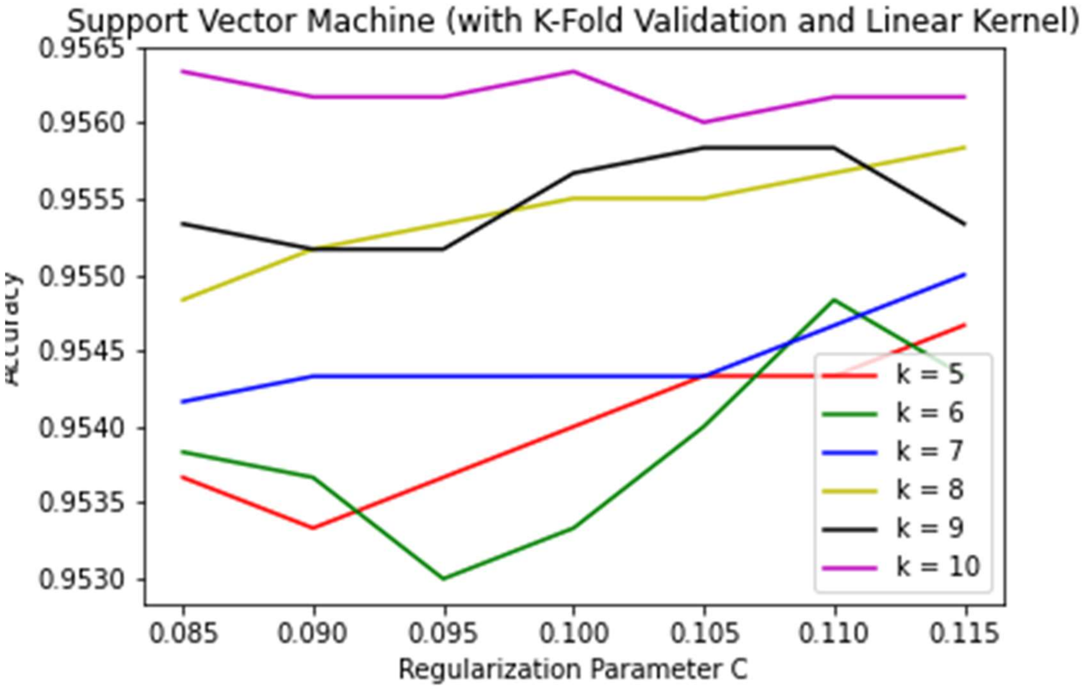
We got the C value which is about 0.1 in the last two sections and k value will be 5,6,7,8,9 and 10 which should not increase the training time deeply.



In the logistic regression, the greatest accuracy is about 0.115 which both k=7 and k=10 is on the top that reached 95.4%, the lowest accuracy is at C = 0.085 and k = 5 which the accuracy is about 94.9%. In general, separate to 7 groups or 10 groups make the training accuracy on the top of the list. All lines in chart is increasing may declare that the right C number is higher than 0.115 which tester should take another test run to measure the true C value for Logistic Regression.



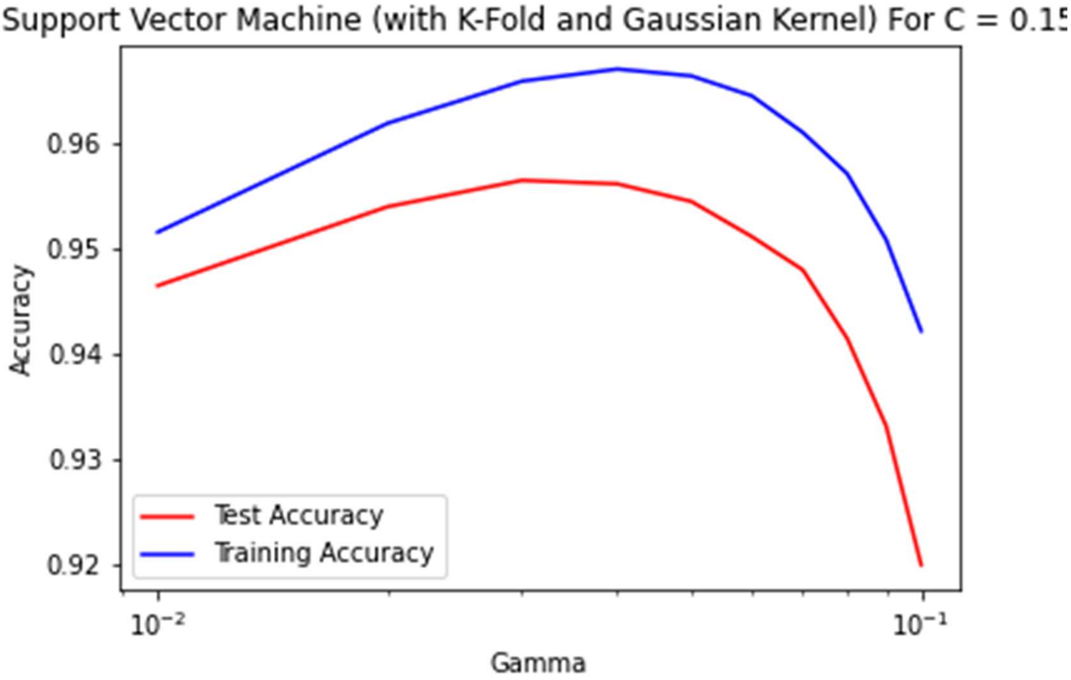
After the last test, I increase the C value to 0.13 which have a better chart which k=10 is un the top of the chart and line of K=5 on the bottom of the chart. Most line is smoother, and the highest point is about 0.955.



The chart above is the K-Fold validation for SVM, the information is straightforward. The highest accuracy is K =10 which both C=0.085 and C= 0.1 is the highest point about 0.9565 in the chart. The lowest accuracy is below the line for K=6 which reaches 0.9530 on C = 0.095. The overall graph shows a stable situation that accuracy got higher with more group is generate to training and testing.

4. Gaussian Support Vector Machine

After comparing with two methods, we will use an SVM with K-Fold and Gaussian on this dataset, K value will be 10, and the C value will be 0.15 which shows the accuracy with gamma. The test accuracy is about 0.96 in this situation, but there is a large gap between test accuracy and training accuracy.



In conclusion, SVM has higher accuracy in data training, especially SVM with K-Fold and Gaussian which have a 1% of advantage, but it takes a much longer time to train which will take 8 times than the Logistic Regression.