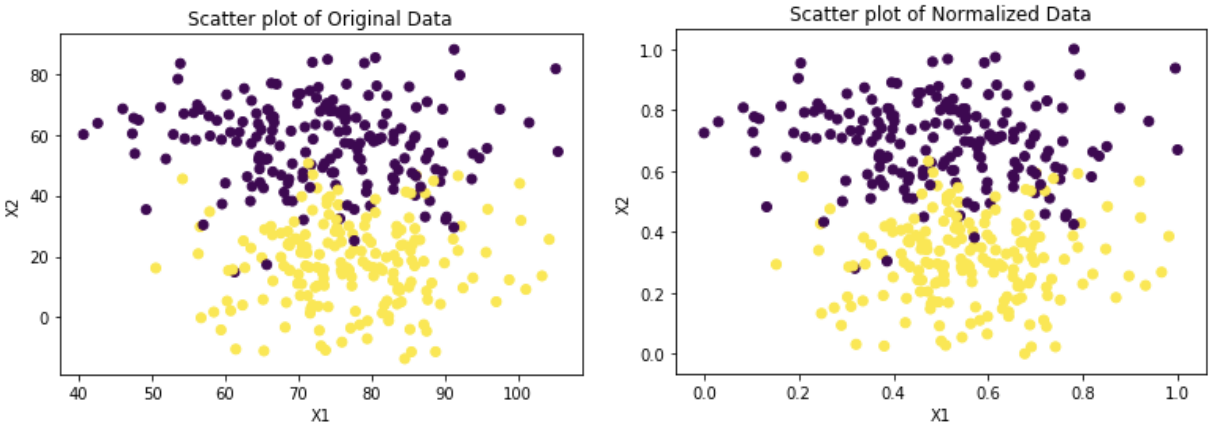
CSC 591 IOT Analytics
Project 5 SVM
Name: Vidhisha Jaswani
Unity ID: vjaswan

Task 1. Scatter Diagram



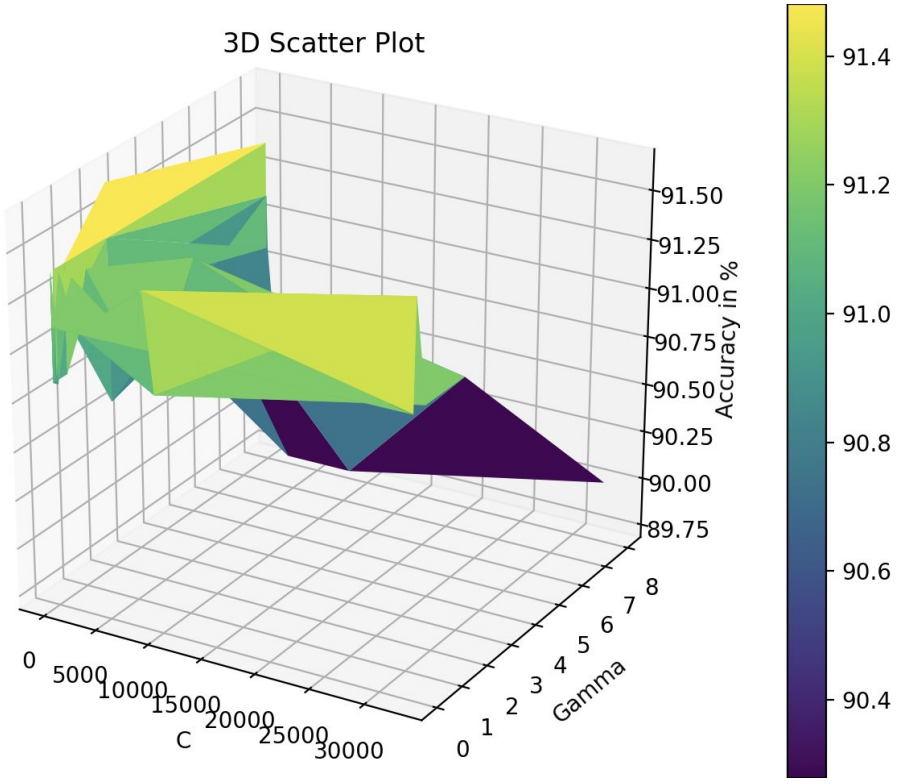
- Above is a scatter plot for the original data with X1 on X-axis and X2 on Y-axis.
- Different classes have different color and since we only have 2 classes, 2 colors are depicted above.
- By the original observation, it seems that data is not well separated and it is possible that SVM gives a lower accuracy for classification of this data.
- Data has then been normalised to the [0,1] scale. This is important in SVM to avoid having attributes with higher ranges dominate those with lower ranges. This also avoids computational complexity.
- Scatter plot for the normalized data looks no different than the original plot.

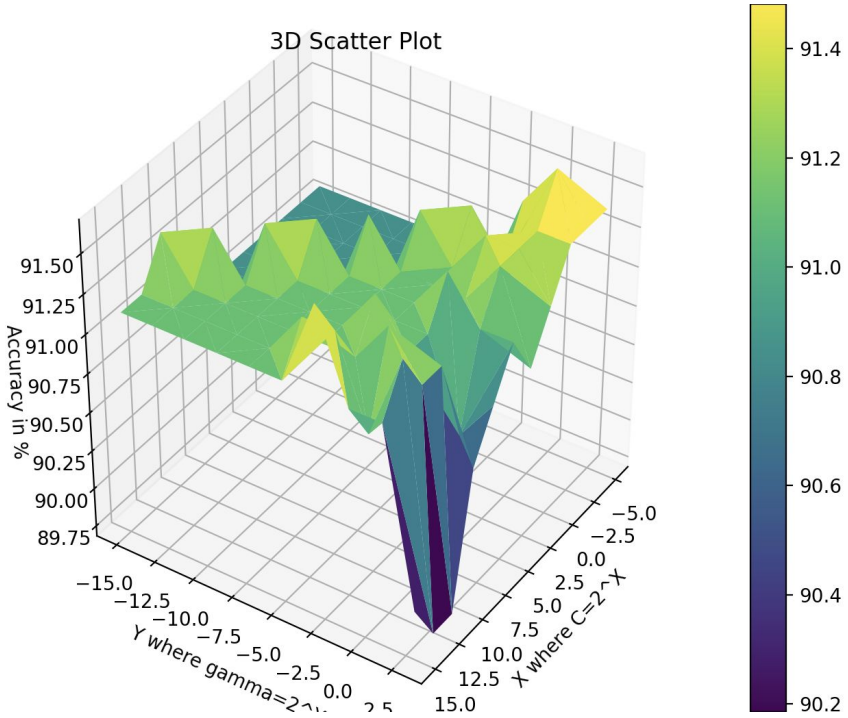
Task 2. Best values of C and γ and show how you obtained them. Specifically, a) discuss the range of values you used, b) give the 3D plot of percent accuracy vs C and γ, and c) identify the best value(s).

- First, I ran the SVM for
  - Kernel: RBF
  - C : $2^{-5}, 2^{-3}, 2^{-1}, \ldots..2^{13}, 2^{15}$ i.e in steps of 2
  - Gamma: $2^{-15}, 2^{-13}, 2^{-11}, \ldots.., 2^{3}, 2^{5}$ i.e in steps of 2
  - 5-folds cross validation

- Grid Search was done that does an exhaustive search over the values of C and Gamma. It also does a k-fold cross validation for each possible value of C and gamma when we specify the cv parameter in Python.

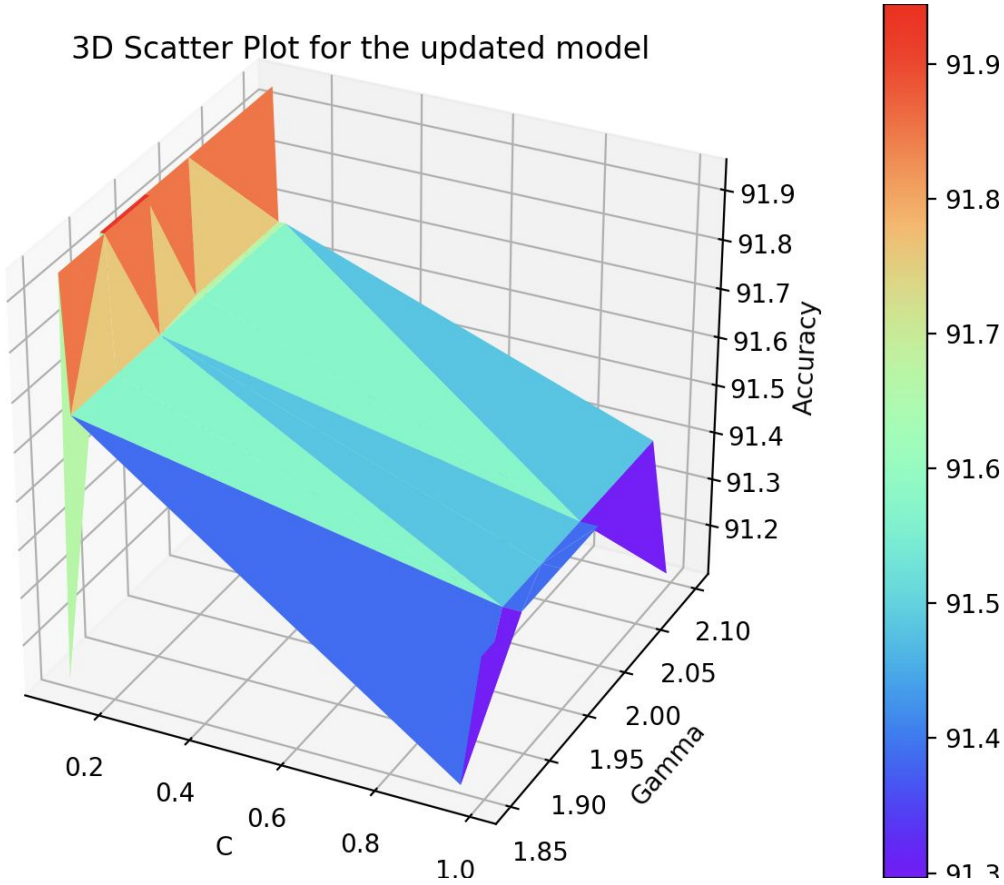
- Results: The best accuracy which is 91.66% was obtained for parameters {'C': 0.125, 'gamma': 2, 'kernel': 'rbf'}. Below is 3-D scatter plot for the above ranges of C, Gamma, and obtained accuracies.





- Above is the same plot with C=2^X and Gamma=2^Y.
- It can be seen from the 3-D Scatter plot that the highest accuracy is obtained the yellow region which around gamma~2-4 and low values of C which also aligns the best parameters obtained. The color of the surface shows the accuracy percentage.
- Second, I ran SVM for the below parameters:
    - C : 0.03125,0.125,0.5,2
    - Gamma: 0.125,0.5,2,4,8
- The ranges are just a finer version of the ranges I took in the first run. The best parameters were obtained for {'C': 0.125, 'gamma': 2, 'kernel': 'rbf'} with accuracy 91.66% which is same as the accuracy obtained earlier. Therefore, we should refine our parameters.
- Next, I took the updated parameters to be:
    - C : 0.95,0.99,0.10,0.110,0.125
    - Gamma: 0.125,0.5,2,4,8

  i.e increasing in steps of 0.1 or more. The best parameters now obtained were different with better accuracies which are {'C': 0.1, 'gamma': 1.9, 'kernel': 'rbf'} and accuracy 91.94% which is a little improvement than before. I also varied C and Gamma and the best parameters were finally at C=0.1 and Gamma=1.9. The 3-D plot for this range is as below:



The color of the surface shows the accuracy. There is top red line where the accuracy of 91.94% is obtained.

- The reason why we did not obtain higher accuracy could be because of what we discussed above that the two classes merge as visible in the scatter plot.
- Conclusion: Best value of C=0.1, Gamma=1.9, and accuracy=91.94%.