ML LAB – 7

NAME: Mrinal Pandey
SRN: PES2UG23CS353

MOONS DATASET QUESTIONS:

1. The Linear kernel performed poorly on the moons dataset because it can only create a straight decision boundary. The moons dataset has two interleaving crescent shapes that are not linearly separable, so a linear decision boundary can't correctly split the classes, leading to lower accuracy and misclassifications.

2. The RBF kernel produced smooth, flexible decision boundaries that closely fit the curved shape of the moons dataset, effectively separating the classes. The Polynomial kernel also captured non-linearity but generated more complex, sometimes irregular boundaries depending on its degree. This could cause overfitting or underfitting relative to the RBF kernel, which by using a radial basis function, adapts more naturally to the shape of the data.

BANKNOTE DATASET QUESTIONS:

1. The linear kernel often performs best for the banknote dataset, as the data is almost linearly separable when using features such as variance and skewness, making complex kernels unnecessary.

2. The polynomial kernel could underperform due to overfitting on noise or unnecessary complexity since the dataset is mostly linearly separable, and polynomial kernels can introduce too many degrees of freedom.

HARD vs. SOFT MARGIN QUESTIONS:

1. The soft margin is wider because it allows some misclassifications to increase the margin size.

2. To handle noisy or overlapping data better and to avoid overfitting, which improves generalization to unseen data.

3. The hard margin model is more likely to overfit because it tries to perfectly separate the training data without tolerance for misclassifications, capturing noise as if it were a pattern.

4. The soft margin model is generally more reliable for new data since it balances margin maximization with some tolerance to errors, helping it generalize better beyond the training set.

TRAINING RESULTS:

MOONS DATASET:

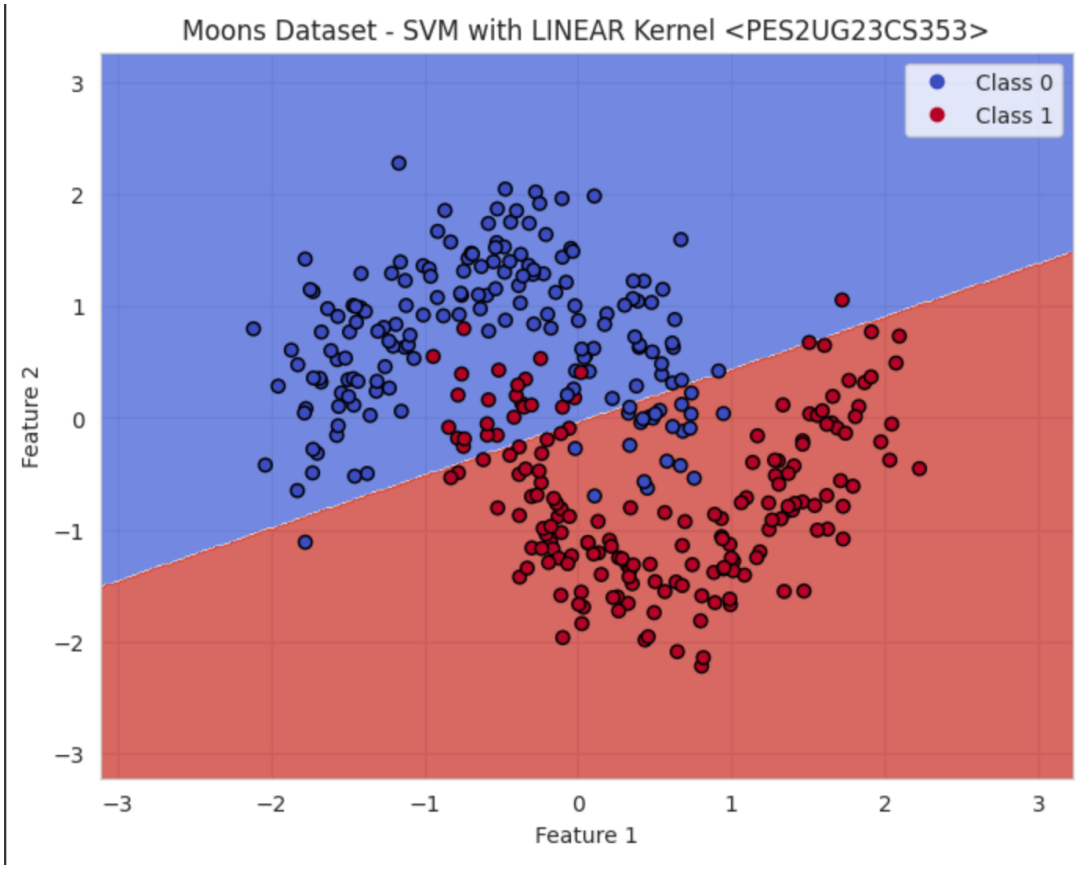
```
SVM with LINEAR Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

           0       0.85      0.89      0.87        75
           1       0.89      0.84      0.86        75

    accuracy                           0.87       150
   macro avg       0.87      0.87      0.87       150
weighted avg       0.87      0.87      0.87       150

----------------------------------------

SVM with RBF Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

           0       0.95      1.00      0.97        75
           1       1.00      0.95      0.97        75

    accuracy                           0.97       150
   macro avg       0.97      0.97      0.97       150
weighted avg       0.97      0.97      0.97       150

----------------------------------------

SVM with POLY Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

           0       0.85      0.95      0.89        75
           1       0.94      0.83      0.88        75

    accuracy                           0.89       150
   macro avg       0.89      0.89      0.89       150
weighted avg       0.89      0.89      0.89       150

----------------------------------------
```

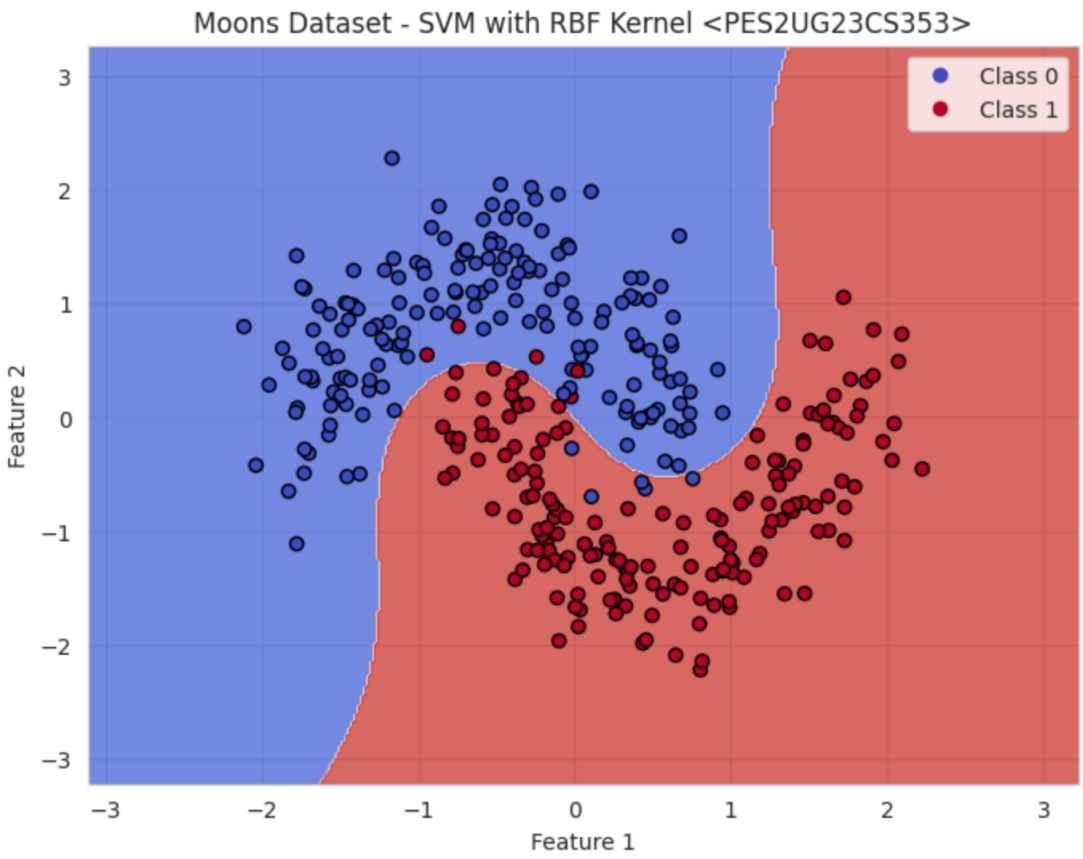
BANKNOTE DATASET:

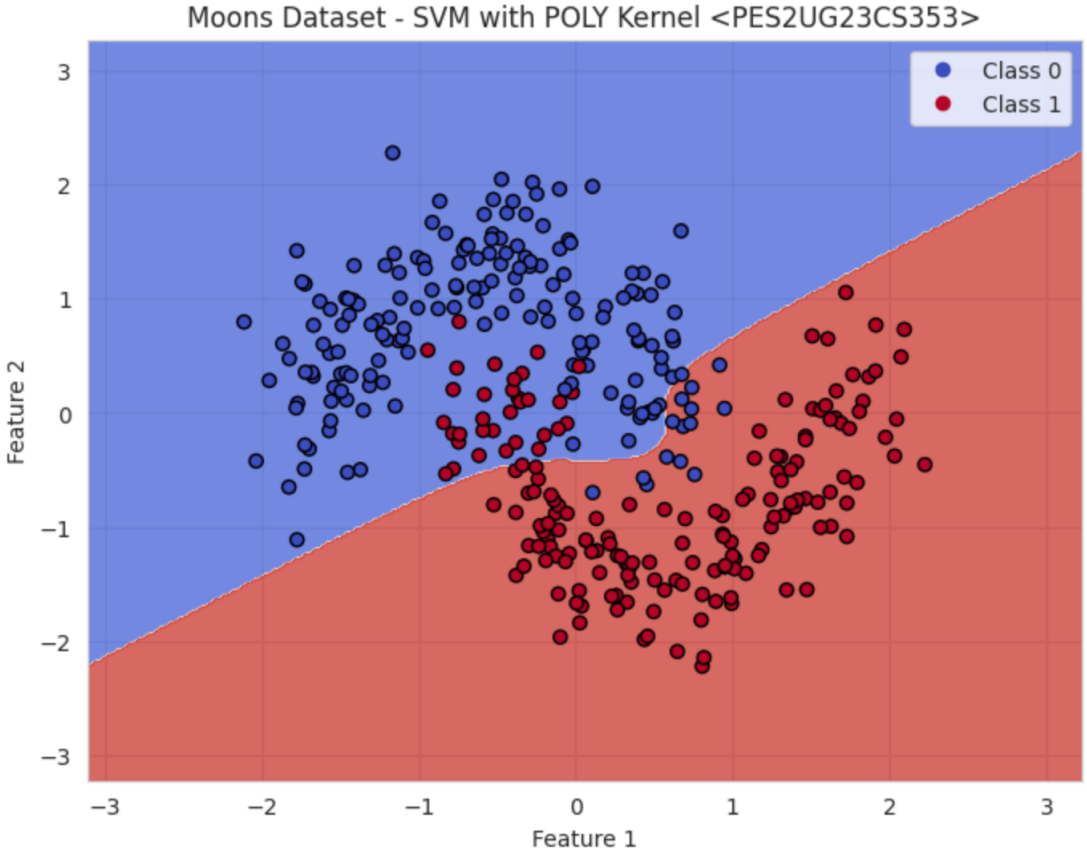
```
SVM with LINEAR Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

      Forged       0.90      0.88      0.89       229
     Genuine       0.86      0.88      0.87       183

    accuracy                           0.88       412
   macro avg       0.88      0.88      0.88       412
weighted avg       0.88      0.88      0.88       412

----------------------------------------

SVM with RBF Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

      Forged       0.96      0.91      0.94       229
     Genuine       0.90      0.96      0.93       183

    accuracy                           0.93       412
   macro avg       0.93      0.93      0.93       412
weighted avg       0.93      0.93      0.93       412

----------------------------------------

SVM with POLY Kernel <PES2UG23CS353>
              precision    recall  f1-score   support

      Forged       0.82      0.91      0.87       229
     Genuine       0.87      0.75      0.81       183

    accuracy                           0.84       412
   macro avg       0.85      0.83      0.84       412
weighted avg       0.85      0.84      0.84       412

----------------------------------------
```

DECISION BOUNDARY VISUALIZATIONS:

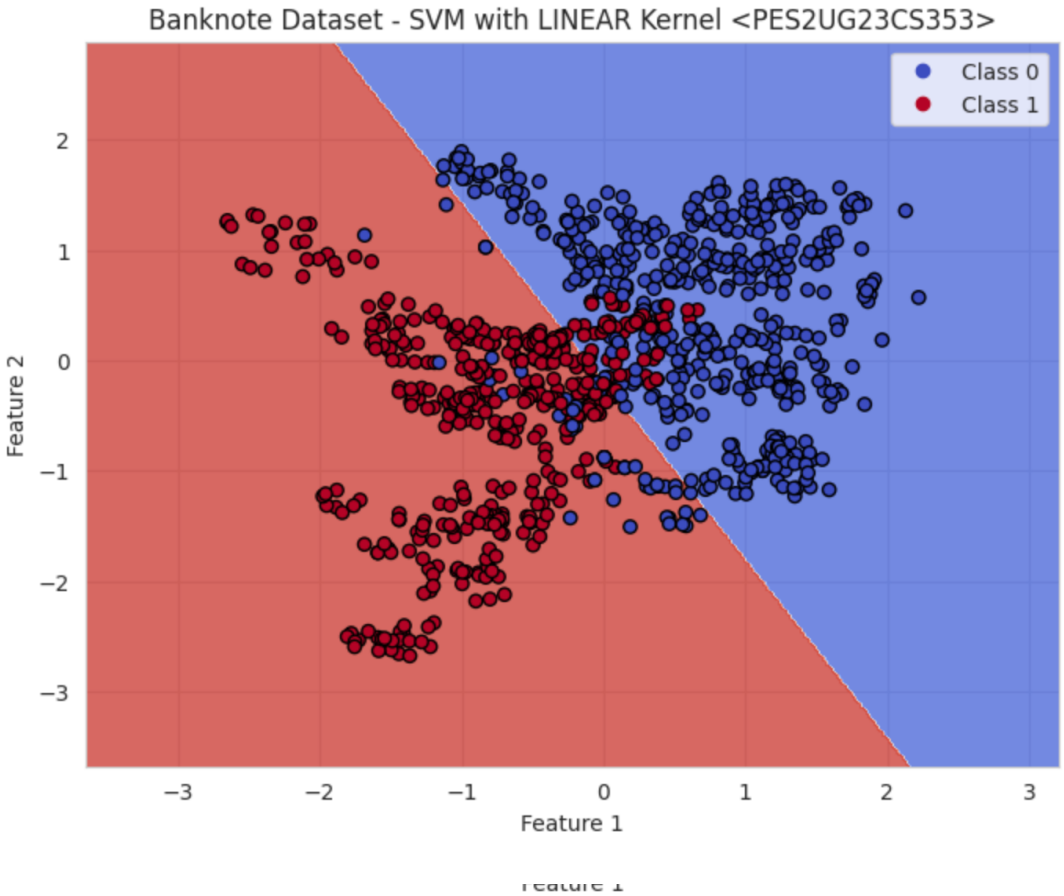
MOONS DATASET:

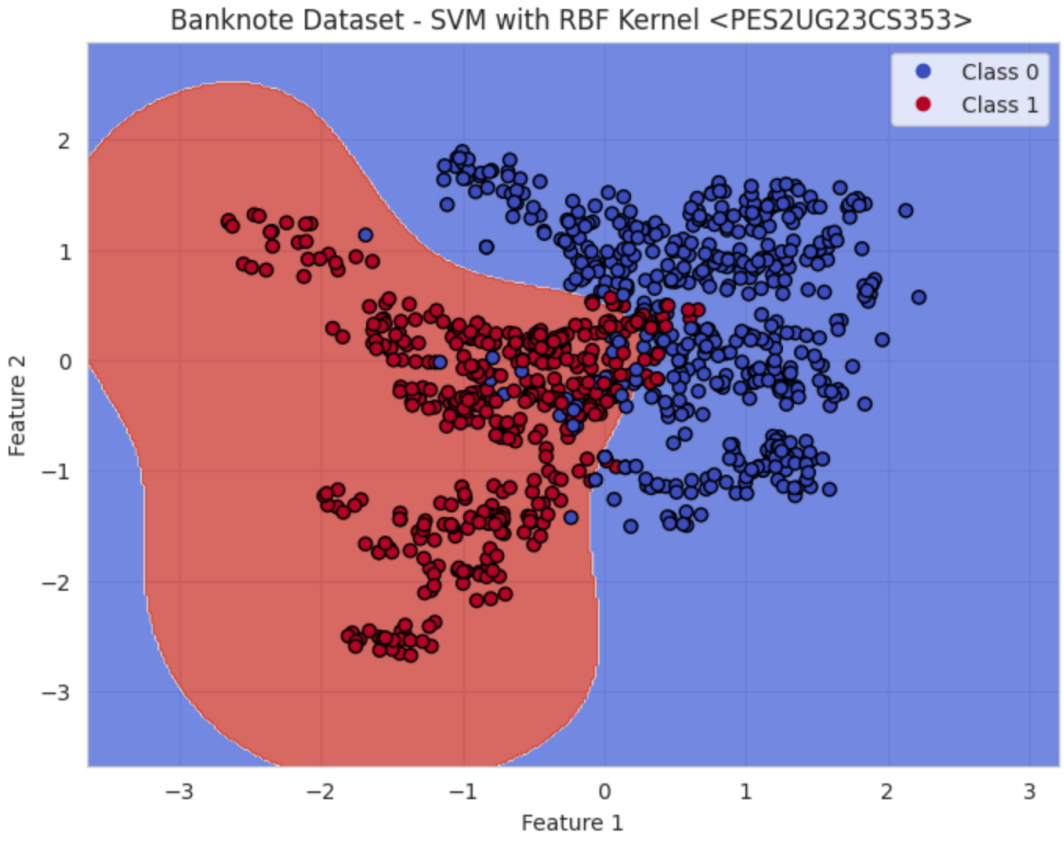


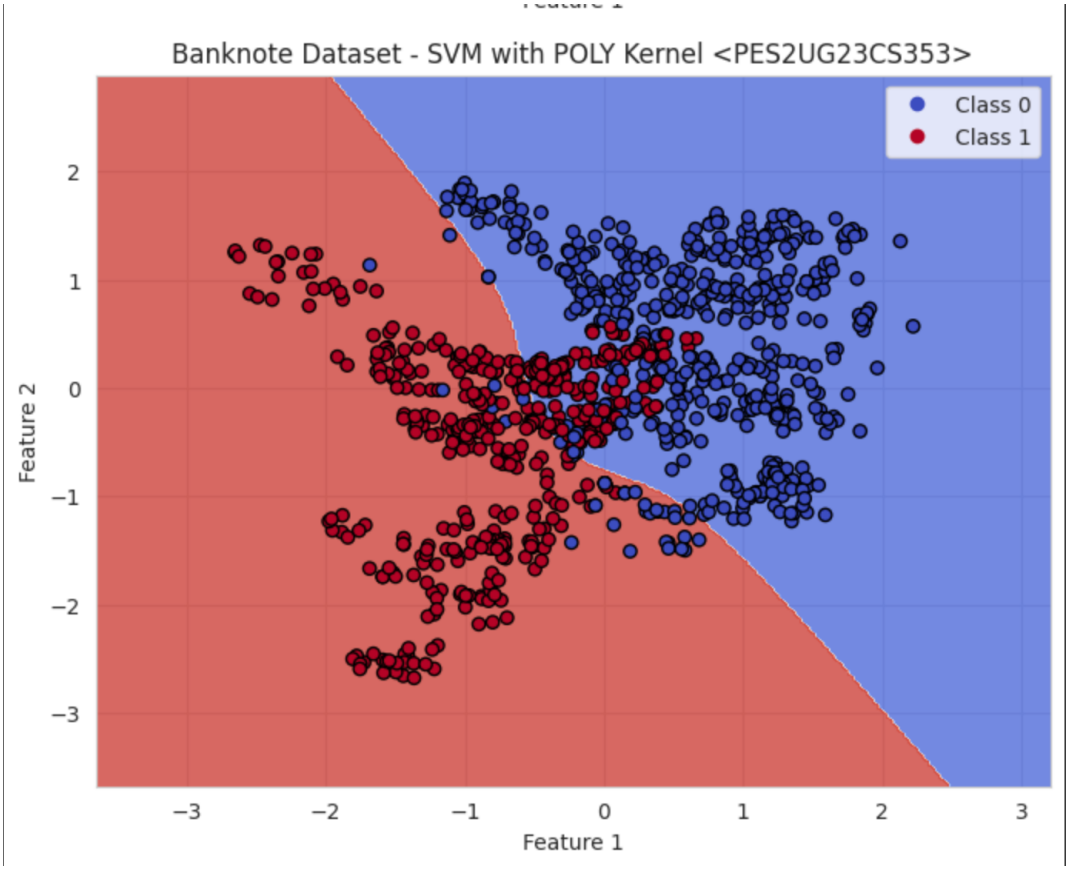

Moons Dataset - SVM with RBF Kernel <PES2UG23CS353>
Class 0
Class 1
Feature 2
Feature 1
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Moons Dataset - SVM with POLY Kernel <PES2UG23CS353>
Class 0
Class 1
Feature 2
Feature 1
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

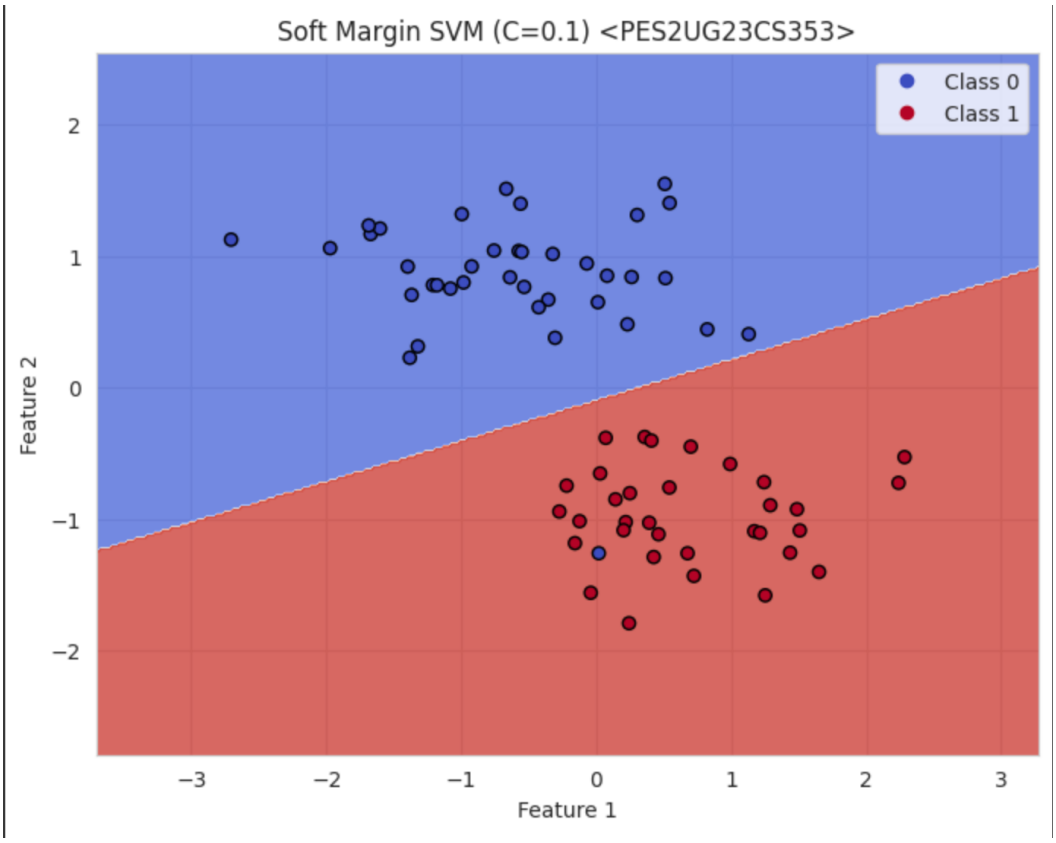
BANKNOTE DATASET:

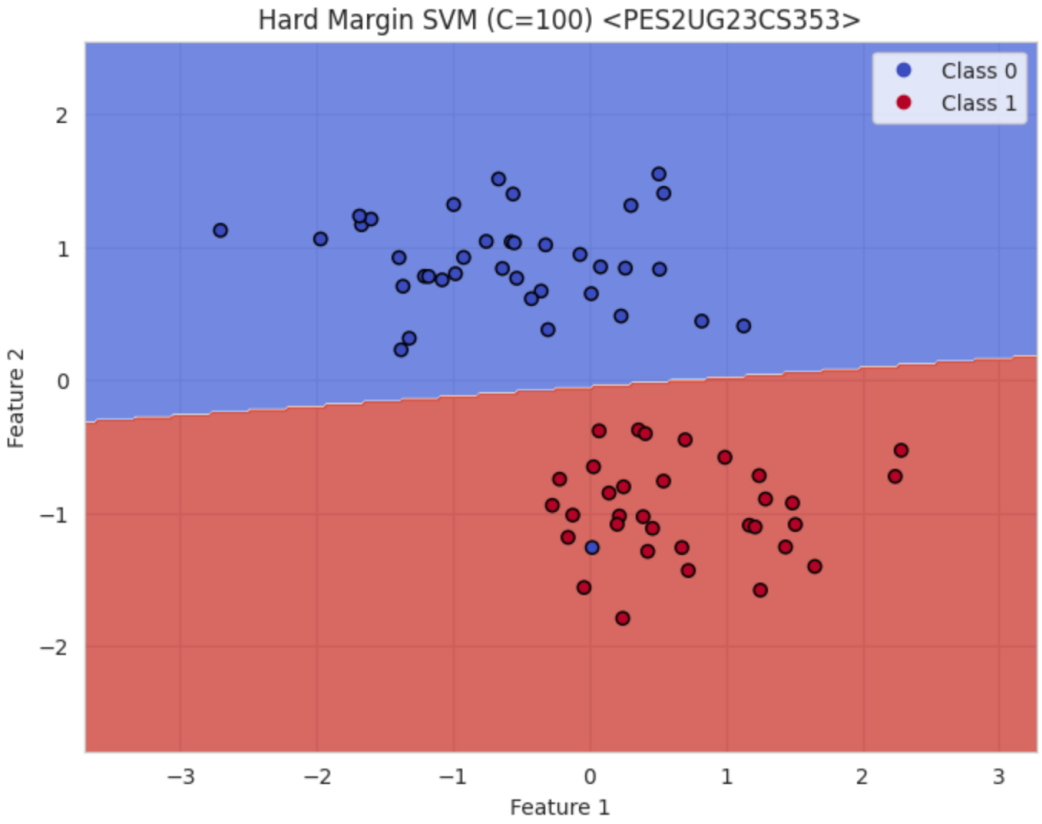






MARGIN ANALYSIS:





Analysis Questions

1. Compare the two plots. Which model, the "Soft Margin" (C=0.1) or the "Hard Margin" (C=100), produces a wider margin?

Ans-: From the plots, the Soft Margin (C=0.1) has a wider margin.

2. Look closely at the "Soft Margin" (C=0.1) plot. You'll notice some points are either inside the margin or on the wrong side of the decision boundary. Why does the SVM allow these "mistakes"? What is the primary goal of this model?

Ans-: It allows mistake to avoid overfitting and to handle non separable data.

The primary goal is better generalisation.

3. Which of these two models do you think is more likely to be overfitting to the training data? Explain your reasoning.

Ans-: Hard Margin (C=100) because it tries to fit the training data perfectly, so its

sensitive to outliers and noise.

4. Imagine you receive a new, unseen data point. Which model do you trust more to classify it correctly? Why? In a real-world scenario where data is often noisy, which value of C (low or high) would you generally prefer to start with?

Ans-: For a new unseen data point, Soft Margin (C=0.1) is more trustworthy because it's less overfit. In real world noisy data, starting with a low C is generally preferred because it prioritizes a wider margin and is more robust to noise.