ML Lab Week 10

SVM Lab

Name: Monisha Sharma

SRN: PES2UG23CS906

Class: 5F

<u>Moons Dataset Questions (2 questions):</u>

1. Inferences about the Linear Kernel's performance.

   The Linear kernel performs poorly on the Moons dataset because the data is non-linearly separable. The decision boundary is a straight line, which cannot capture the curved half-moon shapes.

   Metrics like accuracy, precision, and recall are lower compared to RBF and Polynomial kernels, confirming that linear separation is insufficient.
2. Comparison between RBF and Polynomial kernel decision boundaries.

   The RBF kernel captures the curved shape of the moons dataset more naturally, creating a smooth, non-linear boundary that separates the two classes accurately. The Polynomial kernel can capture some curvature but may underperform if the degree is not optimal or the data is noisy.

   RBF usually provides higher classification metrics for non-linear datasets like Moons.

<u>Banknote Dataset Questions (2 questions):</u>

1. In this case, which kernel appears to be the most effective?

   For the Banknote dataset, which is mostly linearly separable, the Linear kernel is highly effective. It achieves high accuracy because the two classes can be separated well with a straight line in the scaled feature space.

   RBF and Polynomial may perform slightly worse due to overfitting or unnecessary complexity.
2. Why might the Polynomial kernel have underperformed here?

   The Polynomial kernel may overfit on small or noisy datasets when linear separation is sufficient. Its curved decision boundary is unnecessary for linearly separable data, leading to misclassifications and lower generalization performance.

   A higher-degree polynomial can increase model complexity without improving accuracy.

<u>Hard vs. Soft Margin Questions (4 questions):</u>

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

   The Soft Margin (C=0.1) produces a wider margin. It allows some misclassifications to maximize the distance between the closest points of the classes, increasing generalization.

2 Why does the soft margin model allow "mistakes"?

Soft margin SVM allows misclassifications to handle noisy or imperfect data. The primary goal is to maximize the margin between classes while tolerating some errors, which reduces overfitting and improves performance on unseen data.

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

The Hard Margin (C=100) is more likely to overfit because it tries to classify all points correctly, including outliers. This results in a narrower margin and poor generalization to new or noisy data.

4 Which model would you trust more for new data and why?

The Soft Margin model (low C) is more reliable for new data. It generalizes better by allowing some tolerance for noise and outliers. In real-world scenarios, it is recommended to start with a low C value to prevent overfitting.

Screenshots:

Training:

Moons Dataset (3 screenshots):

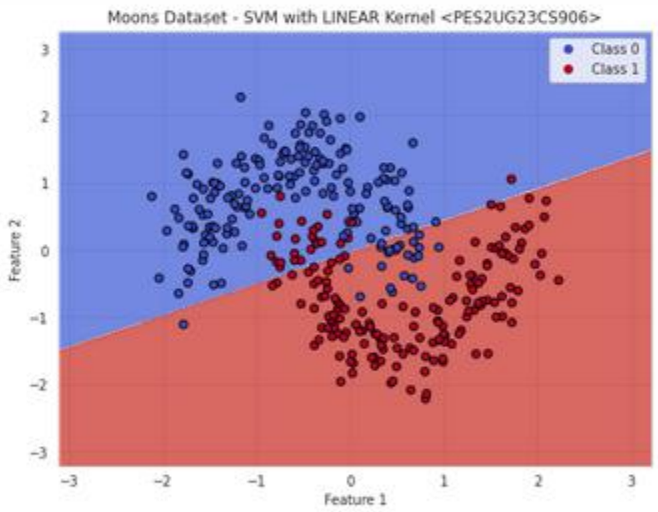
```
SVM with LINEAR Kernel <PES2UG23CS906>
              precision    recall  f1-score   support

           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 <PES2UG23CS906>
              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 <PES2UG23CS906>
              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

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

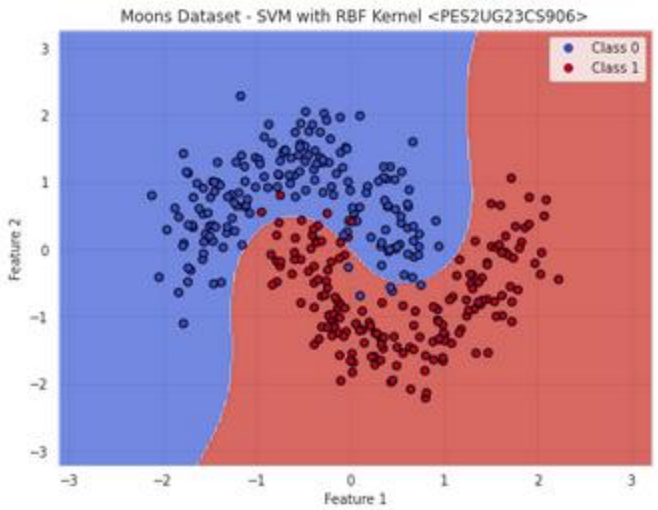
Banknote Dataset (3 screenshots):

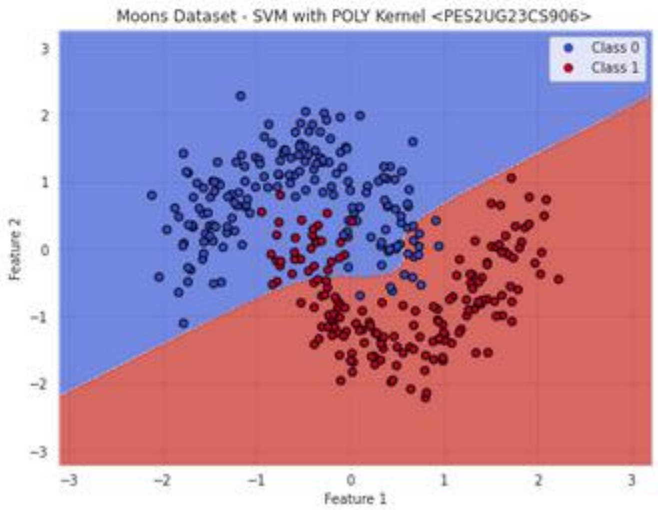
```
SVM with LINEAR Kernel <PES2UG23CS906>
              precision    recall  f1-score   support

      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 <PES2UG23CS906>
              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 <PES2UG23CS906>
              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:

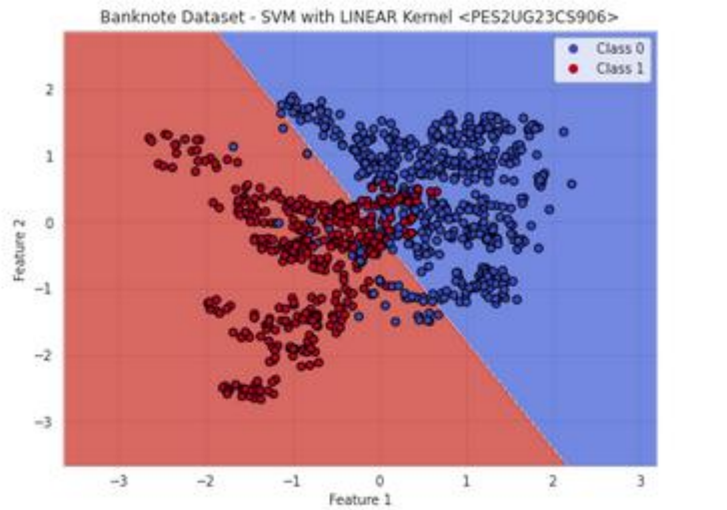
Moons Dataset (3 plots):

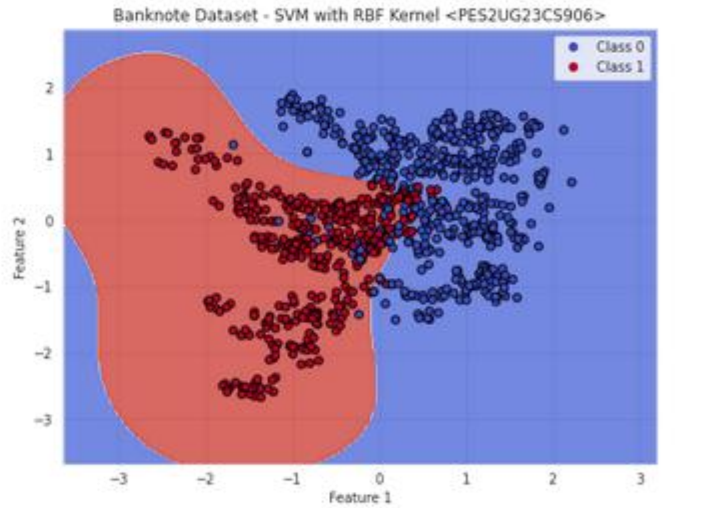


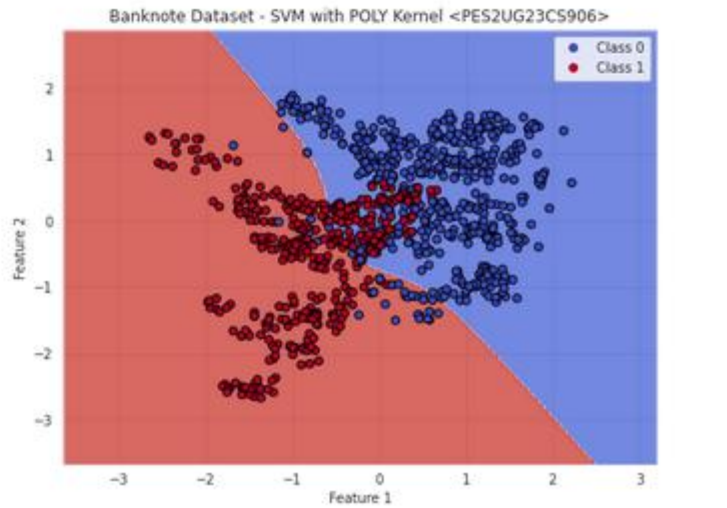




Banknote Dataset (3 plots):







Margin Analysis (2 plots):