ML Lab Week 10: SVM Classifier Lab Report

Name: Lakshya Pachisia
SRN: PES1UG23CS325
Section: F

2. Analysis Questions

i) Analysis Questions for Moons:

1. Based on the metrics and the visualizations, what inferences about the performance of the Linear Kernel can you draw?
2. Compare the decision boundaries of the RBF and Polynomial kernels. Which one seems to capture the shape of the data more naturally?

Answers:-

1. Based on the metrics and visualizations for the Moons dataset, the Linear Kernel performs reasonably well with an accuracy of 0.87. However, the visualization shows that the linear boundary struggles to effectively separate the two crescent shapes of the Moons dataset, indicating it's not ideal for this non-linearly separable data.

2. Comparing the decision boundaries of the RBF and Polynomial kernels on the Moons dataset, the RBF kernel appears to capture the shape of the data more naturally. The RBF boundary closely follows the curves of the moon shapes, while the polynomial boundary is more rigid and doesn't conform as well to the data's non-linear structure.

ii) Analysis Questions for Banknote:

1. In this case, which kernel appears to be the most effective?
2. The Polynomial kernel shows lower performance here compared to the Moons dataset. What might be the reason for this?

Answers:-

1. In the case of the Banknote Authentication dataset, the RBF kernel appears to be the most effective. It achieved the highest accuracy (0.93) and well-balanced precision and recall for both classes, as seen in the classification report. The visualization also shows a clear separation of the classes.

2. The Polynomial kernel shows lower performance on the Banknote dataset

compared to the Moons dataset likely because the relationship between the selected features ('variance' and 'skewness') and the target variable ('class') is not well-captured by a polynomial function in this 2D projection. While the Polynomial kernel can create curved boundaries, the specific polynomial degree used might not be suitable, or the optimal separation might be better achieved with a different kernel like RBF, which is more flexible in capturing complex, non-linear relationships without explicitly defining the polynomial degree.

iii) Analysis Questions for Hard vs Soft Margin

1. Compare the two plots. Which model, the "Soft Margin" (C=0.1) or the "Hard Margin" (C=100), produces 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?
3. Which of these two models do you think is more likely to be overfitting to the training data? Explain your reasoning.
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?

Answers:-

1. Comparing the two plots, the "Soft Margin" model (C=0.1) clearly produces a wider margin than the "Hard Margin" model (C=100).

2. The "Soft Margin" (C=0.1) plot shows some points inside the margin or on the wrong side of the decision boundary. The SVM allows these "mistakes" because the primary goal of a soft margin model is to find a balance between maximizing the margin width and minimizing the classification errors. It prioritizes a wider margin for better generalization, even if it means misclassifying a few training points, especially in the presence of

noise or outliers.

3. The "Hard Margin" model (C=100) is more likely to be overfitting to the training data. This is because a large C value forces the model to try and classify every training point correctly, resulting in a very narrow margin that is highly influenced by individual data points, including outliers. This tight fit to the training data makes it less likely to generalize well to unseen data.

4. In the case of a new, unseen data point, the "Soft Margin" model (C=0.1) is generally more trustworthy to classify it correctly. This is because its wider margin makes it more robust to variations and noise in the data. In a real-world scenario where data is often noisy, you would generally prefer to start with a low value of C (soft margin) to avoid overfitting and achieve better generalization performance.

3. Screenshots

- **Training Results**

  i) Moons Dataset

1. **Classification Report for SVM with LINEAR Kernel with SRN**

```
SVM with LINEAR Kernel PES1UG23CS325
              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
```

2. **Classification Report for SVM with RBF Kernel with SRN**

```
SVM with RBF Kernel PES1UG23CS325
              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
```

3. **Classification Report for SVM with POLY Kernel with SRN**

```
SVM with POLY Kernel PES1UG23CS325
              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
```

ii) Banknote Dataset

4. Classification Report for SVM with LINEAR Kernel

```
SVM with LINEAR Kernel PES1UG23CS325
              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
```

5. Classification Report for SVM with RBF Kernel

```
SVM with RBF Kernel PES1UG23CS325
              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
```

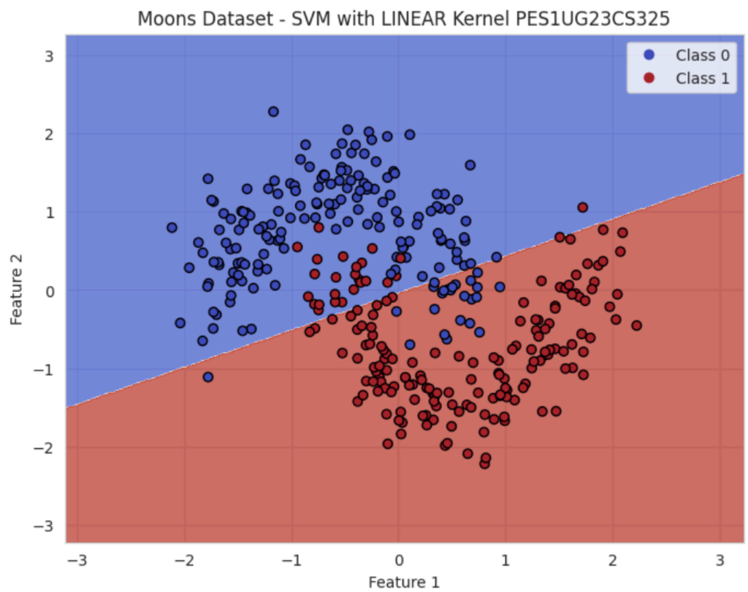
6. **Classification Report for SVM with POLY Kernel**

```
SVM with POLY Kernel PES1UG23CS325
              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
```

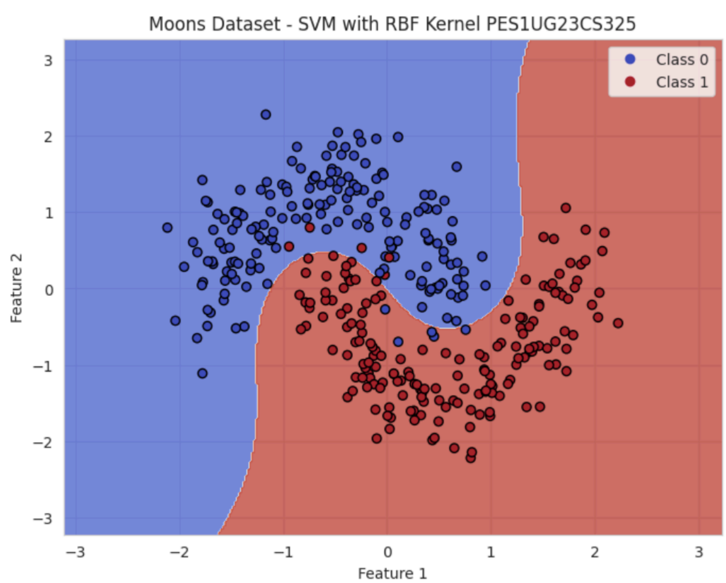
- **Decision Boundary Visualizations**

  i) Moons Dataset

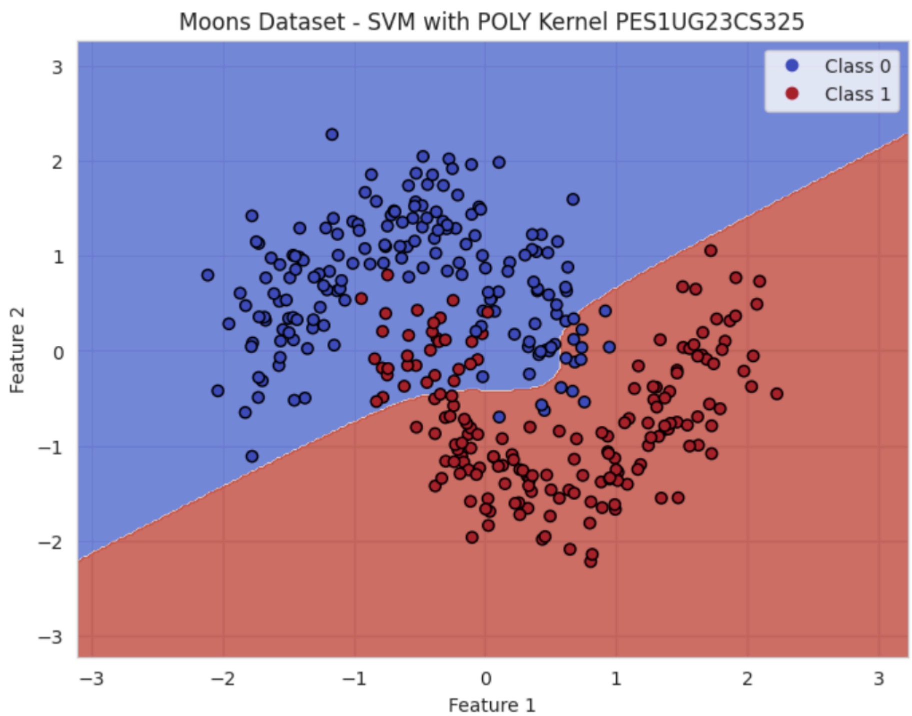
  7. Classification Report for SVM with LINEAR Kernel with SRN



  8. Classification Report for SVM with RBF Kernel with SRN

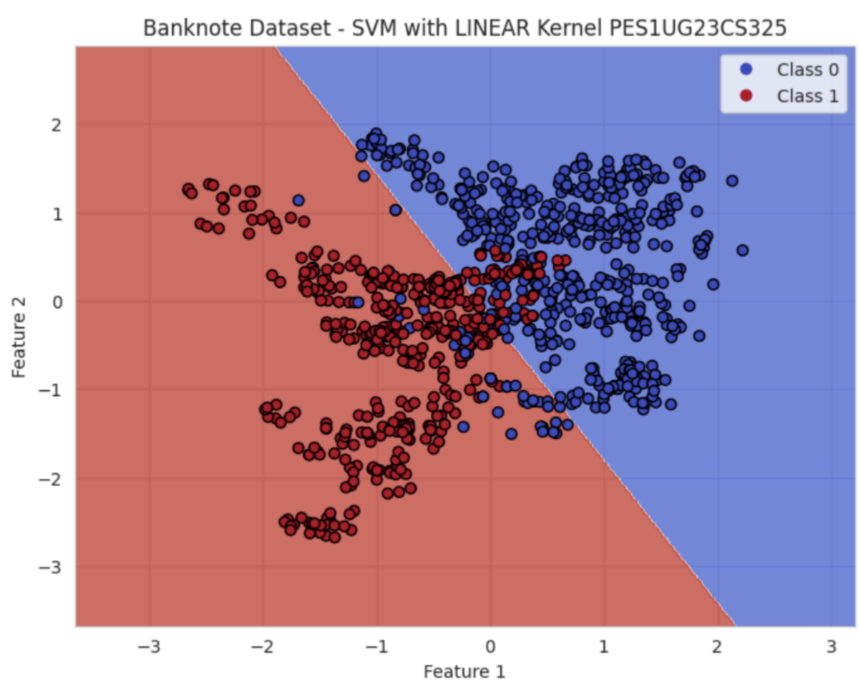


  9. Classification Report for SVM with POLY Kernel with SRN

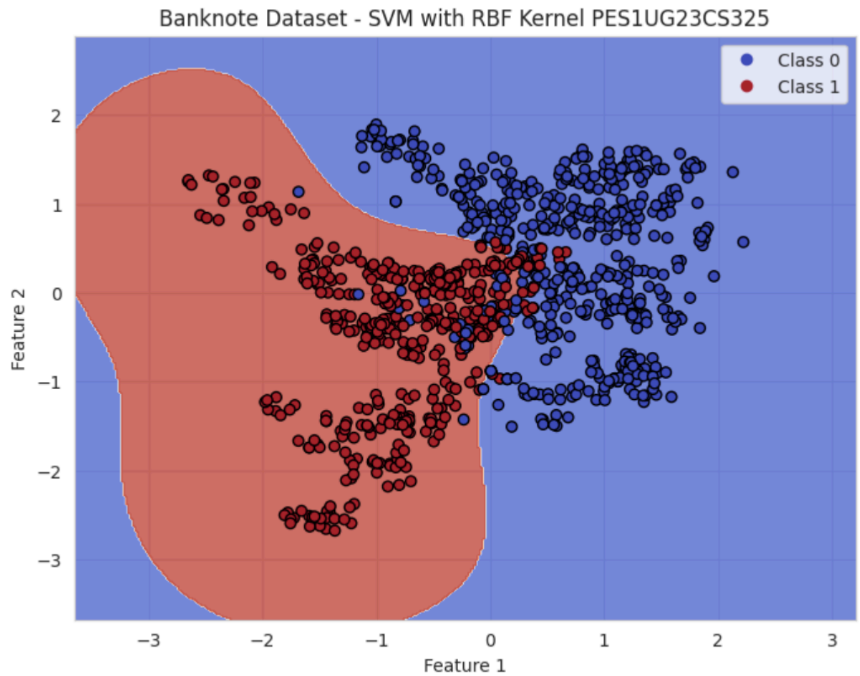


ii) Banknote Dataset

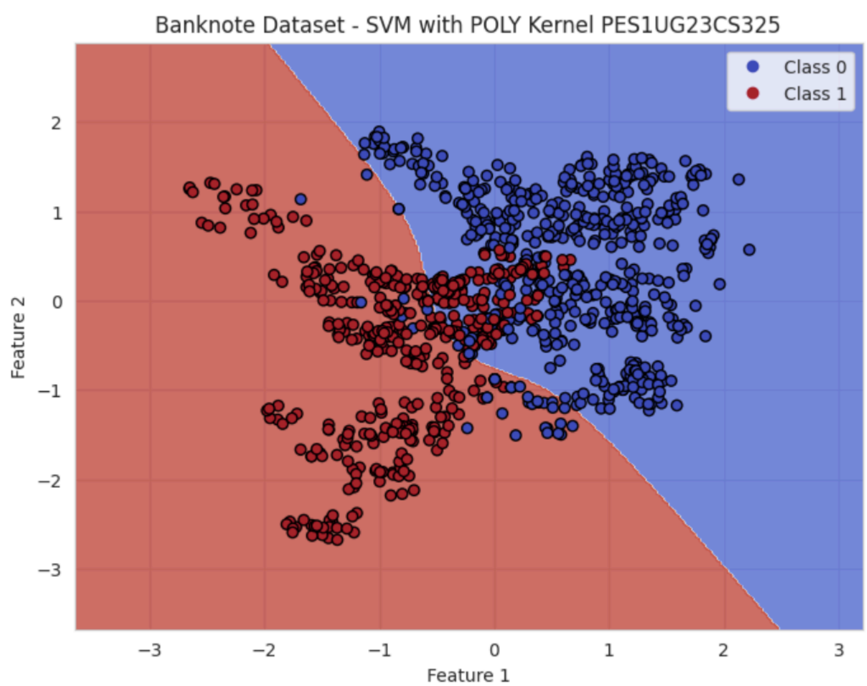
10. Classification Report for SVM with LINEAR Kernel



11. Classification Report for SVM with RBF Kernel

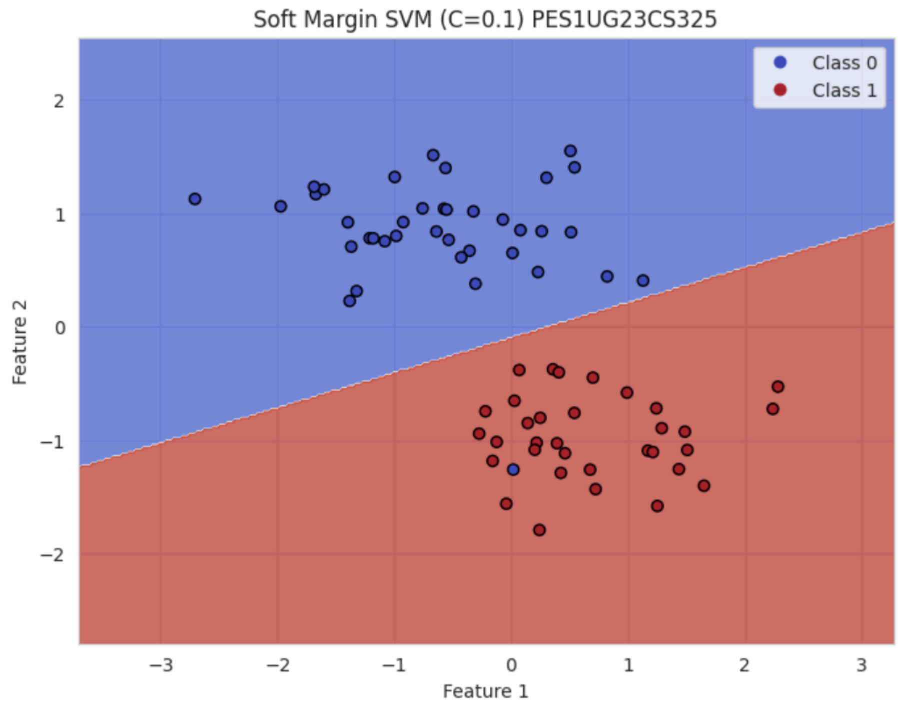


12. Classification Report for SVM with POLY Kernel



iii) Margin Analysis

13. Classification Report for SVM with POLY Kernel



14. Classification Report for SVM with POLY Kernel