Project Title:
Week 7 Lab – SVM Classifier Lab

Name:
Pinisetti Sudhiksha

SRN:
PES2UG23CS916

Course:
UE23CS352A: Machine Learning

Date:
09-10- 2025

Analysis Questions

Analysis Questions for Moons:

1. Based on the metrics and the visualizations, what inferences about the performance of the Linear Kernel can you draw?

From the visualization of the Linear Kernel, it can be observed that the decision boundary is a straight line dividing the two classes. Since the Moons dataset is non-linear in nature, this kind of boundary is not able to properly capture the curved pattern of the data. As a result, several points from both classes are misclassified near the boundary. This shows that the Linear Kernel underfits the data and does not adapt well to its structure. While it may provide a simple baseline, its performance is clearly limited for this type of dataset.

2. Compare the decision boundaries of the RBF and Polynomial kernels. Which one seems to capture the shape of the data more naturally?

When comparing the RBF and Polynomial kernels, the RBF Kernel seems to follow the shape of the data more naturally. The decision boundary created by the RBF Kernel curves smoothly around the two classes, successfully separating most of the points with very few misclassifications. On the other hand, the Polynomial Kernel does not capture the crescent shape as effectively. Its boundary is more rigid and resembles the linear decision boundary to some extent, which leads to a higher chance of misclassification along the curve. Overall, the RBF Kernel provides a better fit for the Moons dataset compared to both the Linear and Polynomial kernels, because it adapts to the non-linear pattern of the data.

Analysis Questions for Banknote:

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

From the visualizations, it can be observed that the RBF kernel gives the most effective decision boundary. It creates a smooth and flexible curve that better separates the two classes (Class 0 and Class 1) compared to the Linear and Polynomial kernels. The Linear and Polynomial kernels both show almost straight boundaries that cut through the center of the data, causing more misclassified points. On the other hand, the RBF kernel adapts to the distribution of the data more closely, reducing misclassifications and achieving a better fit overall.

2. The Polynomial kernel shows lower performance here compared to the Moons dataset. What might be the reason for this?

The Polynomial kernel performs worse here compared to the Moons dataset because the banknote data is more linearly separable with some overlapping regions. In this case, a high-degree polynomial boundary may not provide additional benefits and can even make the decision boundary unnecessarily complex. In contrast, the Moons dataset is highly non-linear, so a Polynomial kernel was more effective there. But for the banknote data, the

Polynomial kernel does not capture any extra useful patterns, leading to similar or worse performance compared to Linear and RBF kernels.

Analysis Questions for Hard and Soft Margins:

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

When comparing the two plots, it is clear that the Soft Margin SVM (C = 0.1) has a wider margin compared to the Hard Margin SVM (C = 100).
This is because a small C value allows the model to tolerate more misclassifications and focus on maximizing the width of the margin.
In contrast, the Hard Margin SVM uses a very large C value, which forces the model to minimize classification errors strictly, resulting in a narrower 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?

In the Soft Margin plot, some points can be seen inside the margin or on the wrong side of the decision boundary. This happens because a small C value allows the SVM to be more flexible, giving up on perfectly classifying every training point. The primary goal of the Soft Margin model is not to classify every single point correctly, but to maximize the margin between classes while still maintaining a good overall fit. This helps the model generalize better to unseen data.

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

Between the two models, the Hard Margin SVM (C = 100) is more likely to overfit the training data. A large C value forces the model to correctly classify almost all training points, which can lead to a decision boundary that is too specific to the training set. This makes the model more sensitive to noise and less generalizable.

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?

For new, unseen data points, the Soft Margin SVM is more reliable.
This is because it is more tolerant of noise and doesn’t try to perfectly fit every training point.
In real-world datasets, which are often noisy and imperfect, starting with a lower C value is generally a better choice.
A low C helps the model generalize better instead of memorizing the training data.

Screenshots

Training Results

Moons Dataset

```
SVM with LINEAR Kernel <PES2UG23CS916>
              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 <PES2UG23CS916>
              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 <PES2UG23CS916>
              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 <PES2UG23CS916>

	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 <PES2UG23CS916>

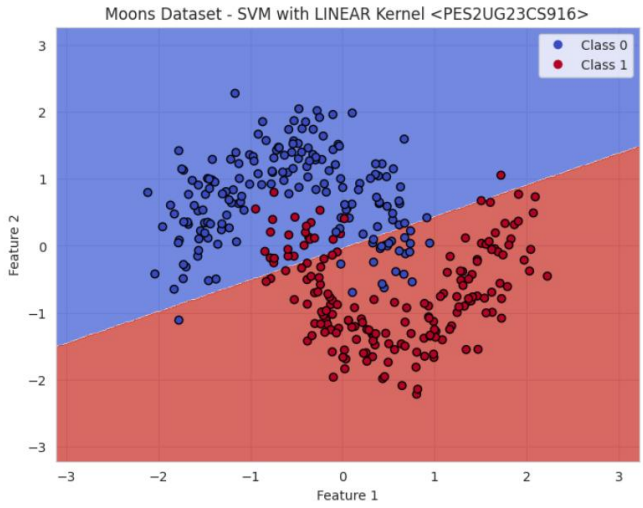
	precision	recall	f1-score	support
Forged	0.96	0.91	0.94	229
Genuine	0.90	0.96	0.93	183
accuracy			0.93	412
macro avg	0.93	0.93	0.93	412
weighted avg	0.93	0.93	0.93	412

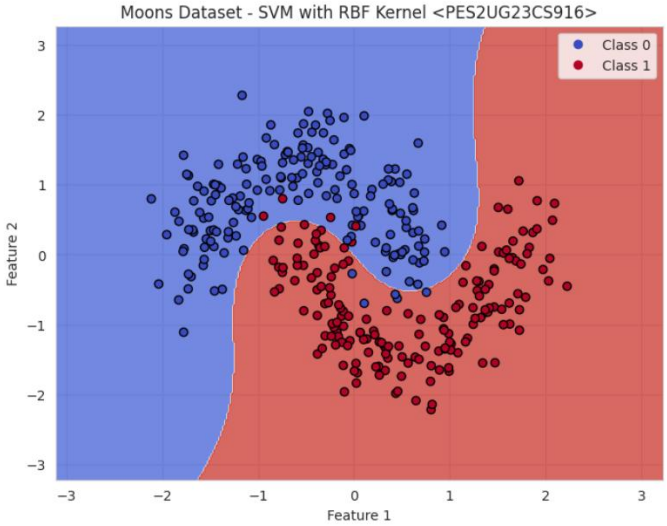
SVM with POLY Kernel <PES2UG23CS916>

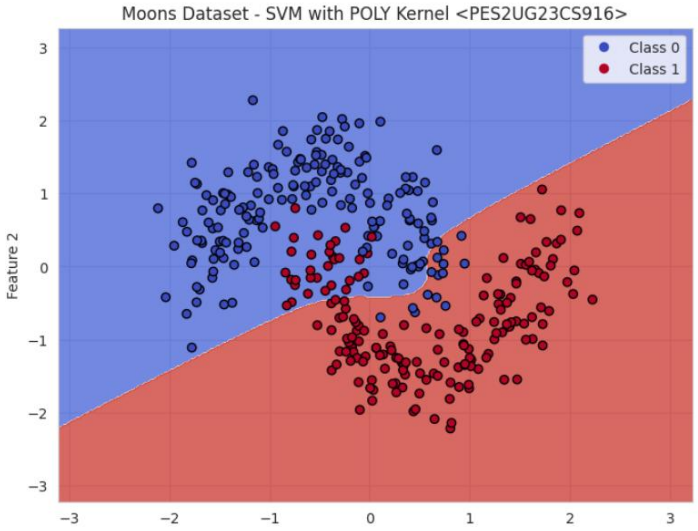
	precision	recall	f1-score	support
Forged	0.82	0.91	0.87	229
Genuine	0.87	0.75	0.81	183
accuracy			0.84	412
macro avg	0.85	0.83	0.84	412
weighted avg	0.85	0.84	0.84	412

Decision Boundary Visualizations

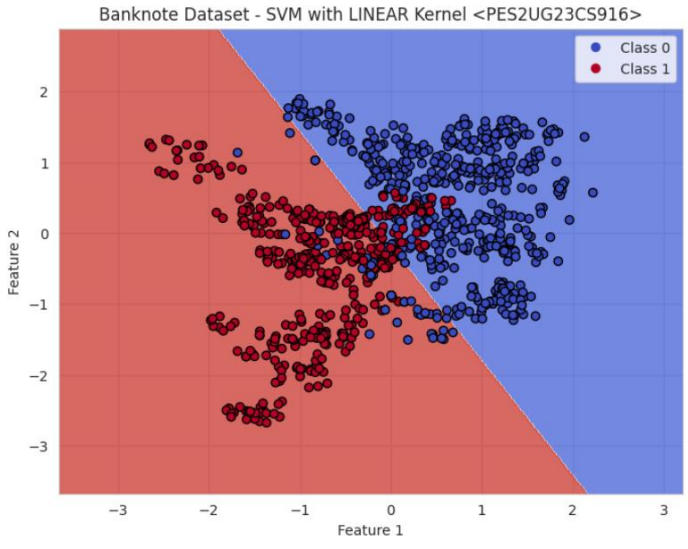
Moons Dataset

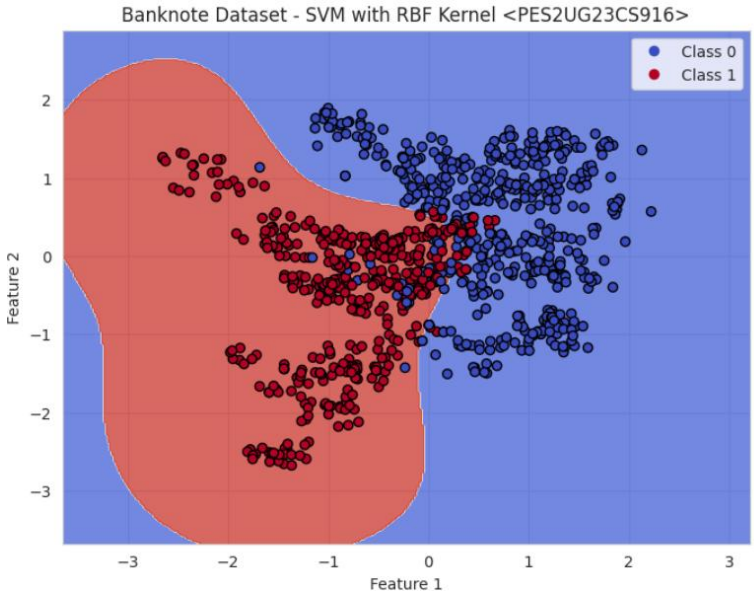


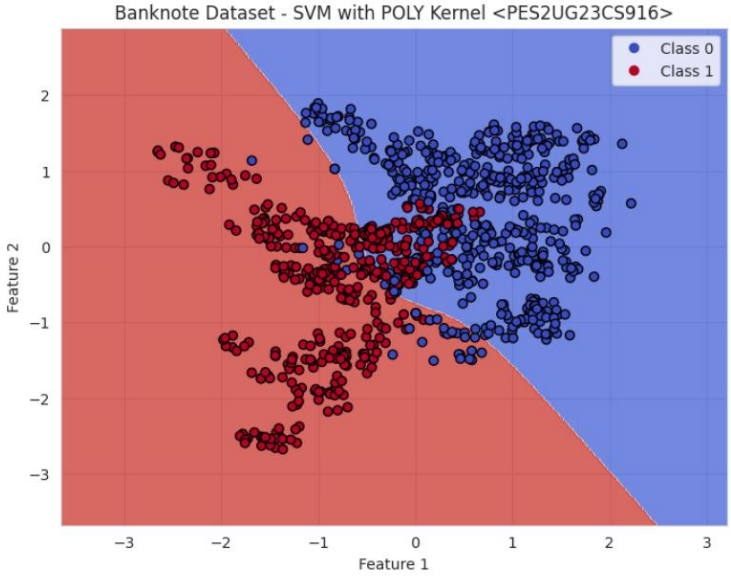




Banknote Dataset







Margin Analysis