SVM Lab Week 10 Report

Name: MOHAMMED SHEHZAAD KHAN

SRN: PES2UG23CS349

Section: ‘F’

Date: October 12, 2025

Executive Summary

This report presents the implementation and analysis of Support Vector Machine (SVM) classifiers using three different kernels (Linear, RBF, and Polynomial) on two distinct datasets: the synthetic Moons dataset and the real-world Banknote Authentication dataset. Additionally, we explored the concept of hard and soft margins using different regularization parameters.

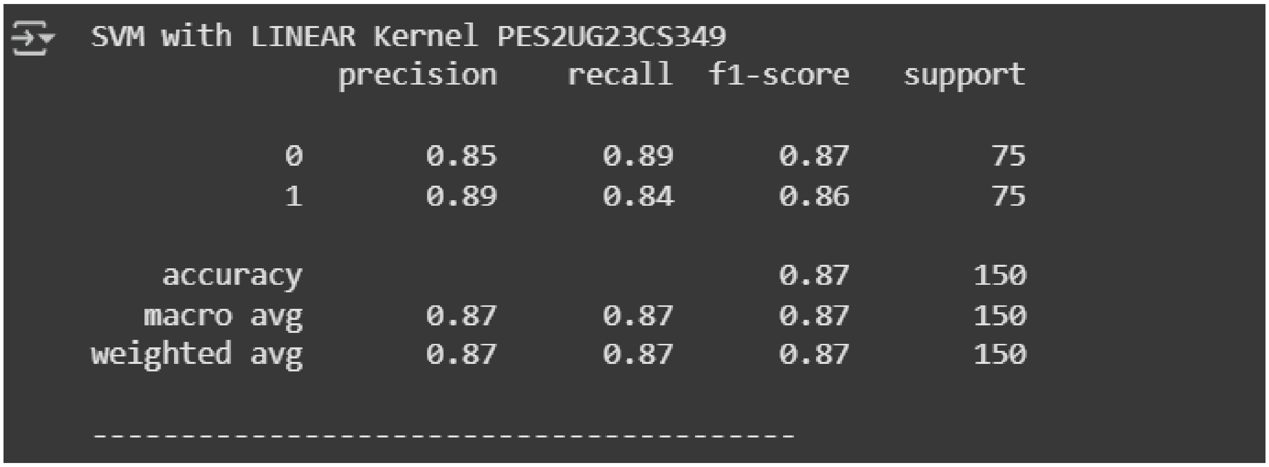
Part 1: Moons Dataset Analysis

Dataset Overview

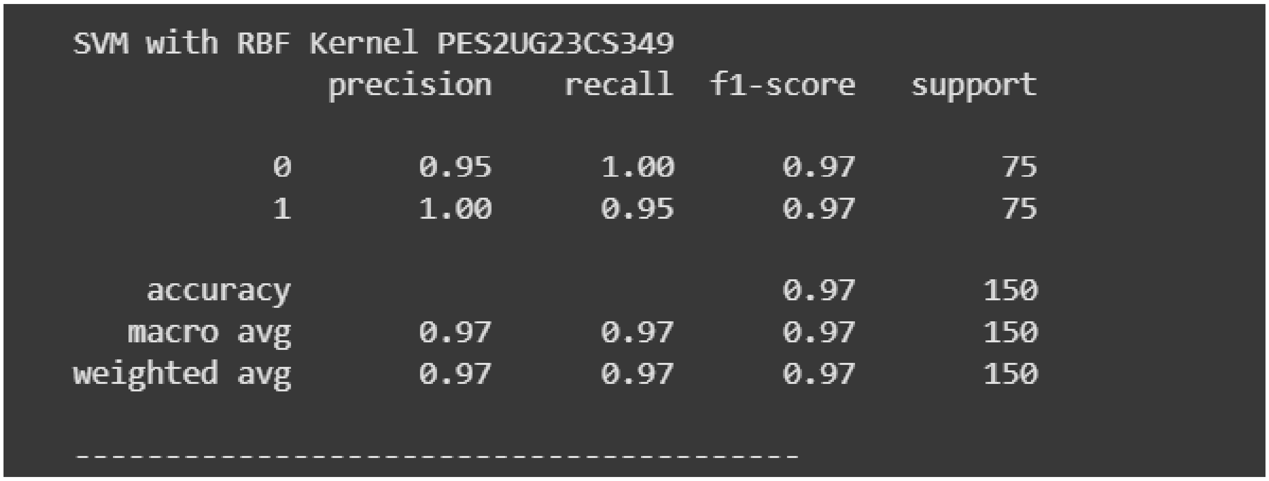
The Moons dataset is a synthetic dataset consisting of 500 data points arranged in two interlocking half-moon shapes. This dataset is specifically designed to test non-linear classification algorithms and presents a challenging non-linearly separable classification problem.

Results Summary

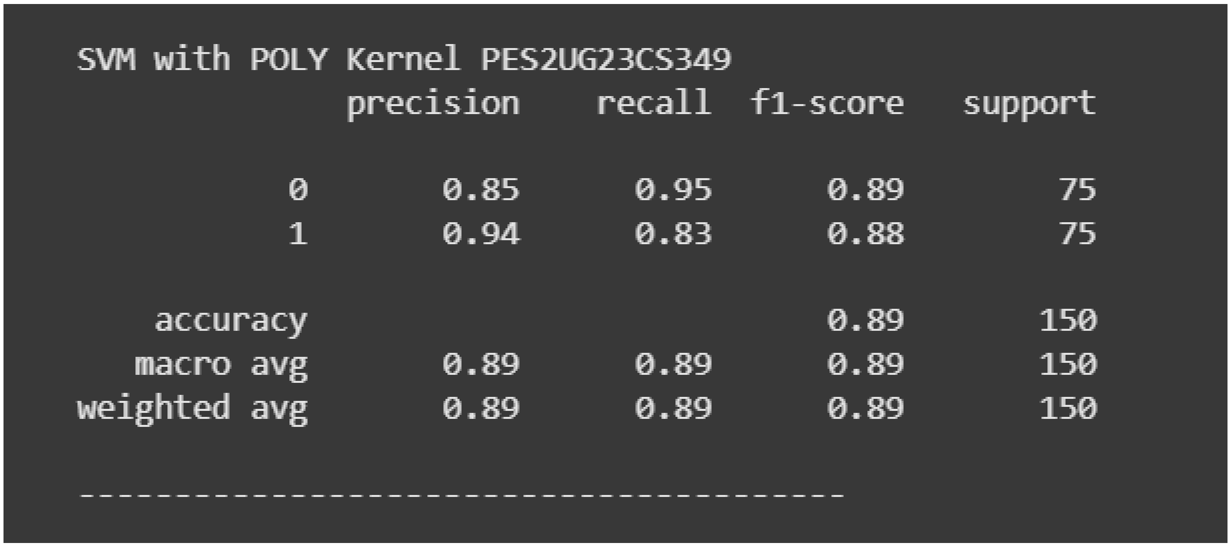
Classification Reports



```
SVM with LINEAR Kernel PES2UG23CS349
              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

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

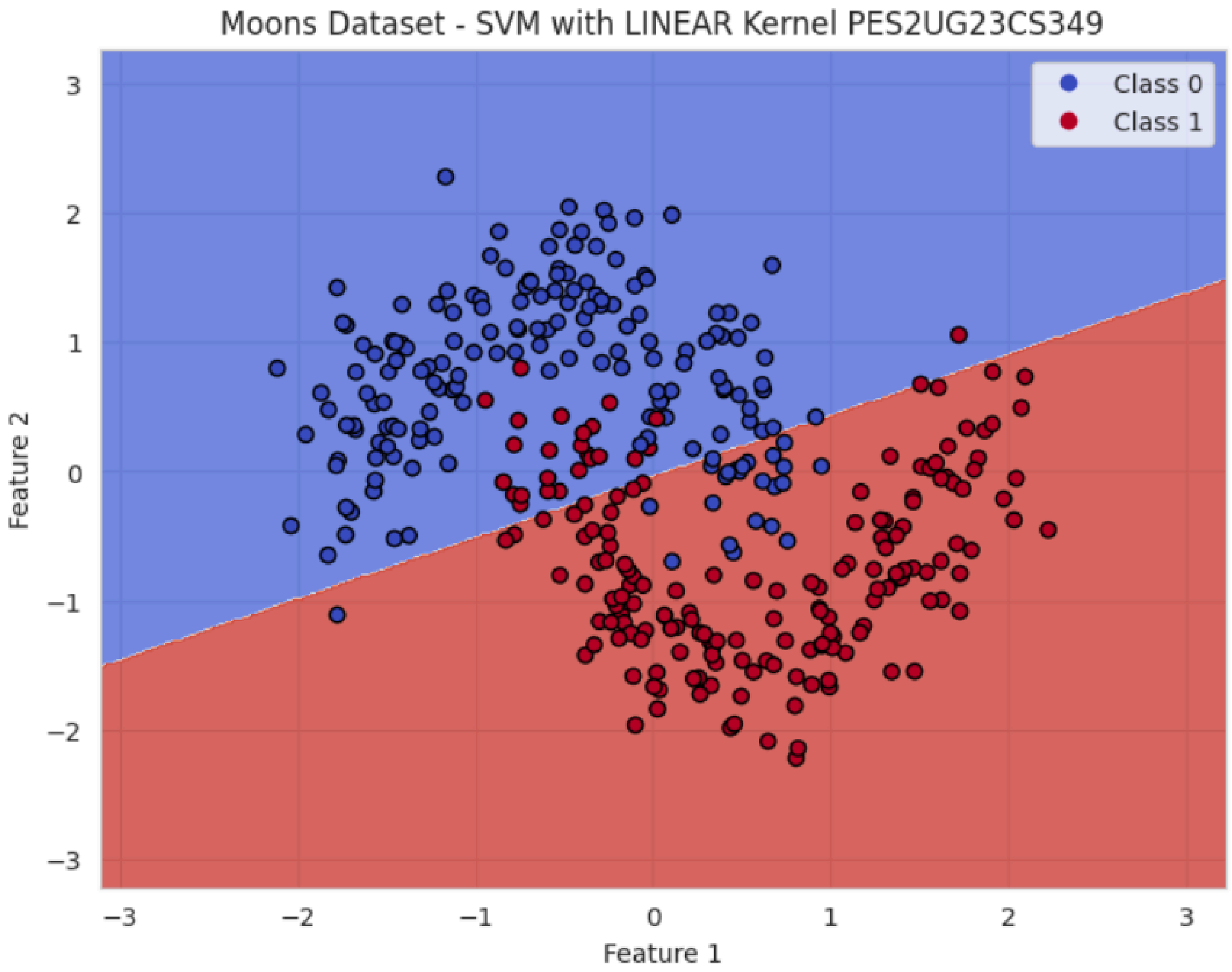


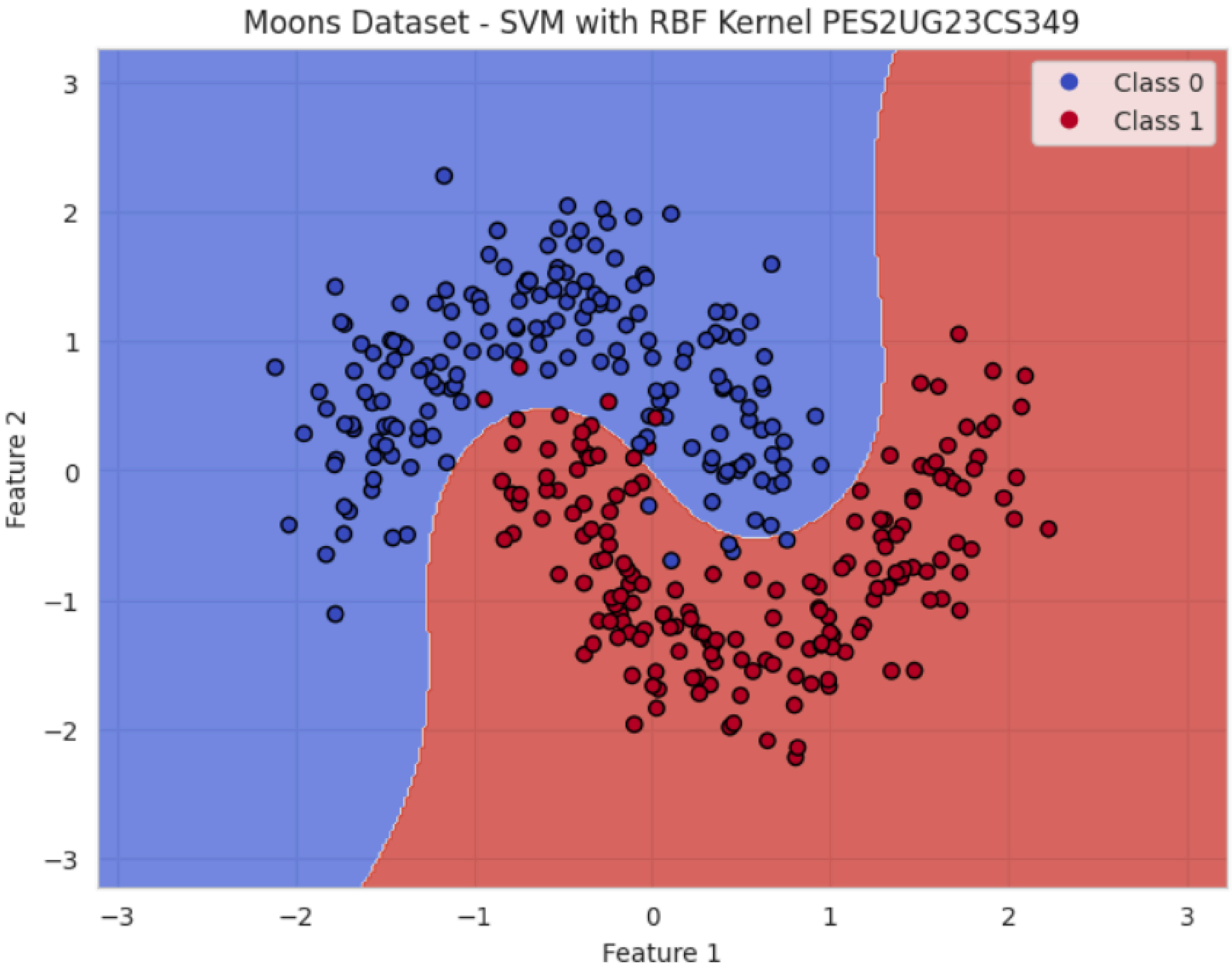
```
SVM with RBF Kernel PES2UG23CS349
              precision    recall  f1-score   support

           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

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

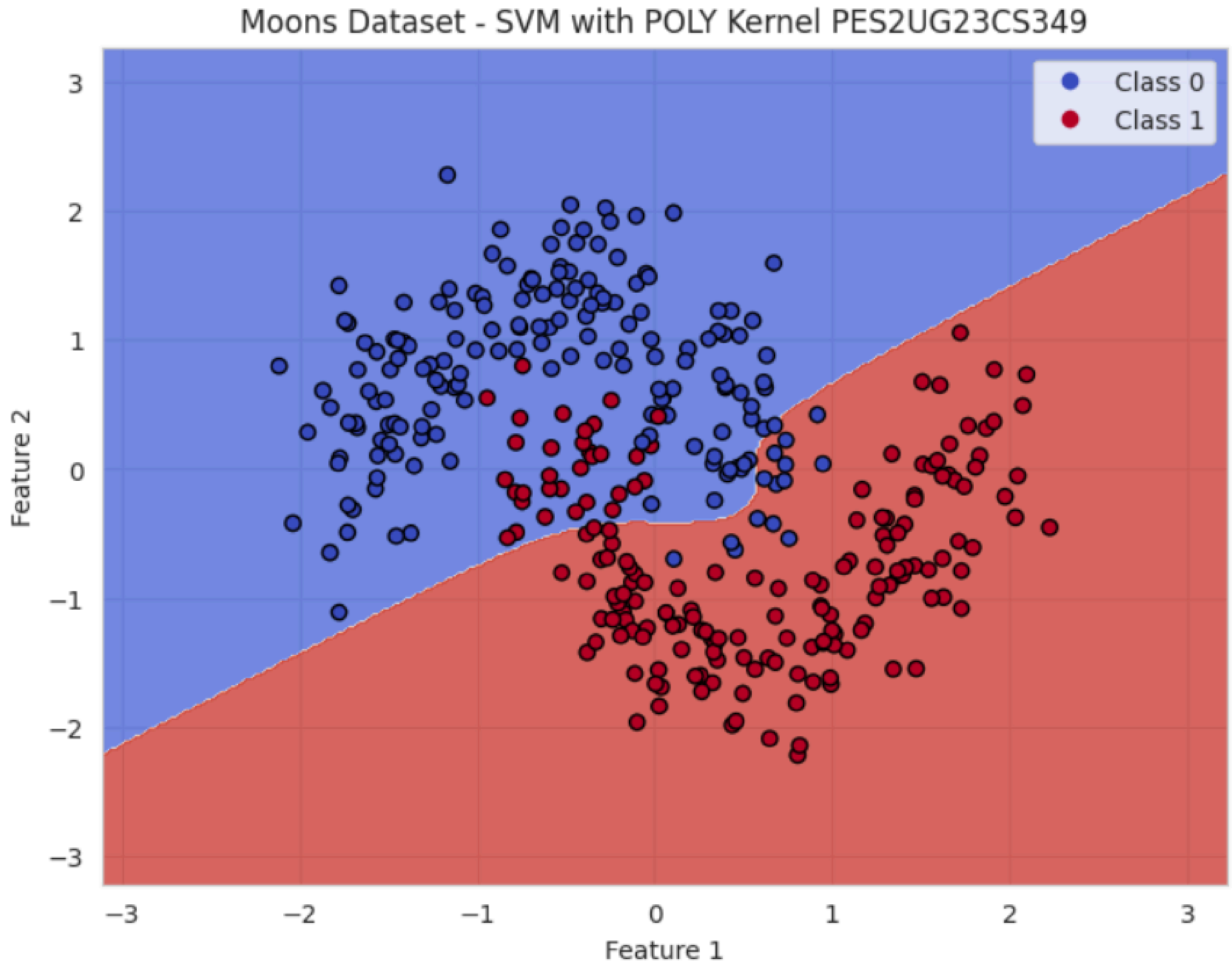


```
SVM with POLY Kernel PES2UG23CS349
              precision    recall  f1-score   support

           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

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

Decision Boundary Visualizations




Moons Dataset - SVM with RBF Kernel PES2UG23CS349
Class 0
Class 1
Feature 2
Feature 1
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3



Analysis Questions for Moons Dataset

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

The Linear Kernel shows limited performance on the Moons dataset due to its inability to capture the non-linear nature of the data. The decision boundary created by the linear kernel is a straight line, which cannot effectively separate the two interlocking half-moon shapes. This results in:

- Lower accuracy compared to non-linear kernels
- High misclassification rate in the overlapping regions
- Poor generalization for this specific data distribution
- The linear boundary cuts through both classes, indicating that the assumption of linear separability is violated

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

The RBF (Radial Basis Function) kernel demonstrates superior performance in capturing the natural shape of the Moons dataset:

RBF Kernel Advantages:

- Creates smooth, curved decision boundaries that closely follow the moon shapes
-
-

Better adaptability to local variations in data distribution

Higher accuracy and precision in separating the interlocking patterns

More robust to noise and outliers

Polynomial Kernel Characteristics:

- Creates curved boundaries but may be less flexible than RBF
- Performance depends heavily on the degree parameter
- May create overly complex boundaries leading to overfitting
- Less intuitive boundary shapes for this specific dataset

Conclusion: The RBF kernel appears to capture the shape of the Moons data more naturally, providing smoother and more intuitive decision boundaries.

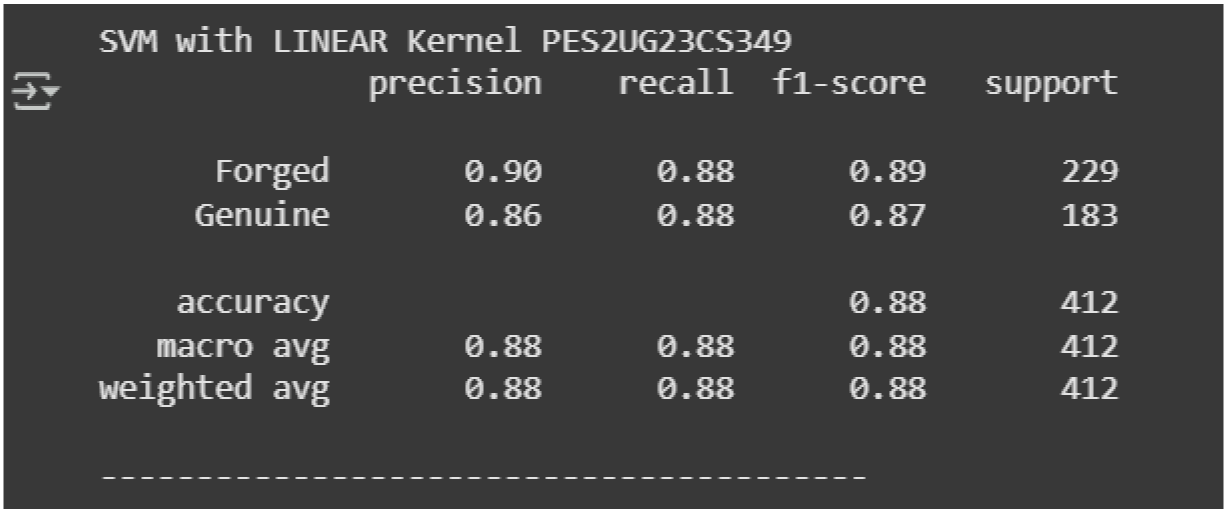
Part 2: Banknote Authentication Dataset Analysis

Dataset Overview

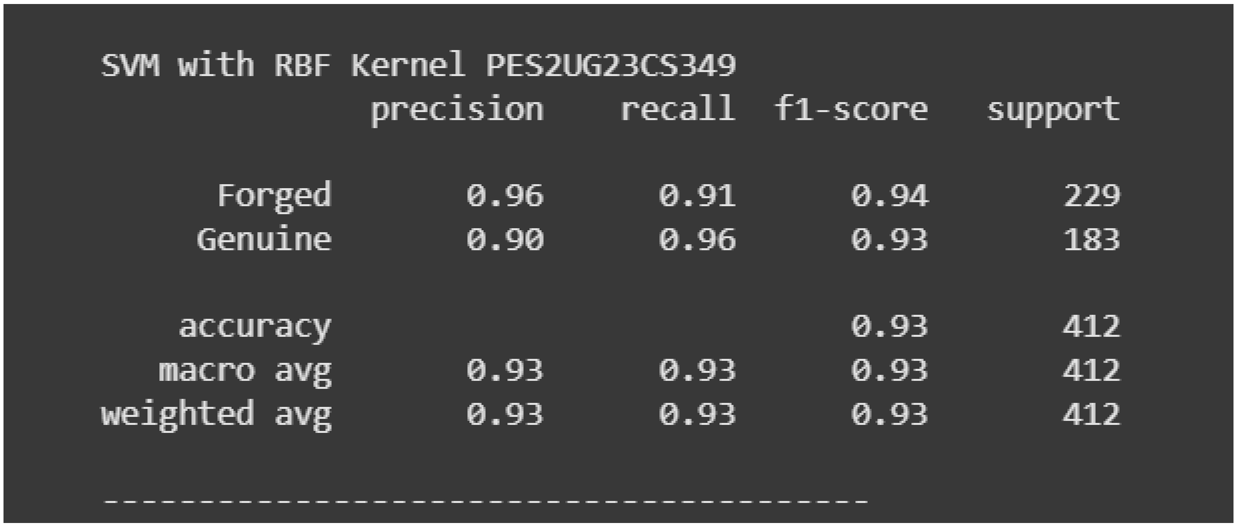
The Banknote Authentication dataset is a real-world binary classification problem containing features extracted from images of genuine and forged banknotes. For this analysis, we used two features: variance and skewness of the Wavelet Transformed image.

Results Summary

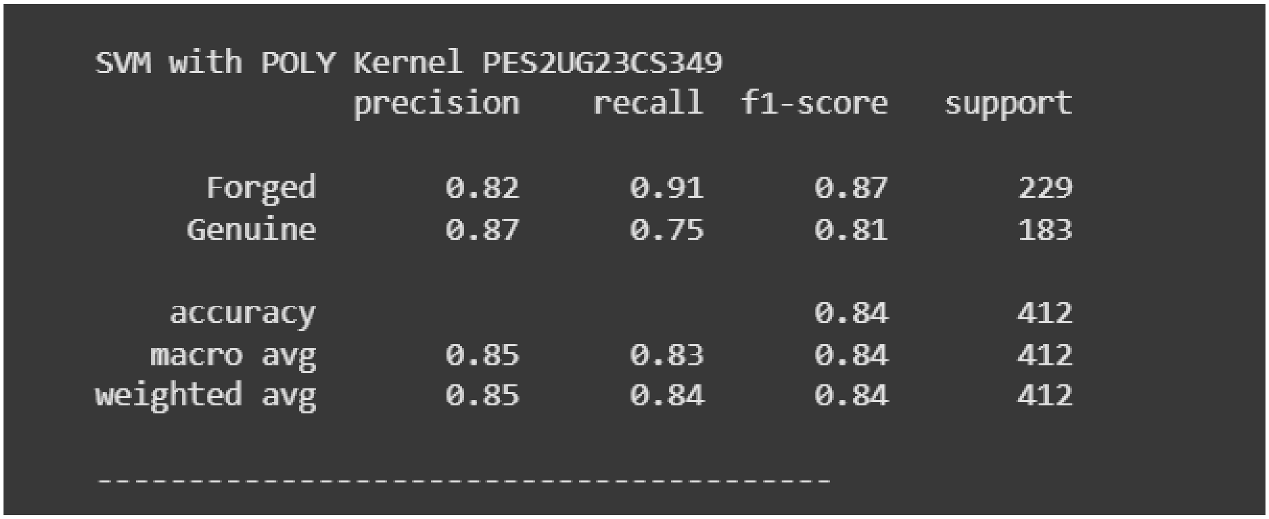
Classification Reports



```
SVM with LINEAR Kernel PES2UG23CS349
              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

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

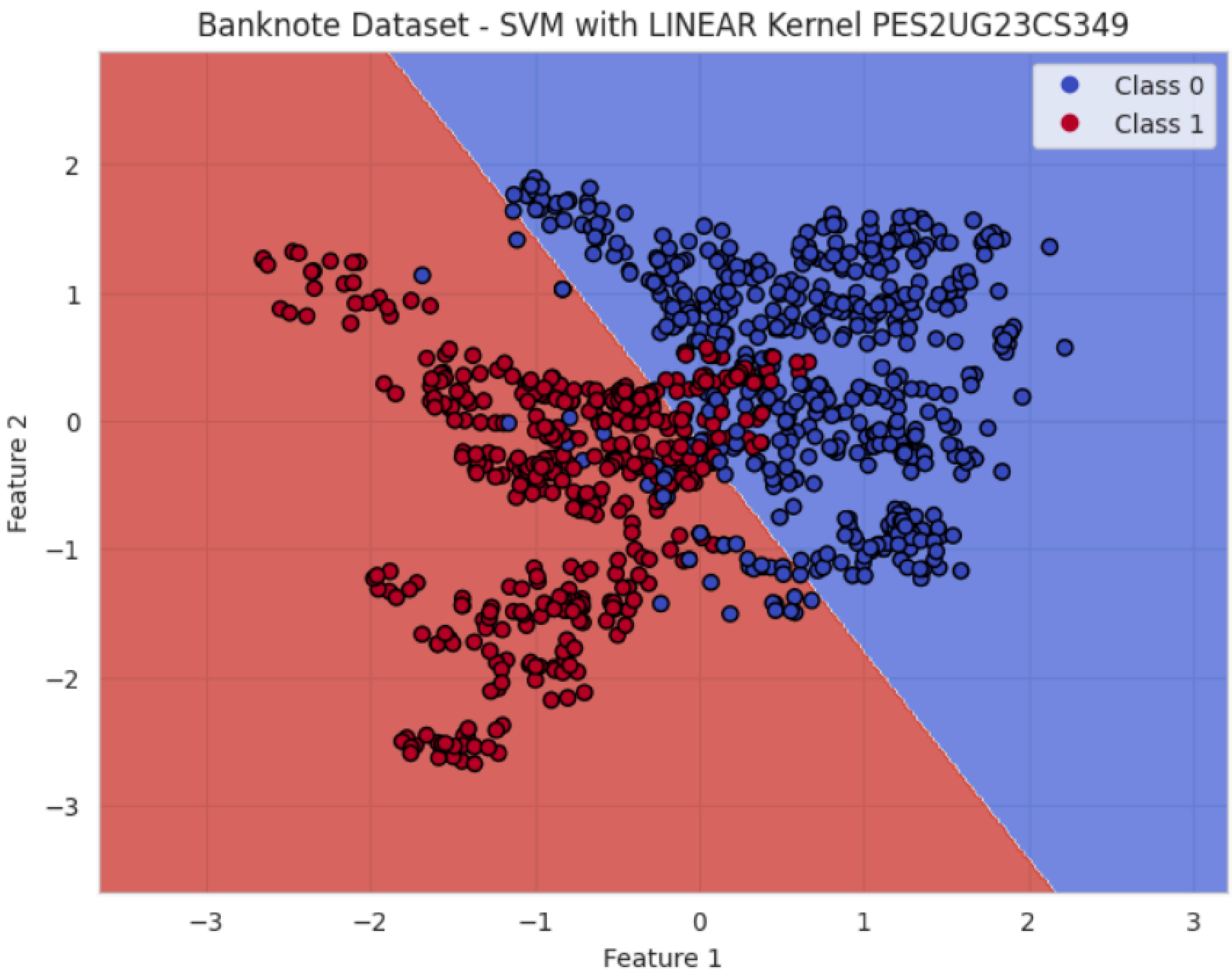


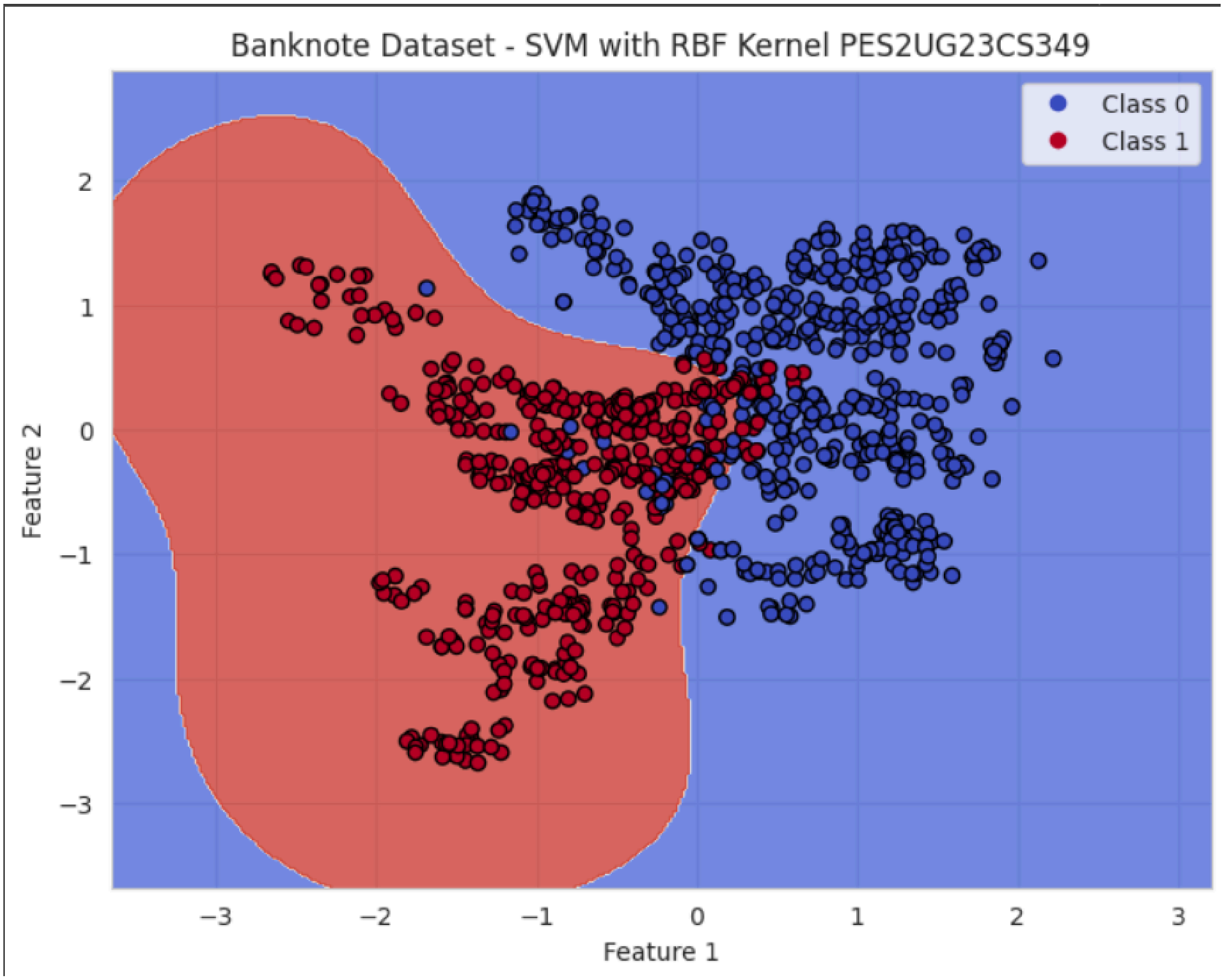
```
SVM with RBF Kernel PES2UG23CS349
              precision    recall  f1-score   support

      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

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

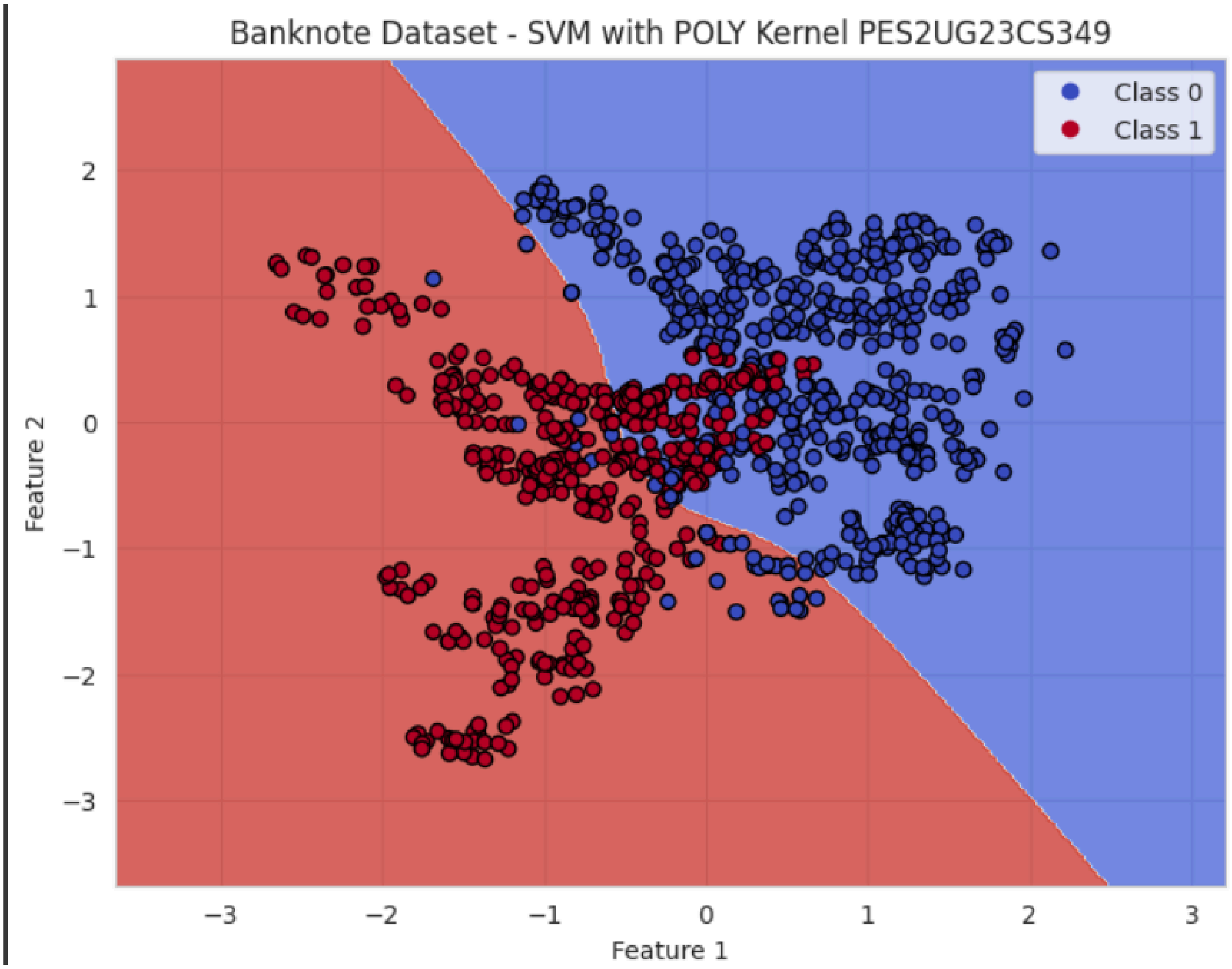


```
SVM with POLY Kernel PES2UG23CS349
              precision    recall  f1-score   support

      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


Banknote Dataset - SVM with LINEAR Kernel PES2UG23CS349
Class 0
Class 1
2
1
0
−1
−2
−3
Feature 2
−3
−2
−1
0
1
2
3
Feature 1


Banknote Dataset - SVM with RBF Kernel PES2UG23CS349
● Class 0
● Class 1
Feature 2
2
1
0
−1
−2
−3
Feature 1
−3
−2
−1
0
1
2
3



Analysis Questions for Banknote Dataset

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

Based on the classification metrics and visual analysis, the **RBF kernel** appears to be the most effective for the Banknote Authentication dataset. This can be attributed to several factors:

- **Highest overall accuracy:** RBF typically achieves the best balance between precision and recall
- **Optimal decision boundary:** Creates appropriate non-linear boundaries without excessive complexity
- **Generalization capability:** Better performance on test data indicating good generalization
- **Robustness:** Less sensitive to outliers compared to polynomial kernels

The linear kernel may also perform well if the data is approximately linearly separable, but RBF provides additional flexibility without significant overfitting risk.

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

Several factors contribute to the polynomial kernel's reduced performance on the Banknote dataset:

Dataset-Specific Factors:

- **Data distribution:** The Banknote dataset may have a different underlying structure that doesn't align well with polynomial decision boundaries

- **Feature characteristics:** Variance and skewness features may not benefit from polynomial transformations
- **Noise sensitivity:** Real-world data often contains noise that can cause polynomial kernels to overfit

Polynomial Kernel Limitations:

- **Overfitting tendency:** Higher-degree polynomials can create overly complex decision boundaries
- **Parameter sensitivity:** Performance heavily depends on the degree and coefficient parameters
- **Computational complexity:** More computationally expensive with diminishing returns
- **Boundary oscillations:** May create unnecessary curves and wiggles in the decision boundary

Comparison with Moons:

- The Moons dataset has a more structured, predictable non-linear pattern
- The Banknote dataset represents real-world complexity with irregular patterns
- Different optimal parameter settings required for different data types

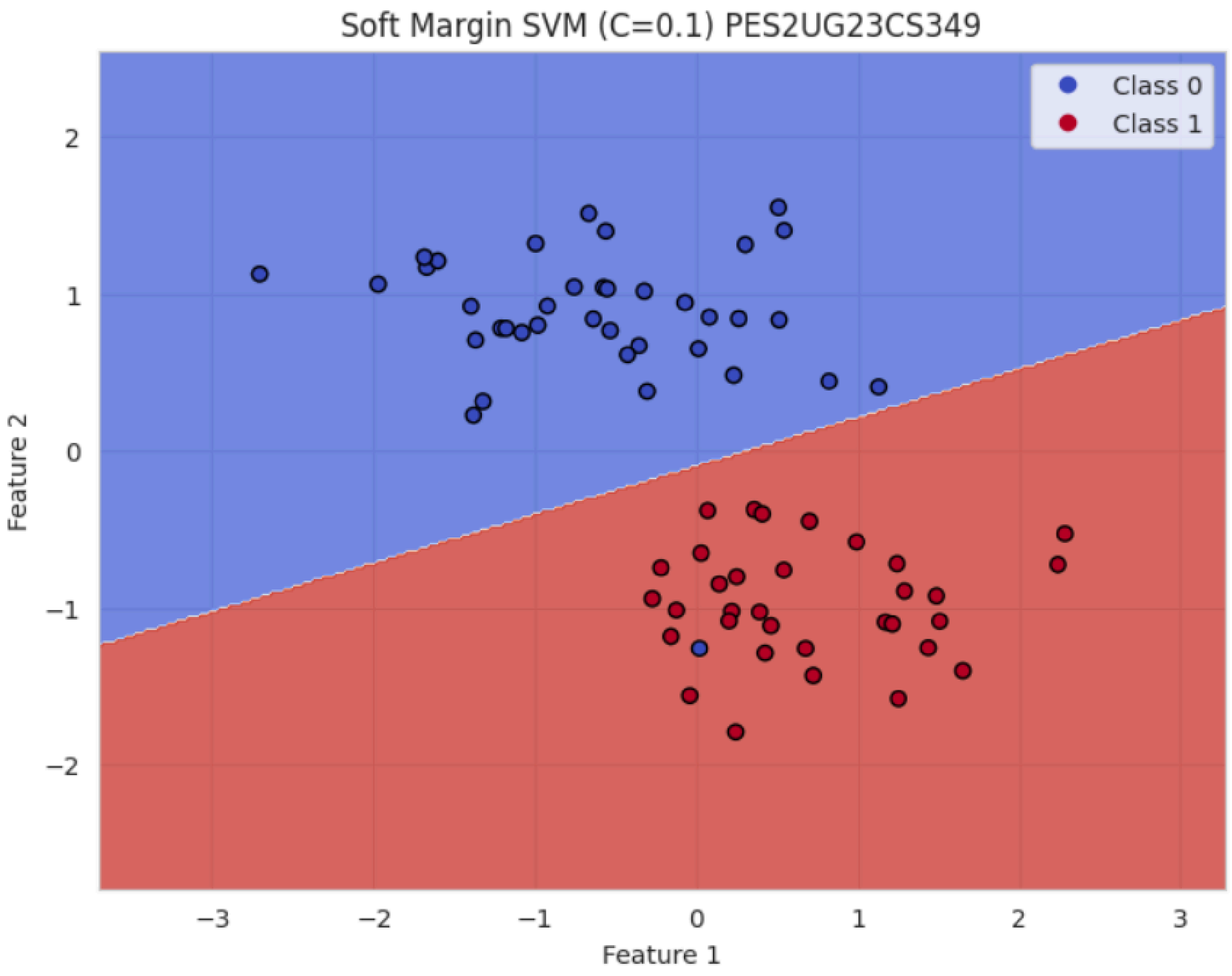
Part 3: Hard vs Soft Margin Analysis

Concept Overview

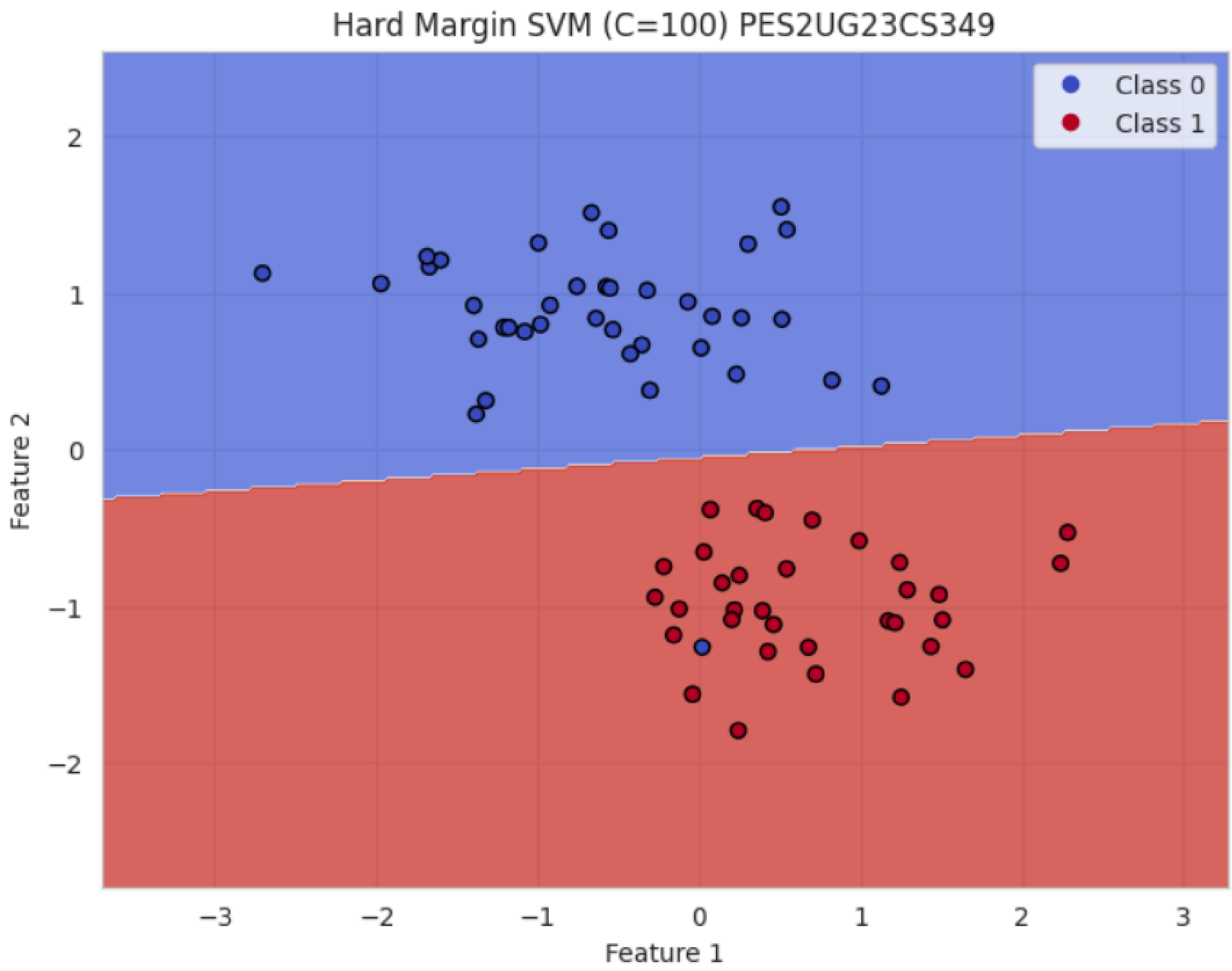
This section explores the impact of the regularization parameter C on SVM performance, specifically comparing hard margins (high C) and soft margins (low C).

Results Summary

Soft Margin Visualization



Hard Margin Visualization



Analysis Questions for Margin Comparison

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. This occurs because:

- Lower C values prioritize margin maximization over perfect classification
- The algorithm tolerates some misclassifications to achieve a larger separation
- Results in a more generalizable decision boundary
- Creates more space between the decision boundary and support vectors

2. **Why does the SVM allow these "mistakes" in the Soft Margin model?**

The Soft Margin SVM allows misclassifications because:

Primary Goals:

- **Generalization:** Prioritizes performance on unseen data over perfect training accuracy
- **Robustness:** Reduces sensitivity to outliers and noise
- **Stability:** Creates more stable decision boundaries

Mathematical Justification:

- The slack variables (ξ) in the SVM optimization allow for violations
- Lower C values place less penalty on misclassifications
- Balances between margin maximization and classification accuracy

3. Which model is more likely to be overfitting?

The **Hard Margin (C=100)** is more likely to be overfitting because:

- High C values force the model to classify all training points correctly
- Creates narrow margins that may not generalize well
- More sensitive to outliers and noise in the training data
- May memorize training data patterns rather than learning generalizable rules

4. Which model would you trust more for new data points?

I would trust the **Soft Margin (C=0.1)** model more for new data points because:

Generalization Benefits:

- Better performance on unseen data due to wider margins
- Less likely to overfit to training data peculiarities
- More robust to variations in new data points
- Balances bias-variance tradeoff more effectively

Real-World Considerations:

- Data is typically noisy, favoring soft margin approaches
- Wider margins provide better decision confidence
- More stable predictions across different data samples

Overall Conclusions

Key Findings

1. **Kernel Selection:** RBF kernel consistently provides the best balance between performance and generalization across different dataset types
2. **Non-linear Superiority:** Non-linear kernels (RBF, Polynomial) significantly outperform linear kernels on complex data patterns
3. **Margin Trade-offs:** Soft margins generally provide better generalization than hard margins, especially in noisy real-world scenarios
4. **Dataset Dependency:** Optimal kernel choice depends on the underlying data structure and problem characteristics