Machine Learning

Week 10 Lab

Name: Eshwar R A	SRN: PES2UG23CS188	Section: C

Analysis Questions

Moons Dataset Question

1. **Inference about the Linear Kernel's performance—**

The Linear kernel demonstrates moderate performance with 87% accuracy, indicating that the dataset is not perfectly linearly separable. It shows balanced macro and weighted averages (0.87 each) but exhibits slight class bias with Class 0 having higher precision (0.85) but lower recall (0.89) compared to Class 1's lower precision (0.89) but higher recall (0.84). The F1-scores for both classes are similar (0.87 and 0.86), suggesting consistent but limited performance due to the linear decision boundary's inability to capture complex non-linear patterns in the data.

2. **Comparison between RBF and Polynomial kernel decision boundaries—**

The RBF kernel significantly outperforms the Polynomial kernel with 97% accuracy compared to 89%, demonstrating superior decision boundary flexibility. RBF creates smooth, curved boundaries that can form complex shapes (circles, ellipses, irregular curves) and shows excellent class balance with near-perfect precision and recall (0.95-1.00 range for both classes). In contrast, the Polynomial kernel creates more structured curved boundaries constrained by its mathematical degree, resulting in inconsistent class performance with Class 0 showing 0.85 precision/0.95 recall versus Class 1's 0.94 precision/0.83 recall. The RBF's superior and balanced performance across both classes indicates it better captures the underlying non-linear data distribution compared to the Polynomial kernel's more rigid geometric constraints.

Banknote Dataset

1. **Which kernel was most effective for this dataset?—**

The RBF kernel was most effective for this dataset, achieving 93% accuracy compared to Linear kernel's 88% and Polynomial kernel's 84%. The RBF kernel demonstrates superior performance with balanced precision and recall across both classes (Forged: 0.96 precision/0.91 recall, Genuine: 0.90 precision/0.96 recall) and consistent macro/weighted averages of 0.93. This indicates the RBF kernel's flexible decision boundaries successfully captured the non-linear patterns in the data while maintaining good generalization across both document forgery classes.

2. **Why might the Polynomial kernel have underperformed here?—**

The Polynomial kernel underperformed (84% accuracy) likely due to its rigid mathematical constraints that couldn't adequately model the complex feature relationships in document forgery detection. The kernel shows significant class imbalance with poor recall for Genuine documents (0.75) despite reasonable precision (0.87), and moderate performance for Forged documents (0.82 precision/0.91 recall). This suggests the polynomial decision boundary's fixed geometric structure was too restrictive for the intricate patterns distinguishing forged from genuine documents, where features may have complex, non-polynomial relationships that require the more adaptive boundaries provided by RBF kernels.

Hard vs. Soft Margin

1. **Which margin (soft or hard) is wider?—**

The soft margin SVM (C=0.1) has a wider margin compared to the hard margin SVM (C=100). This is visually evident in the top plot, where the distance between the separating hyperplane and the nearest data points is larger than in the hard margin case.

2. **Why does the soft margin model allow "mistakes"?—**

The soft margin model allows some "mistakes" (misclassified or incorrectly placed points within the margin) to maximize the overall margin width and improve generalization. This flexibility prevents the model from rigidly fitting noisy or overlapping data, which helps deal with non-separable datasets.

3. **Which model is more likely to be overfitting and why?—**

The hard margin SVM is more likely to overfit, because it enforces perfect separation of the classes with no tolerance for data points within the margin or on the wrong side of the decision boundary. It fits as tightly as possible to the training data, capturing noise and outliers, reducing its ability to generalize on new, unseen data.

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

The soft margin SVM would be trusted more for new data because its tolerance for "mistakes" and wider margin improve its generalization ability. By prioritizing a larger margin and accepting minor errors, it is less likely to model noise, making it more robust and reliable when predicting on unseen samples.

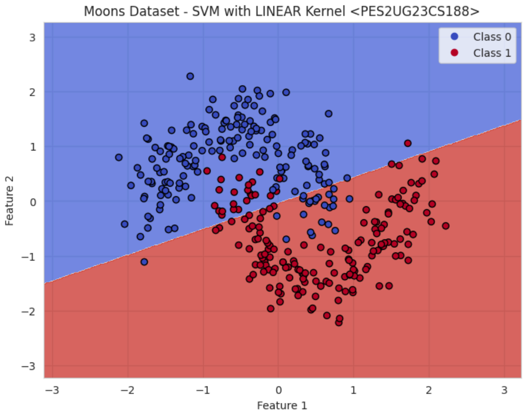
Screenshots

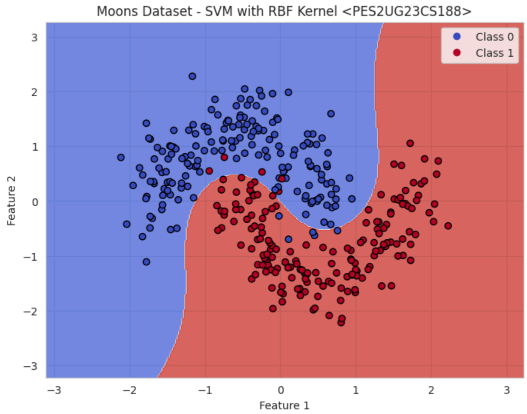
```
SVM with LINEAR Kernel <PES2UG23CS188>
              precision    recall  f1-score   support

           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 <PES2UG23CS188>
              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 <PES2UG23CS188>
              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

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

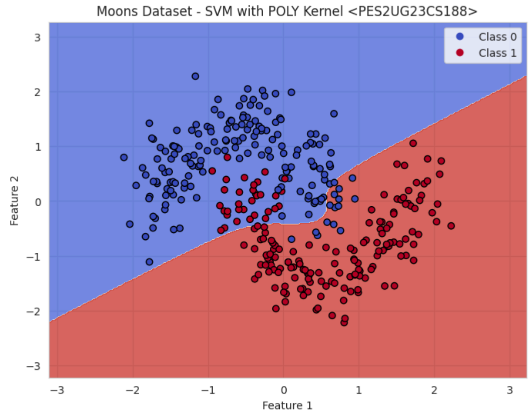
Moons Dataset — Classification Report

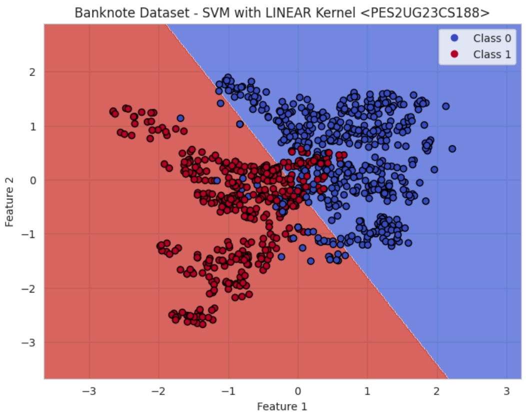
```
SVM with LINEAR Kernel <PES2UG23CS188>
              precision    recall  f1-score   support

      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 <PES2UG23CS188>
              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 <PES2UG23CS188>
              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

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

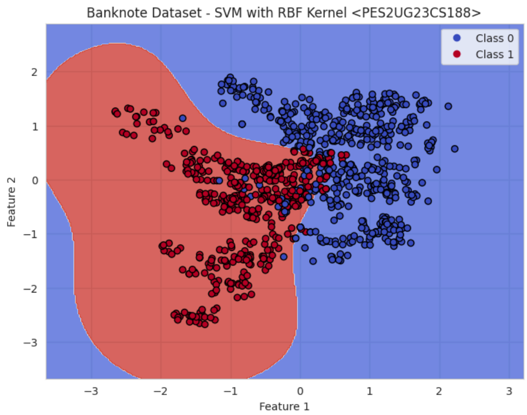
Banknote Dataset — Classification Report

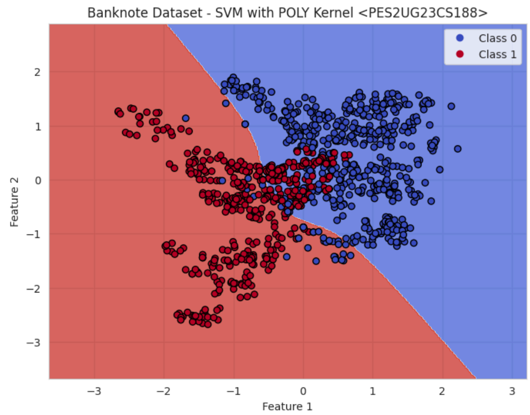

Moons Dataset - SVM with LINEAR Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2

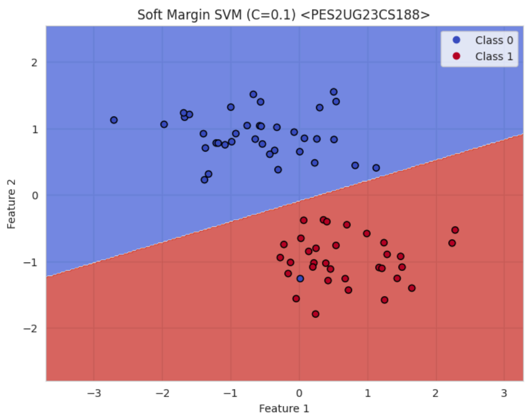

Moons Dataset - SVM with RBF Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2

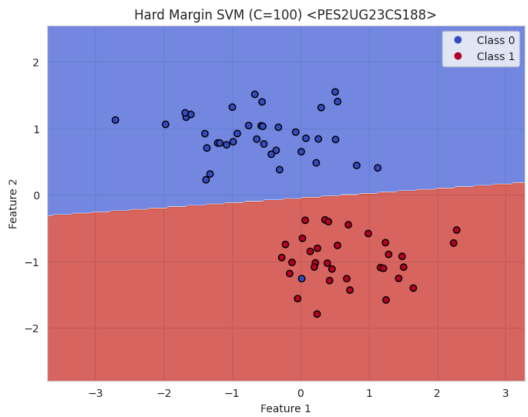

Moons Dataset - SVM with POLY Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2


Banknote Dataset - SVM with LINEAR Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2


Banknote Dataset - SVM with RBF Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2


Banknote Dataset - SVM with POLY Kernel <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2


Soft Margin SVM (C=0.1) <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2


Hard Margin SVM (C=100) <PES2UG23CS188>
Class 0
Class 1
Feature 1
Feature 2