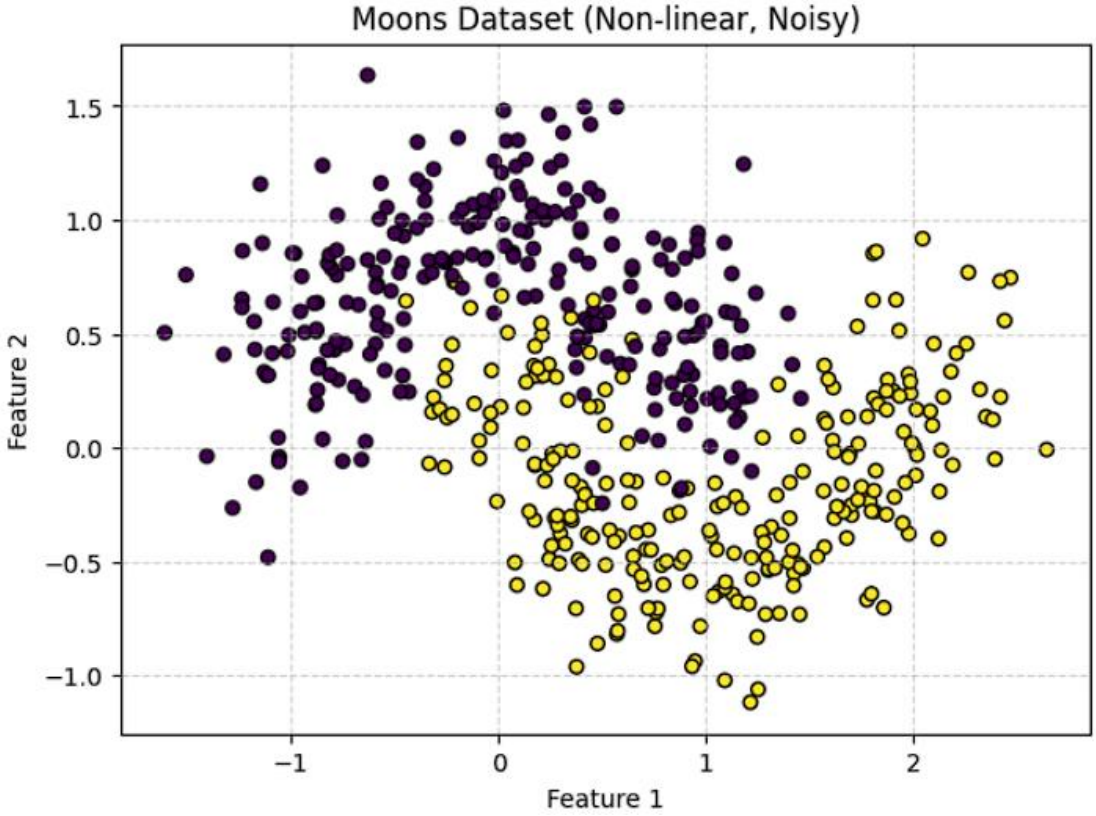
Support Vector Machine (SVM) Implementation

[Student Name: Anish Kumar Student ID: 23/CS/057 Course: BTech CSE]

1. Analysis of Exploratory Data Analysis (EDA)

Moons Dataset EDA



The Exploratory Data Analysis (EDA) involved generating a non-linear "moons" dataset consisting of 500 samples. This dataset has two features and two target classes. A scatter plot was created to visualize the data, color-coded by class.

- **Class Separation:** The plot clearly shows that the data is **non-linearly separable**. The two classes form distinct crescent shapes that are intertwined. A simple straight line cannot effectively separate these two groups.
- **Ease of Distinction:** While the clusters are distinct, their non-linear relationship makes them impossible to distinguish with a basic linear classifier. This dataset is specifically designed to test a model's ability to find a complex, non-linear decision boundary.

2. Model Classification Results

Analysis of "Accuracy vs. Kernel"

- **Best Performance:** The **RBF kernel** provided the best performance. The default RBF model achieved 94.67% accuracy, which was significantly higher than the other kernels. This is because the RBF kernel is highly flexible and capable of creating the complex, non-linear decision boundary required to separate the "moons" dataset.
- **Sub-optimal Kernels:**
    - **Linear Kernel (High Bias):** A linear kernel can only create a straight line. As seen in the decision boundary plot, this is not sufficient to separate the crescent shapes, resulting in a low accuracy of 85.33%. This model is **underfitting** the data.
    - **Polynomial Kernel:** The polynomial kernel (degree 3) performed better than the linear kernel (87.33%) but was still inferior to the RBF kernel. This suggests that while it can create a non-linear boundary, its specific shape was not as well-suited to this data as the RBF kernel's.

3. Hyperparameter Tuning (GridSearchCV)

Final Accuracy (k=3)

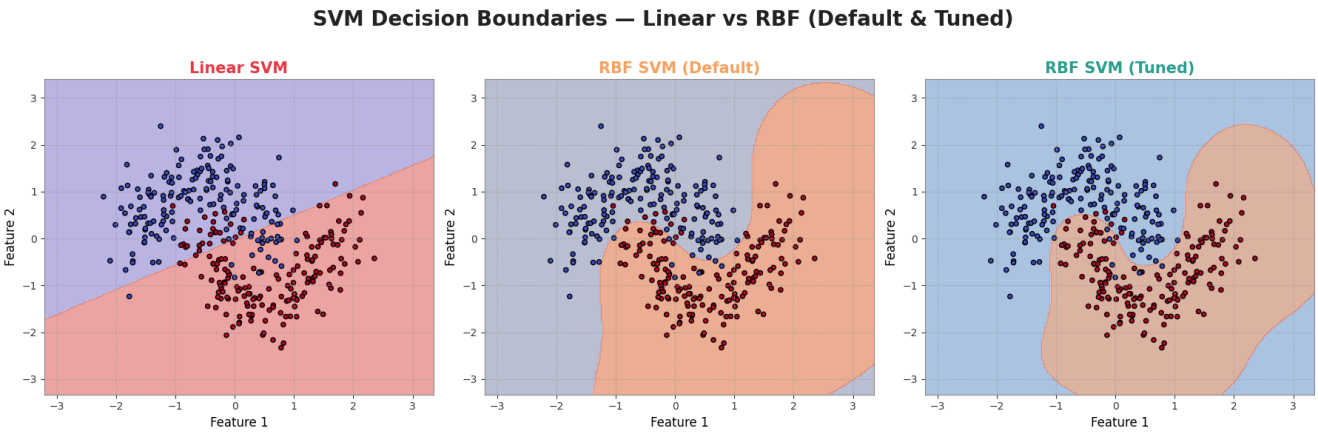
The final, optimized classification accuracy was achieved by using GridSearchCV to tune the RBF kernel's C and gamma hyperparameters.

Accuracy: 97.00%

Analysis of GridSearchCV Results

The GridSearchCV systematically tested 16 combinations of C ([0.1, 1, 10, 100]) and gamma ([0.1, 1, 10, 100]) to find the best model.

- The search confirmed that the best parameters were **C=1** and **gamma=1**.
- This tuned model achieved an accuracy of 97.00% on the validation set, a notable improvement from the default RBF kernel's 94.67%. This indicates that the default gamma was slightly too high (creating a boundary that was a bit too complex), and gamma=1 provided a better-generalized decision boundary.



4. Conclusion and Key Learnings

Summary

- **Linear Kernel Accuracy:** 85.33%
- **RBF Kernel (Default) Accuracy:** 94.67%
- **Polynomial Kernel Accuracy:** 87.33%
- **Tuned RBF Kernel (C=1, gamma=1) Accuracy: 97.00%**

This lab involved applying and tuning Support Vector Machine (SVM) classifiers and provided several key insights:

1. **EDA is Crucial:** The EDA plot was essential. It immediately showed that the data was non-linear, which guided the model selection process. It made it clear from the start that a linear kernel would fail and a non-linear kernel like RBF or Polynomial would be necessary.
2. **Kernel Implementation:** The core of the lab was understanding the "kernel trick." The experiment showed how different kernels (Linear, Poly, RBF) create vastly different decision boundaries and how choosing the right kernel for the data's shape is the most critical step.
3. **Hyperparameters are Critical:** The experiments showed that C (the regularization parameter) and gamma (the kernel coefficient) are critical for performance. The default RBF (94.67%) was good, but GridSearchCV found an optimal combination (C=1, gamma=1) that improved accuracy to 97.00% by finding a better balance between bias and variance.
4. **Bias-Variance Trade-off:** This lab was a practical demonstration of the bias-variance trade-off in SVMs.

- The **Linear kernel** has **high bias** (it's too simple for this data) and **underfits**.
- An RBF kernel with a very **high gamma** would have **high variance** (overfitting) by creating a boundary that is too complex and specific to the training points.
- GridSearchCV helps find the optimal C and gamma to balance this trade-off, creating a model that generalizes well to new data.