Support Vector Machines for Email Classification

Objective:

This assignment involves using Support Vector Machines (SVM) to classify emails as spam or not spam. The goal is to build, train, and evaluate SVM models using different kernel functions and hyperparameter settings, and report the classification performance.

Dataset:

The dataset used is the UCI Spambase dataset, available at:

https://archive.ics.uci.edu/ml/datasets/Spambase

It contains 4601 email instances, each represented by 57 features such as the frequency of specific keywords, average length of capital letters, and other characteristics relevant to spam detection.

You are required to:

- Load the dataset and preprocess the features as needed.
- Randomly split the dataset into 75% training and 25% test data.

Tasks:

1. Preprocessing:

- Normalize or standardize the feature values.
- Ensure the label column (spam or not) is correctly encoded.

2. Model Building:

- Use the sklearn.svm.SVC class to implement the SVM.

- Train models using different kernel functions:

- Linear Kernel

- Polynomial Kernel (e.g., degree 2 or 3)

- RBF (Radial Basis Function) Kernel

3. Hyperparameter Tuning:

- Experiment with different values for C (regularization parameter).

- For RBF kernel: try different gamma values.

- For polynomial kernel: vary the degree.

4. Evaluation:

- Calculate and report classification accuracy for each kernel and setting.

- Use a confusion matrix and possibly other metrics like precision, recall, or F1-score for deeper insights.

5. Analysis:

- Compare the performance of the different kernels.

- Discuss which kernel and hyperparameter combination works best for this task and why.

Deliverables:

- Python code (.ipynb or .py) implementing the above tasks.

- Accuracy results and brief analysis.

- This report in .pdf format summarizing your approach.

Tools and Libraries:

- Python

- scikit-learn

- pandas, NumPy, matplotlib (optional for visualization)