DA5401 A7: Multi-Class Model Selection using ROC and Precision-Recall Curves

Objective: This assignment challenges you to apply and interpret **Receiver Operating Characteristic (ROC)** curves and **Precision-Recall Curves (PRC)** for model selection in a multi-class, complex classification environment. You will compare a diverse set of classifiers, including those known to perform poorly, requiring careful analysis of the curves rather than relying solely on simple accuracy.

---

1. Problem Statement

You are a machine learning scientist tasked with classifying land cover types using satellite image data. You have the **UCI Landsat Satellite dataset**, a multi-class problem (6 classes) known to be non-trivial due to high feature dimensionality and potential class overlap.

Your primary goal is to perform model selection by comparing various classifier types. You must use **ROC** and **PRC** analysis, adapted for the multi-class setting (e.g., using one-vs-rest averaging), to determine the best-performing and worst-performing models, paying special attention to the performance across different decision thresholds.

You will submit a Jupyter Notebook with your complete code, visualizations, and a plausible story that explains your findings. The notebook should be well-commented, reproducible, and easy to follow.

Dataset:

- **Landsat Satellite Dataset:** A classic multi-class problem with six primary classes. Ignore the class label that says “all types present”.
    - **Citation:** Blake, C. and Merz, C.J. (1998). *UCI Repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science.
    - **Download Link (UCI ML Repository):** [UCI Machine Learning Repository - Satimage]

Model Classes for Comparison (Minimum Five):

Model Class	Python Library/Reference	Performance Expectation

1. K-Nearest Neighbors (KNN)	sklearn.neighbors.KNeighborsClassifier	Moderate/Good
2. Decision Tree Classifier	sklearn.tree.DecisionTreeClassifier	Moderate
3. Dummy Classifier (Prior)	sklearn.dummy.DummyClassifier	Baseline (Likely AUC < 0.5 for minority classes)
4. Logistic Regression	sklearn.linear_model.LogisticRegression	Good/Baseline (Linear model benchmark)
5. Naive Bayes (Gaussian)	sklearn.naive_bayes.GaussianNB	Poor/Varies (Simple assumptions fail in complex data)
6. Support Vector Machine (SVC)	sklearn.svm.SVC	Good (Requires probability=True for ROC/PRC)

The following model classes should be used, spanning different performance levels and ensuring at least one model is likely to perform worse than random chance (AUC < 0.5). The expected performance of a few model classes is also shown at https://archive.ics.uci.edu/dataset/146/statlog+landsat+satellite.

---

2. Tasks

Part A: Data Preparation and Baseline [5]

1. **Load and Prepare Data:** Load the Landsat dataset. **Standardize** the features (X).
2. **Train/Test Split:** Split the data into training and testing sets.
3. **Train All Models:** Train one instance of each of the six specified model classes on the training data. *Note: For the **Dummy Classifier**, use the 'prior' strategy. For **SVC**, ensure you set the parameter* `probability=True` *to enable probability predictions for ROC/PRC analysis.*
4. **Baseline Evaluation:** Calculate the simple **Overall Accuracy** and **Weighted F1-Score** for all six models on the test set. Observe which models perform poorly.

Part B: ROC Analysis for Model Selection [20]

1. **Multi-Class ROC Calculation [3]:** Explain how the **One-vs-Rest (OvR)** approach is used to generate ROC curves and calculate the Area Under the Curve (AUC) in a multi-class setting.
2. **Plotting ROC [12]:** Generate a single plot displaying the OvR ROC curves for all six models, averaging the False Positive Rate (FPR) and True Positive Rate (TPR) across all six classes (Macro-average or Weighted-average ROC).
3. **ROC Interpretation [5]:**
   - Identify the model with the highest **Macro-averaged AUC**.
   - Identify the model with AUC < 0.5. Explain what AUC < 0.5 implies conceptually and why a model might exhibit this performance.

Part C: Precision-Recall Curve (PRC) Analysis [20]

1. **PRC Calculation [3]:** Explain why the **Precision-Recall Curve (PRC)** is a more suitable metric than ROC when dealing with highly imbalanced classes (addressing the conceptual importance even if the imbalance in this dataset is moderate).
2. **Plotting PRC [12]:** Generate a single plot displaying the OvR PRC curves for all six models, averaging the Precision and Recall across all six classes (Macro-average or Weighted-average PRC).
3. **PRC Interpretation [5]:**
   - Identify the model with the highest **Average Precision (AP)** across the classes.
   - Analyze the behavior of the worst-performing model's PRC. Why does the curve drop sharply as Recall increases for poor models?

Part D: Final Recommendation [5]

1. **Synthesis:** Compare the model rankings derived from the initial F1-Score, the ROC-AUC, and the PRC-AP. Do the rankings align? If not, explain the specific trade-offs (e.g., how a model with a high ROC-AUC might still have a poor PRC-AP).
2. **Recommendation:** Based on the comprehensive analysis of all curves, recommend the best model for this classification task, justifying your choice based on performance across different thresholds and the desired balance between precision and recall.

Brownie Points Task [5 points]

1. Additionally, experiment with RandomForest and XGBoost classifiers.
2. Identify and experiment with another model class of your choice whose AUC < 0.5.

3. Submission Guidelines

- The assignment is due in **2 weeks**.
- Submit a single Jupyter Notebook with all your code, visualizations, and answers to the conceptual questions in markdown cells.
- Ensure all code is clean, readable, and reproducible.

Evaluation Criteria:

- Correct training and evaluation of all six models, particularly the SVC with probability estimates.
- Accurate implementation of multi-class ROC and PRC analysis.
- High-quality, clearly labeled plots for both ROC and PRC.
- Demonstrated understanding of the theoretical concepts of AUC, Average Precision, and the interpretation of models that perform worse than random.

Good luck!