Machine Learning Assignment: Rice Species Classification

Total Marks: 50

Due Date: [To be announced]

Submission Format: Jupyter notebook (.ipynb) + PDF report

Overview

You will build and evaluate a binary classifier to distinguish between two visually similar rice varieties: Arborio and Jasmine. This assignment emphasizes independent learning, proper methodology, and critical analysis of your results.

Dataset

You will work with the Rice Image Dataset available on UC Irvine's Machine Learning Repository or Kaggle. The dataset contains grain images with extracted features including:

- Morphological measurements (area, perimeter, major/minor axis length)
- Shape descriptors (eccentricity, convex area, extent)
- Geometric ratios (aspect ratio, roundness)

Filter the dataset to include only Arborio and Jasmine varieties for binary classification.

Learning Objectives

1. Implement machine learning pipelines from scratch using appropriate libraries
2. Perform exploratory data analysis and feature engineering independently
3. Apply multiple classification algorithms and compare their performance
4. Evaluate models using appropriate metrics beyond simple accuracy
5. Document your methodology and justify your decisions

Tasks and Grading Breakdown

Part 1: Data Exploration and Preprocessing (12 marks)

1.1 Data Loading and Initial Analysis (4 marks)

- Load the dataset and filter for Arborio and Jasmine varieties
- Report dataset statistics: number of samples per class, feature distributions
- Identify and handle any missing values or outliers with justification

1.2 Exploratory Data Analysis (5 marks)

- Create at least 3 meaningful visualizations that reveal patterns or differences between the two rice varieties
- Analyze feature correlations and identify potentially redundant features
- Discuss which features appear most discriminative and why

1.3 Data Splitting and Preprocessing (3 marks)

- Split data into training (70%), validation (15%), and test (15%) sets
- Apply appropriate preprocessing (scaling/normalization) with justification
- Ensure reproducibility by setting random seeds

Part 2: Model Implementation (18 marks)

2.1 Baseline Model (4 marks)

- Implement a simple baseline classifier (e.g., Logistic Regression or Decision Tree)
- Train on the training set and evaluate on the validation set
- Report performance metrics (specified in Part 3)

2.2 Advanced Models (10 marks)

- Implement at least TWO additional classification algorithms from different families:
  - Examples: Random Forest, Support Vector Machine, k-Nearest Neighbors, Naive Bayes, Gradient Boosting
- For each model, explain why you selected it and what you expect it to handle well
- Perform hyperparameter tuning using the validation set (document your search process)

2.3 Model Training Documentation (4 marks)

- Clearly document all hyperparameters tested and final selections
- Show training curves or validation performance across different configurations
- Explain your decision-making process for final model selection

Part 3: Evaluation and Metrics (12 marks)

3.1 Comprehensive Metrics (6 marks)

For each model, calculate and report:

- Accuracy
- Precision
- Recall
- Confusion Matrix

Create a comparison table summarizing all models' performance.

3.2 Metric Analysis (4 marks)

- Explain which metric is most important for this rice classification task and why

- Discuss any trade-offs observed (e.g., precision vs. recall)
- Analyze misclassifications: what patterns do you notice in the errors?

3.3 Final Model Evaluation (2 marks)

- Evaluate your best-performing model on the test set
- Compare test performance to validation performance
- Discuss whether your model generalizes well or shows signs of overfitting

Part 4: Critical Analysis and Report (8 marks)

4.1 Methodology Justification (3 marks)

- Defend your choice of preprocessing techniques
- Explain why your selected models were appropriate for this problem
- Discuss any challenges encountered and how you addressed them

4.2 Results Interpretation (3 marks)

- Which model performed best and why do you think this occurred?
- What do the results tell you about the separability of these rice varieties?
- Suggest at least two ways the model could be improved with additional work

4.3 Documentation Quality (2 marks)

- Clear, well-commented code
- Professional report structure with proper sections
- Proper citations for any external resources consulted

Submission Requirements

1. Jupyter Notebook: Contains all code, visualizations, and inline commentary
2. PDF Report: 4-6 pages summarizing your methodology, results, and analysis (exclude code)
3. README file: Brief instructions on how to run your notebook

Academic Integrity Policy

This is an individual assignment. You must complete all work independently.

Prohibited:

- Using AI tools (ChatGPT, GitHub Copilot, etc.) to generate code or analysis
- Copying code from classmates or online sources without proper attribution
- Submitting work that is not substantially your own

Permitted:

- Consulting official documentation (scikit-learn, pandas, matplotlib, etc.)
- Reading textbook chapters and course materials

- Asking clarifying questions about assignment requirements (not implementation help)
- Using standard library functions as documented

Any violation will result in a grade of zero and referral to the academic integrity office.

Helpful Resources (Official Documentation Only)

- scikit-learn User Guide: https://scikit-learn.org/stable/user_guide.html
- pandas Documentation: https://pandas.pydata.org/docs/
- matplotlib Tutorials: https://matplotlib.org/stable/tutorials/index.html
- Course textbook chapters on classification and evaluation metrics