Iris Dataset Classification Using Machine Learning

1. Problem Overview and Motivation

The objective of this project is to automate plant species identification using the Iris dataset. By applying supervised machine learning techniques, flowers are classified into three species—Iris Setosa, Iris Versicolor, and Iris Virginica—based on morphological measurements. This project demonstrates pattern recognition in biological data and serves as a foundational classification task in machine learning.

2. Dataset Description and Preprocessing

The dataset is obtained from the UCI Machine Learning Repository and consists of 150 samples, equally distributed among three species. Each sample contains four numerical features: Sepal Length, Sepal Width, Petal Length, and Petal Width. An 80/20 train-test split was applied, resulting in 120 training samples and 30 testing samples.

3. Mathematical Formulation

Logistic Regression (Linear Model)

Logistic Regression models the probability of class membership using the sigmoid function:

$\sigma(z) = 1 / (1 + e^{-z})$

where z is the linear combination of inputs defined as:

$z = w^T x + b$

Here, x is the input feature vector, w represents the weight vector, and b is the bias term.

Random Forest (Non-Linear Model)

Random Forest is an ensemble learning technique that combines multiple decision trees. Each tree is trained on a bootstrap sample of the dataset, and only a random subset of features is considered at each split. The final class prediction is obtained using majority voting.

4. Loss Function and Training Process

To optimize classification performance, Cross-Entropy Loss is minimized:

$L = - \Sigma [y \log(p) + (1 - y) \log(1 - p)]$

Training involves forward propagation, loss computation, and iterative parameter updates until convergence.

5. Model Architecture and Justification

Logistic Regression Architecture

- Single-layer linear model
- Weighted sum of input features
- Sigmoid activation function

Random Forest Architecture

- Ensemble of decision trees
- Bagging for variance reduction
- Random feature selection

6. Evaluation Methodology and Results

Model performance was evaluated using accuracy and confusion matrix analysis. The achieved classification accuracy was 100%.

Confusion Matrix Results:
- 10 Iris Setosa samples correctly classified
- 9 Iris Versicolor samples correctly classified
- 11 Iris Virginica samples correctly classified

7. Limitations and Future Improvements

Although the model achieved excellent performance, the dataset is small and may not represent real-world variability. Future improvements include cross-validation, hyperparameter tuning, feature importance analysis, and evaluation on larger datasets.

Conclusion

This project successfully demonstrates an end-to-end machine learning classification pipeline using both linear and non-linear models. The results confirm the effectiveness of supervised learning techniques and provide a strong foundation for advanced machine learning applications.