Statistical Data Analysis: Iris Dataset and Decision Tree Analysis

Elif Solmaz - 220467249

December 2024

1 Introduction

This report presents an analysis of the Iris dataset using decision tree and boosted decision tree models. The Iris dataset is a classic benchmark in machine learning and statistics, containing measurements of three different iris species:
- Setosa, Versicolor, and Virginica.

Each sample includes four features:

1. Sepal Length
2. Sepal Width
3. Petal Length
4. Petal Width

The report focuses on:
- Characterizing the dataset features.
- Analyzing the performance of **Decision Tree (DT)** and **Boosted Decision Tree (BDT)** classifiers.
- Exploring the impact of tree depth and boosting on model accuracy.

2 Characterizing the Iris Dataset

The Iris dataset consists of 150 samples, with 50 samples per species. Below are the statistical summaries and visualizations for the features:

2.1 Feature Statistics

Features	**Mean(cm)**	**Standard Dev. (cm)**
Sepal length	5.84	0.83
Sepal width	3.06	0.43
Petal length	3.76	1.76
Petal width	1.20	0.76

2.2 Correlation Analysis

The correlation matrix reveals significant relationships between features:
- **Strong positive correlation** between Petal Length and Petal Width (0.96).
- **Positive correlation** between Sepal Length and Petal Length (0.87).
- **Moderate negative correlation** between Petal Length and Sepal Width (-0.43).

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1.0	-0.12	0.87	0.82
Sepal Width	-0.12	1.0	-0.43	-0.37
Petal Length	0.87	-0.43	1.0	0.96
Petal Width	0.82	-0.37	0.96	1.0

2.3 Sepal and Petal Analysis

The scatter plots below illustrate the distributions and separability of species based on different feature combinations, where Setosa(purple), Versicolor(green), Virginica(yellow) is presented:

Sepal Length vs Sepal Width

- Weak correlation and higher overlap among species.

Petal Length vs Petal Width

- Clear separation of **Setosa** from **Versicolor** and **Virginica**.

- Overlap between **Versicolor** and **Virginica**.

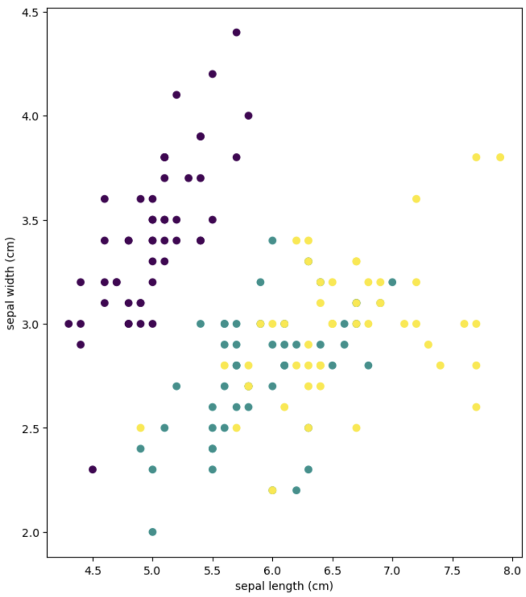


Figure 1:

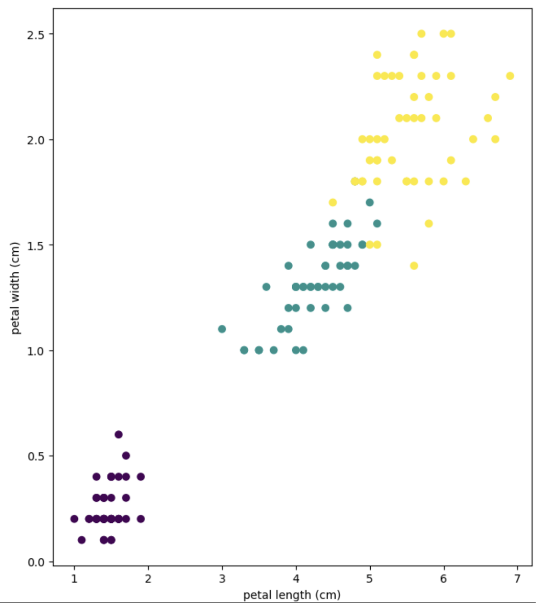


Figure 2:

1. **Figure 1: Sepal Length vs Sepal Width**

   - Shows clustering for Setosa, while Versicolor and Virginica exhibit overlap.

2. **Figure 2: Petal Length vs Petal Width**

   - Demonstrates strong linear separation, particularly for Setosa, highlighting petal dimensions as the most discriminative features.

3 Decision Tree-Based Analysis

3.1 Confusion Matrix Analysis:

The confusion matrix for test samples shows more off-diagonal entries due to:

1. **Overfitting:** The model learns the training data's specific patterns and noise.

2. **Generalization Gap:** The learned decision boundaries may not perfectly classify unseen data.

3.2 Tree Depth Analysis:

We evaluated decision trees with depths ranging from 1 to 4.

Tree Depth	Train Score	Test Score
1	0.6952	0.6000
2	0.9619	0.9111
3	0.9810	0.9810
4	1.0000	0.9778

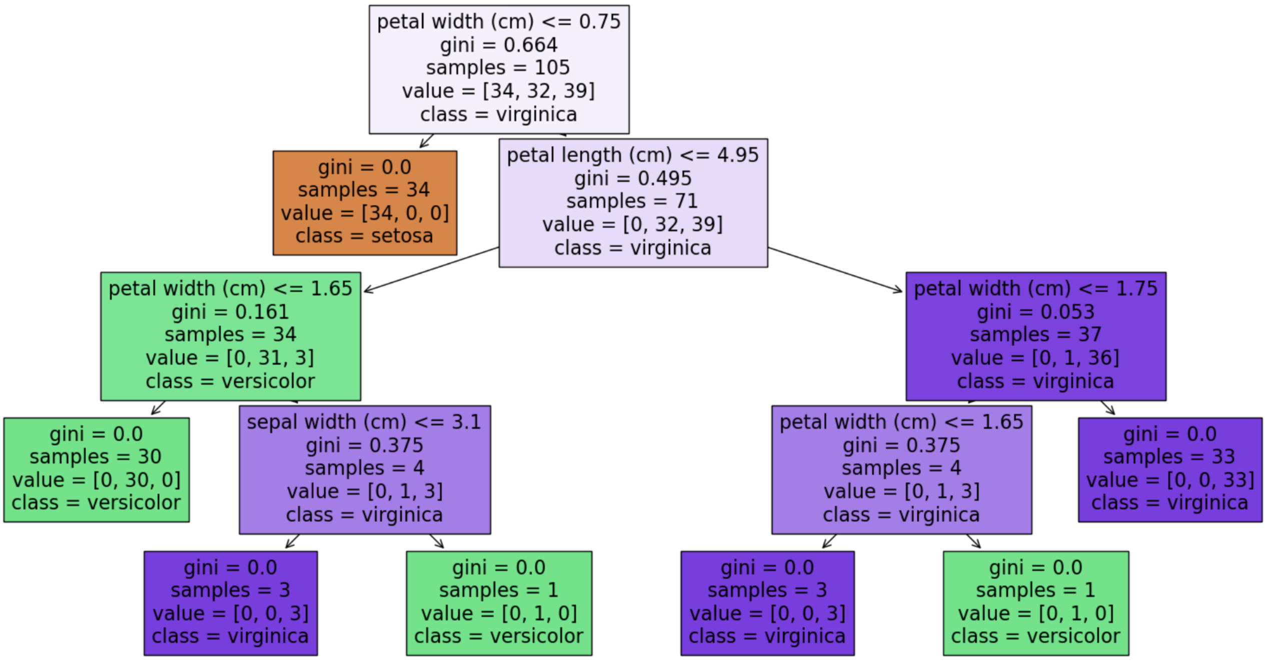
Key Observations:

- Depth 3 provides the best balance of accuracy and generalization.

- Depth 4 achieves perfect training accuracy but shows diminishing returns for test accuracy.

3.3 Visualization of Decision Trees

The diagram below shows the decision tree for depth 4:



3.3.1 Variable Selection:

The decision trees prioritize **Petal Length** and **Petal Width** as splitting criteria:

Depth 1: Uses only petal width

Depth 2-4: Petal Length and Width, with refined splits improving class separation.
This matches our observation that petal measurements provide clearer species separation.

3.3.2 Effective Cuts:

The most effective splits are on **Petal Length** and **Petal Width**, aligning with observations in scatter plots:

- Early splits separate Setosa clearly.

- Deeper splits refine boundaries between **Versicolor** and **Virginica.**

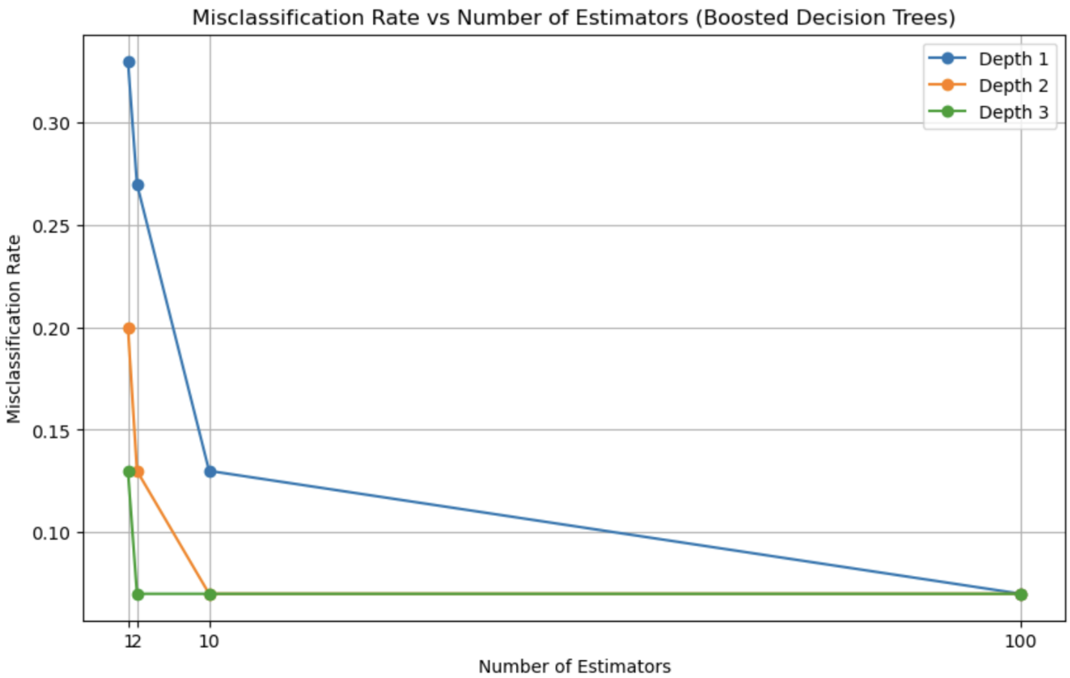
4 Boosted Decision Tree Analysis

4.1 Performance with 50% Train Split:

The following table shows misclassification rates for varying estimators and depths:

Estimators	Depth 1	Depth 2	Depth 3
1	0.33	0.20	0.13
2	0.27	0.13	0.07
10	0.13	0.07	0.07
100	0.07	0.07	0.07

Misclassification Trends The trend graph below illustrates how increasing the number of estimators improves performance by reducing misclassification rates:



Observations:

- Performance improves significantly with the first few estimators.

- Depth 3 achieves the best results, even with fewer estimators.

4.2 Performance with 80% Train Split

With a larger training set, the misclassification rates are more stable:

Estimators	Depth 1	Depth 2	Depth 3
1	0.30	0.17	0.10
2	0.23	0.10	0.07
10	0.10	0.03	0.03
100	0.03	0.03	0.03

4.2.1 Residual Misclassifications

Misclassifications persist due to:

Feature Overlap: Between **Versicolor** and **Virginica**.

Limited Data: Smaller test sets increase sensitivity to individual errors.

5 Conclusion

The analysis of the Iris dataset demonstrates the effectiveness of **Decision Trees** and **Boosted Decision Trees** for classification:

- **Decision Trees:** Achieve high accuracy with simple, interpretable models. Depth 3 strikes the optimal balance.

- **Boosted Trees:** Further improve accuracy by refining splits and correcting misclassifications.

Key Findings:

- **Petal measurements** are the most discriminative features.

- Boosted Decision Trees with **depth 3** and **10 estimators** achieve near-optimal performance.

Future work can explore alternative models (e.g., SVM or kNN) for comparative analysis.