

Lab 9: Building & Visualizing Decision Trees (Wine Dataset)

Tools

Scikit-learn, Matplotlib

Dataset

Wine Dataset (Scikit-learn built-in)

Link: https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-dataset

Lab Tasks

1. Load the Wine dataset using Scikit-learn's `load_wine()` function.
2. Split the dataset into training and testing sets (e.g., 70% training, 30% testing).
3. Train a Decision Tree Classifier with `max_depth=3`.
4. Visualize the trained tree using `plot_tree()`. Use `feature_names` and `class_names` for interpretability.
5. Evaluate the model's accuracy on both training and testing sets.
6. Train another Decision Tree with no depth restriction (`max_depth=None`).
7. Compare the training and testing accuracies of both models.
8. Save both tree visualizations and include them in your lab report.

Exercises

1. According to the root node of your first tree (`max_depth=3`):
 - Which feature was used for the first split?
 - What is the Gini impurity at that node?
2. When you removed the `max_depth` restriction:
 - What happened to the training and test accuracies?
 - Did your model overfit? Explain your reasoning.
3. Use `clf.feature_importances_` to get the importance of each feature.
 - Which features are the most important?
 - Do they match the features you see in your visualization?
4. Suggest two ways to reduce overfitting in a decision tree. (Hint: think about parameters like `min_samples_split`, `min_samples_leaf`, or pruning.)