Vietnam National University,
Ho Chi Minh City

University of Science

Faculty of Information Technology

Project 03: Decision Tree

CS14003 – Introduction to Artificial Intelligence

Ngo Nguyen The Khoa 23127065
Bui Minh Duy 23127040
Nguyen Le Ho Anh Khoa 23127211

April 5, 2025

Contents

1 Group Information

- **Subject:** Introduction to Artificial Intelligence.
- **Class:** 23CLC09.
- **Lecturer:** Bui Duy Dang, Le Nhut Nam.
- **Team members:**

No.	Fullname	Student ID	Email
1	Ngo Nguyen The Khoa	23127065	nntkhoa23@clc.fitus.edu.vn
2	Bui Minh Duy	23127040	bmduy23@clc.fitus.edu.vn
3	Nguyen Le Ho Anh Khoa	23127211	nlhakhoa23@clc.fitus.edu.vn

2 Project Information

- **Name:** Decision Tree Classifier using Scikit-learn.
- **Developing Environment:** Visual Studio Code (Windows, WSL).
- **Programming Language:** Python.
- **Libraries and Tools:**
  - **Libraries:**
    - **scikit-learn:** Machine learning library for training and evaluating decision tree models.
    - **pandas:** Data manipulation and analysis.
    - **numpy:** Numerical operations.
    - **matplotlib**, **seaborn:** Data visualization libraries.
    - **graphviz:** Visualization of decision trees.
  - **Tools:**
    - **Git**, **GitHub**: Source code version control.
    - **Visual Studio Code**: Code editor for Python, LaTeX.
- **Datasets:**
  - **Breast Cancer Wisconsin (Diagnostic)**
  - **Wine Quality**
  - **Car Evaluation**

3 Work Assignment Table

No.	Task Description	Assigned to	Rate
1	Prepare all three datasets (Breast Cancer, Wine Quality, and Additional) with proper preprocessing and stratified splits.	The Khoa	100%
2	Implement and train decision tree models for each dataset with different train/test splits.	Minh Duy	100%
3	Visualize decision trees using Graphviz.	Anh Khoa	100%
4	Evaluate classifiers with classification reports and confusion matrices.	Anh Khoa	100%
5	Analyze impact of tree depth on accuracy (80/20 split, varying max_depth values).	The Khoa	100%
6	Research and integrate additional dataset.	Minh Duy	100%
7	Conduct comparative analysis across the three datasets.	The Khoa	100%
8	Visualize and format results (accuracy tables, charts, dataset distributions, etc.).	Anh Khoa	100%
9	Write and format final report with all results, insights, and figures.	Minh Duy	100%
10	Ensure overall cohesion, proofreading, and prepare final PDF submission.	All	100%

4 Self-evaluation

No.	Task Description	Rate
1	Prepare datasets with stratified splits and visualize class distributions.	100%
2	Train and visualize decision tree models on all datasets using multiple train/test splits.	100%
3	Evaluate decision trees using classification reports and confusion matrices.	100%
4	Analyze the impact of decision tree depth on model accuracy.	100%
5	Research and integrate an additional dataset for training and evaluation.	100%
6	Conduct comparative analysis across all datasets.	100%
7	Create charts, tables, and visualizations to support findings.	100%
8	Write and format the final report with insights and well-organized results.	100%
9	Team collaboration and adherence to project schedule.	100%

5 Dataset Analysis and Experiments

5.1 Breast Cancer Wisconsin Dataset

- **Description:** 569 samples, binary labels (malignant vs. benign), 30 numeric features.
- **Preprocessing:** stratified shuffle & split at 40/60, 60/40, 80/20, 90/10.

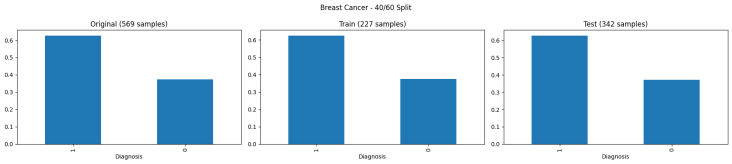

Figure 1: Breast Cancer: class distribution (40/60 split).

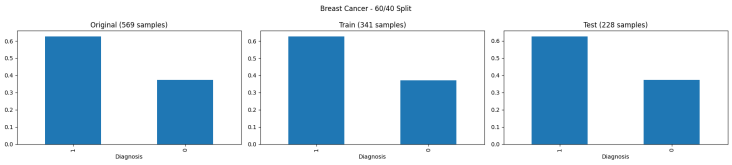

Figure 2: Breast Cancer: class distribution (60/40 split).

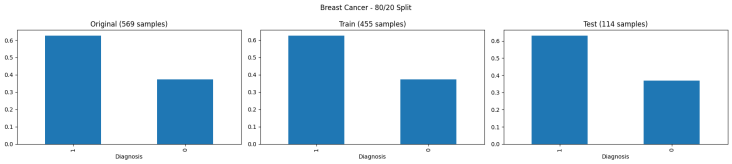

Figure 3: Breast Cancer: class distribution (80/20 split).

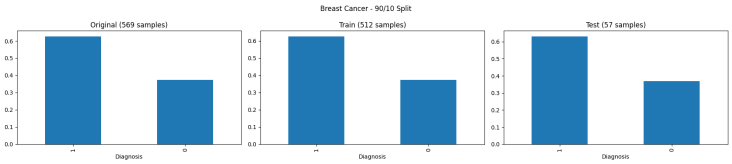

Figure 4: Breast Cancer: class distribution (90/10 split).

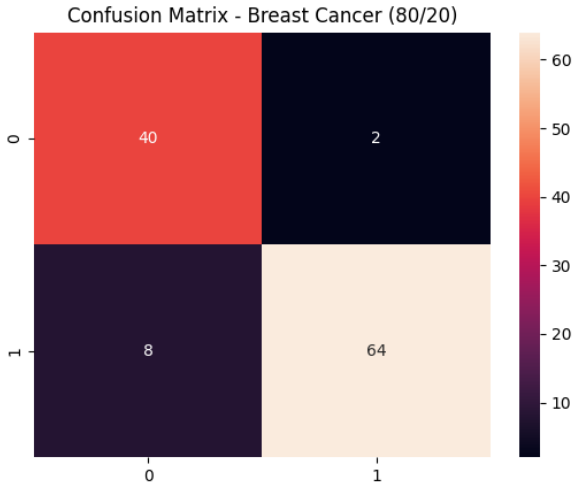

Figure 5: Breast Cancer: confusion matrix (80/20 split).

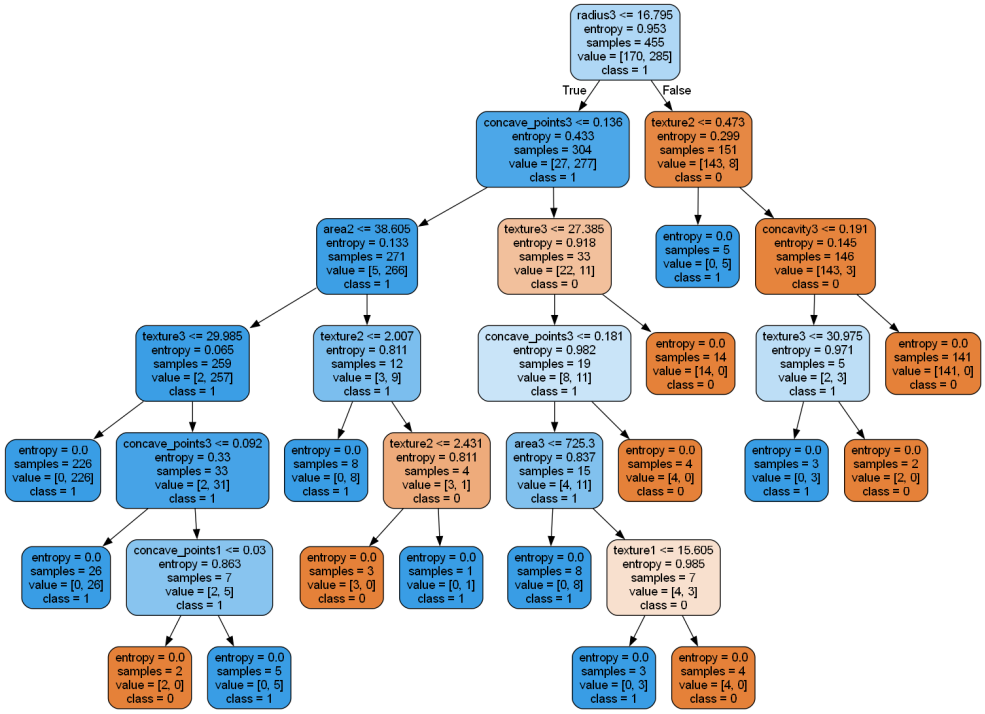

Figure 6: Breast Cancer: decision tree (base) for 80/20 split.

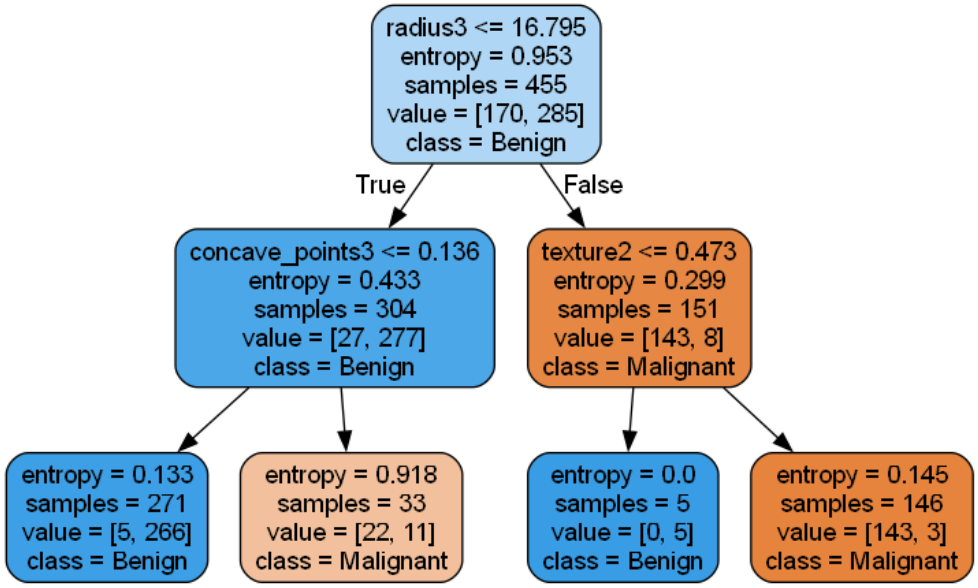

Figure 7: Breast Cancer: decision tree with `max_depth`=2 (80/20 split).

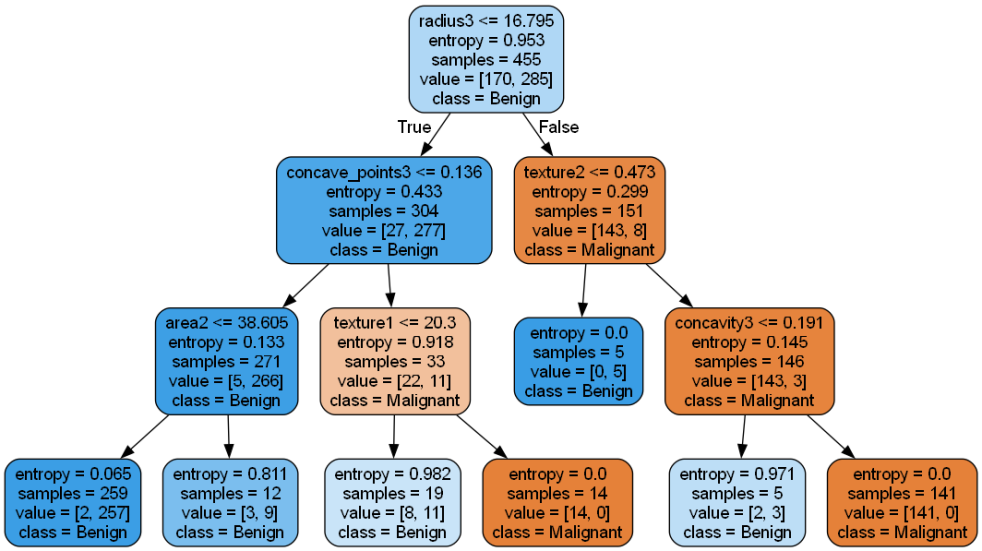

Figure 8: Breast Cancer: decision tree with `max_depth`=3 (80/20 split).

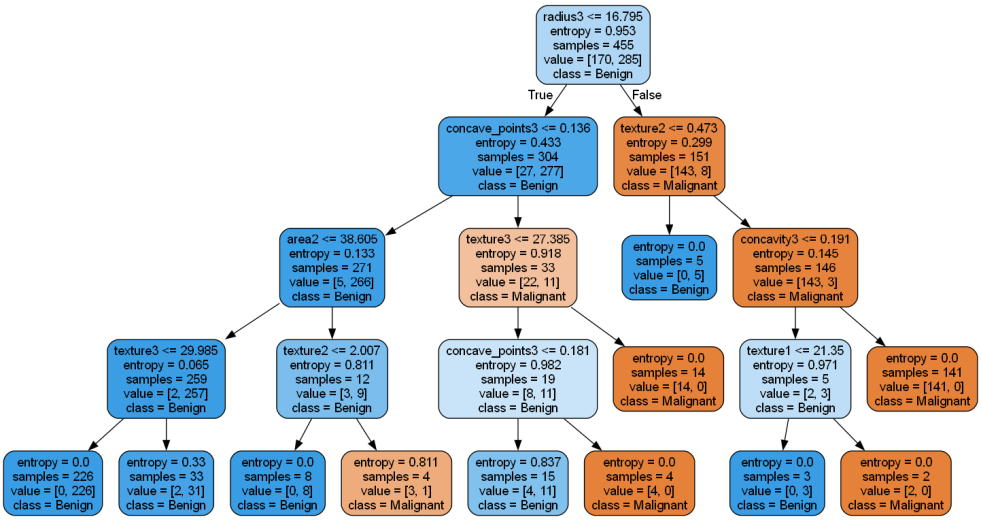

Figure 9: Breast Cancer: decision tree with `max_depth`=4 (80/20 split).

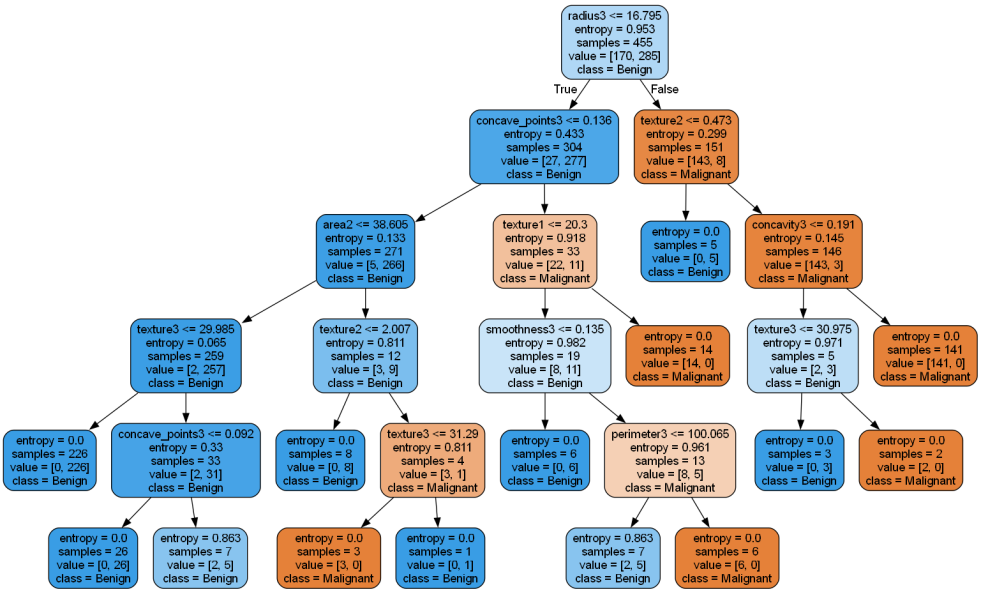

Figure 10: Breast Cancer: decision tree with `max_depth`=5 (80/20 split).

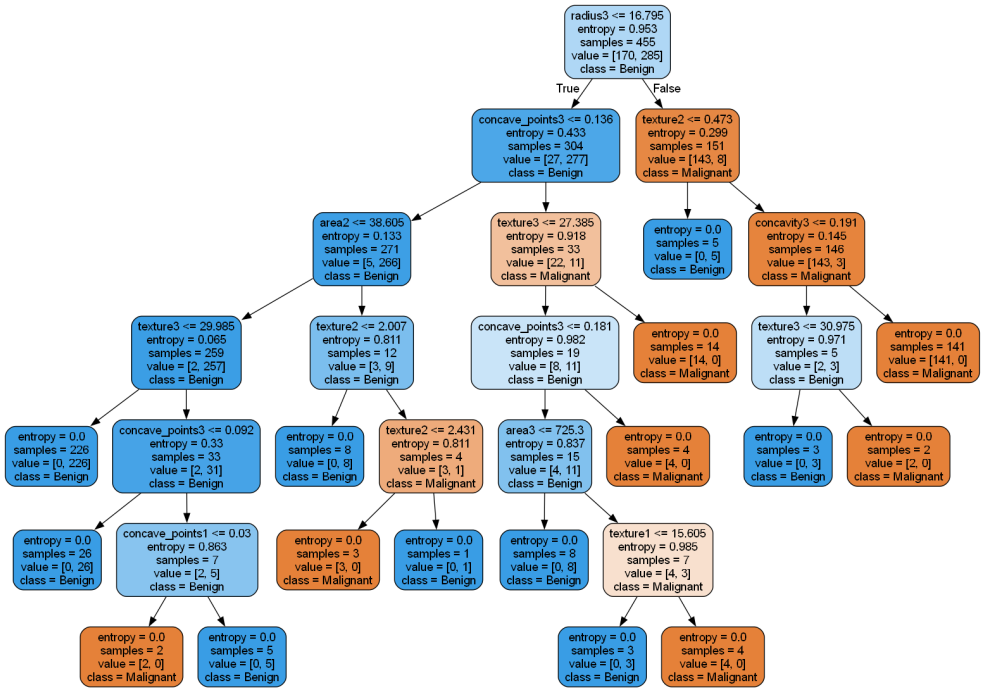

Figure 11: Breast Cancer: decision tree with `max_depth`=6 (80/20 split).

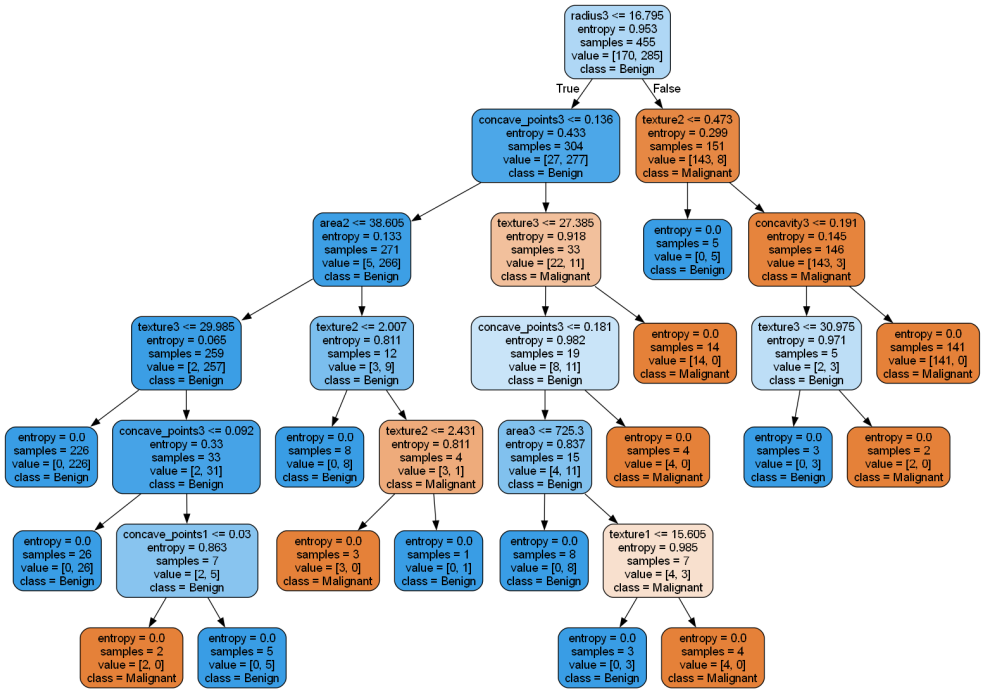

Figure 12: Breast Cancer: decision tree with `max_depth`=7 (80/20 split).

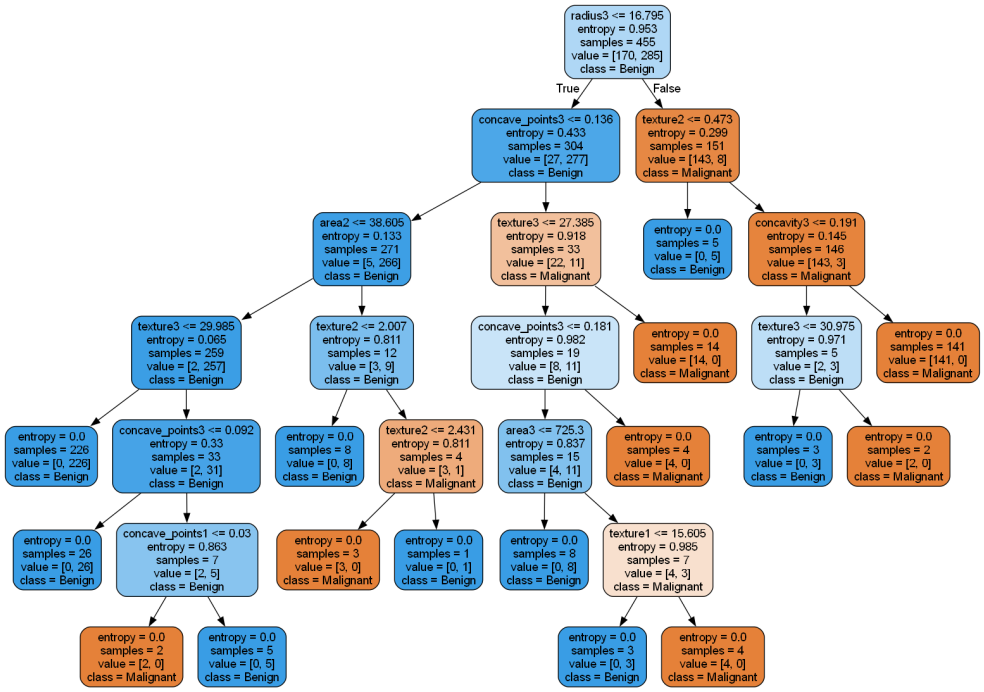

Figure 13: Breast Cancer: decision tree with `max_depth`=None (80/20 split).

5.2 Wine Quality Dataset

- **Description:** 4,898 samples; original scores 0–10 grouped into Low (0–4), Standard (5–6), High (7–10).
- **Preprocessing:** label encoding, stratified splits at 40/60, 60/40, 80/20, 90/10.

Figure 14: Wine Quality: class distribution (40/60 split).

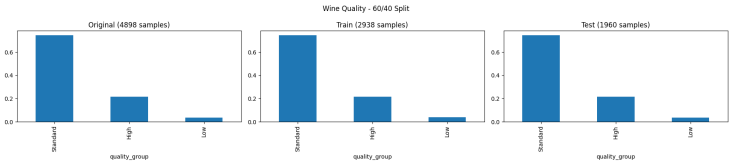

Figure 15: Wine Quality: class distribution (60/40 split).

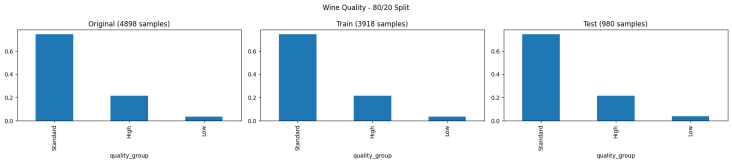

Figure 16: Wine Quality: class distribution (80/20 split).

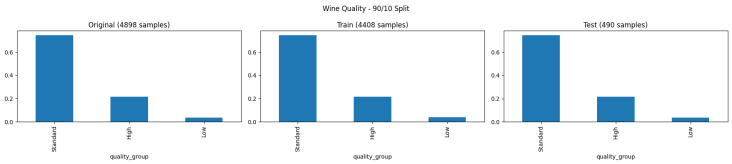

Figure 17: Wine Quality: class distribution (90/10 split).

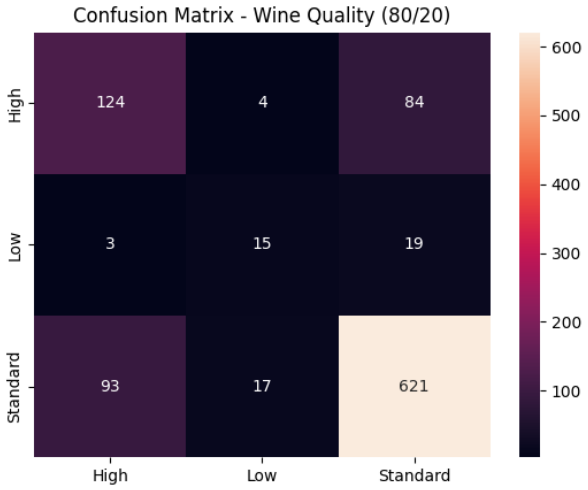

Figure 18: Wine Quality: confusion matrix (80/20 split).

Figure 19: Wine Quality: decision tree (base) for 80/20 split.

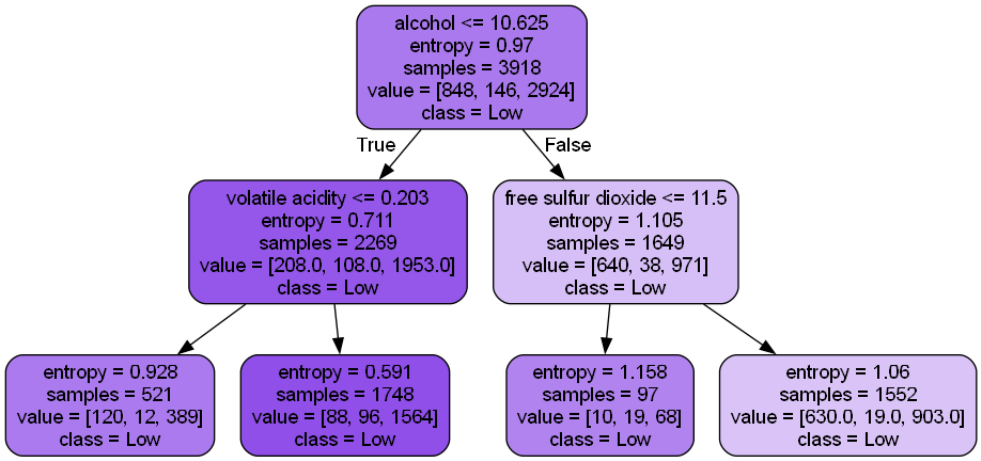

Figure 20: Wine Quality: decision tree with `max_depth`=2 (80/20 split).

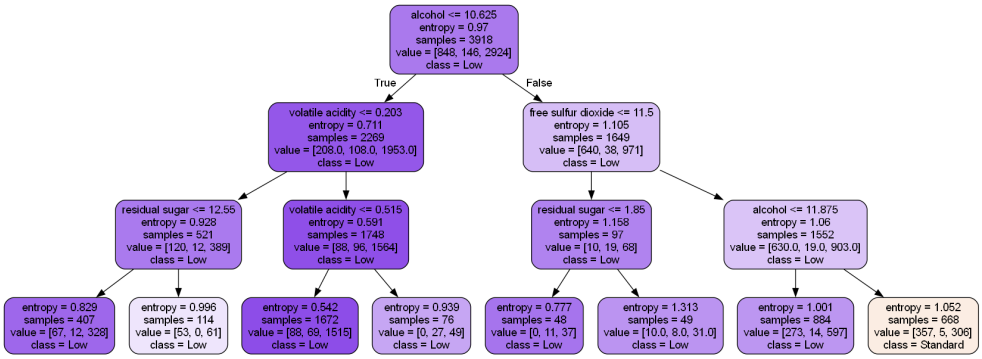

Figure 21: Wine Quality: decision tree with `max_depth`=3 (80/20 split).

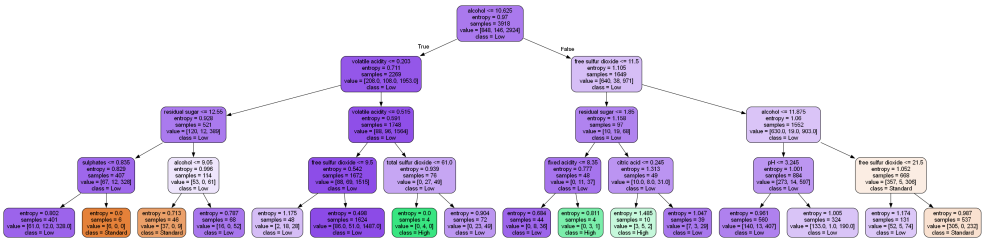

Figure 22: Wine Quality: decision tree with `max_depth`=4 (80/20 split).

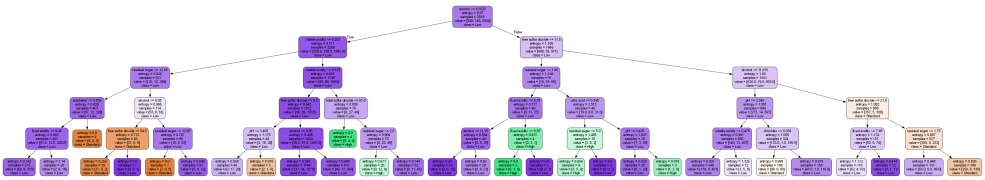

Figure 23: Wine Quality: decision tree with `max_depth`=5 (80/20 split).

Figure 24: Wine Quality: decision tree with `max_depth`=6 (80/20 split).

Figure 25: Wine Quality: decision tree with `max_depth`=7 (80/20 split).

Figure 26: Wine Quality: decision tree with `max_depth`=None (80/20 split).

5.3 Car Evaluation Dataset

- **Description:** 1,728 samples; 4 classes (`unacc`, `acc`, `good`, `vgood`), 6 categorical features.
- **Preprocessing:** label encoding, stratified splits at 40/60, 60/40, 80/20, 90/10.

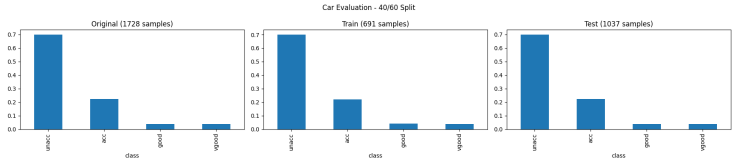

Figure 27: Car Evaluation: class distribution (40/60 split).

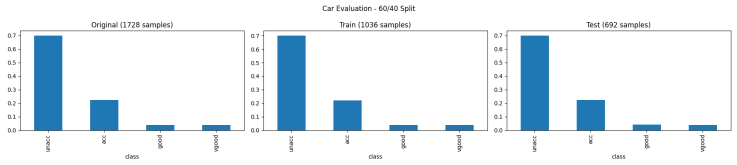

Figure 28: Car Evaluation: class distribution (60/40 split).

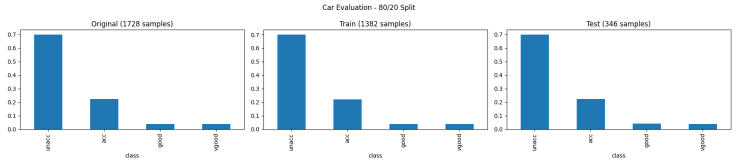

Figure 29: Car Evaluation: class distribution (80/20 split).

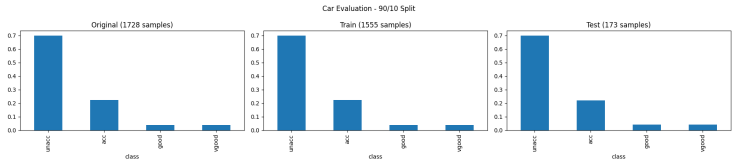

Figure 30: Car Evaluation: class distribution (90/10 split).

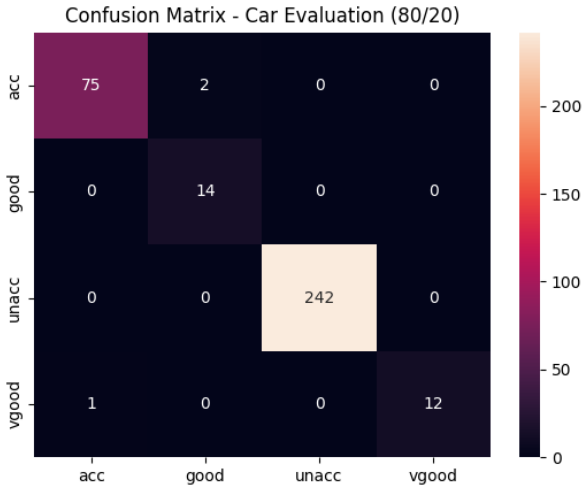

Figure 31: Car Evaluation: confusion matrix (80/20 split).

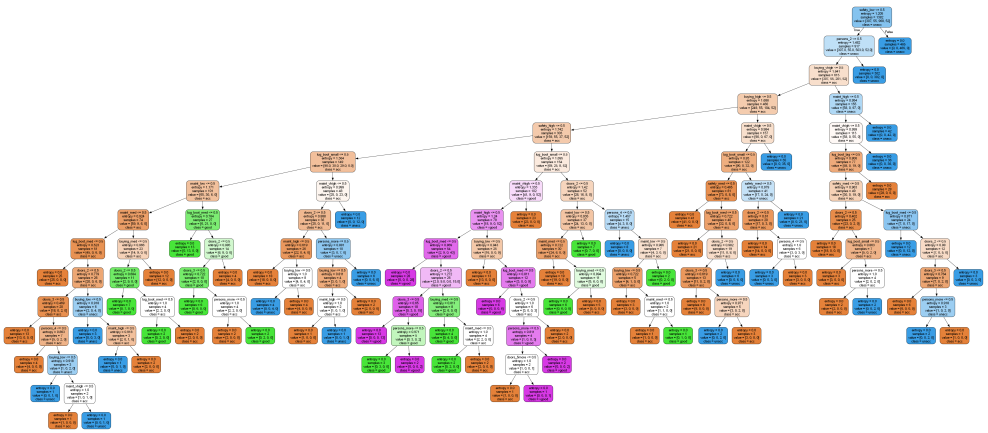

Figure 32: Car Evaluation: decision tree (base) for 80/20 split.

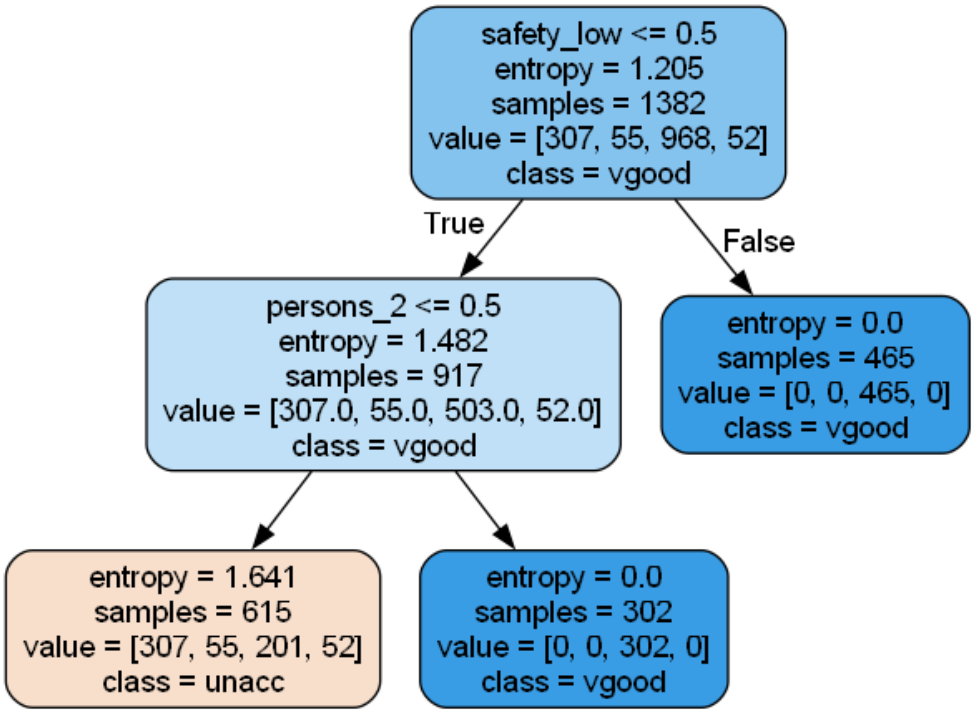

Figure 33: Car Evaluation: decision tree with `max_depth`=2 (80/20 split).

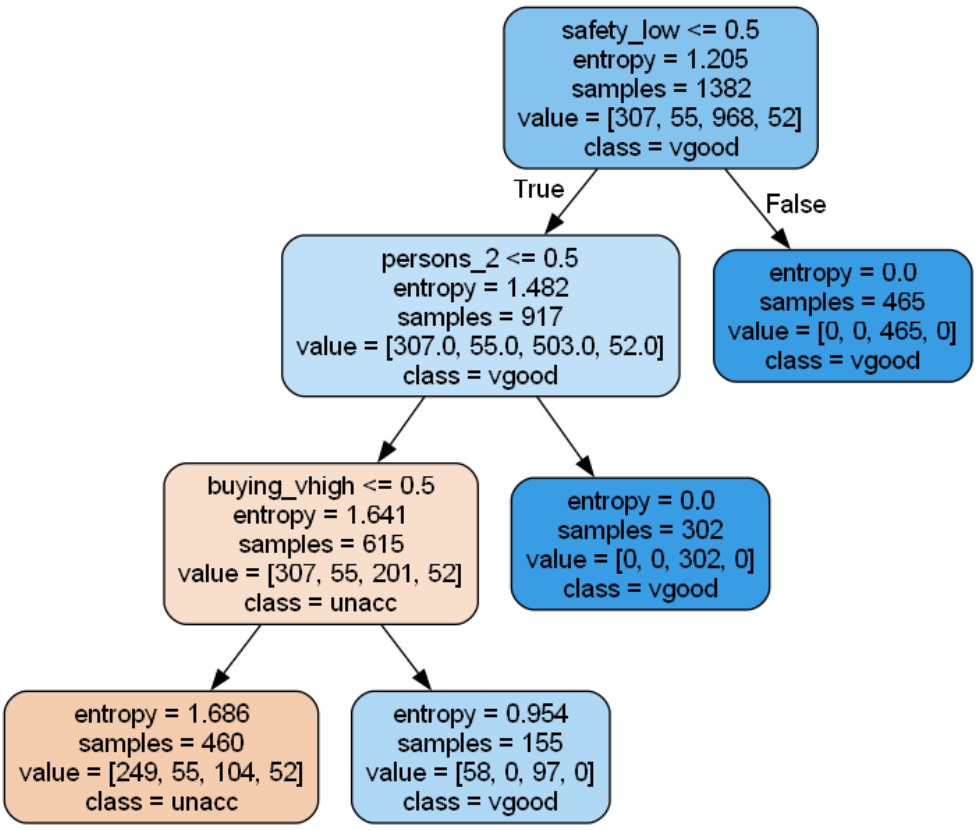

Figure 34: Car Evaluation: decision tree with `max_depth`=3 (80/20 split).

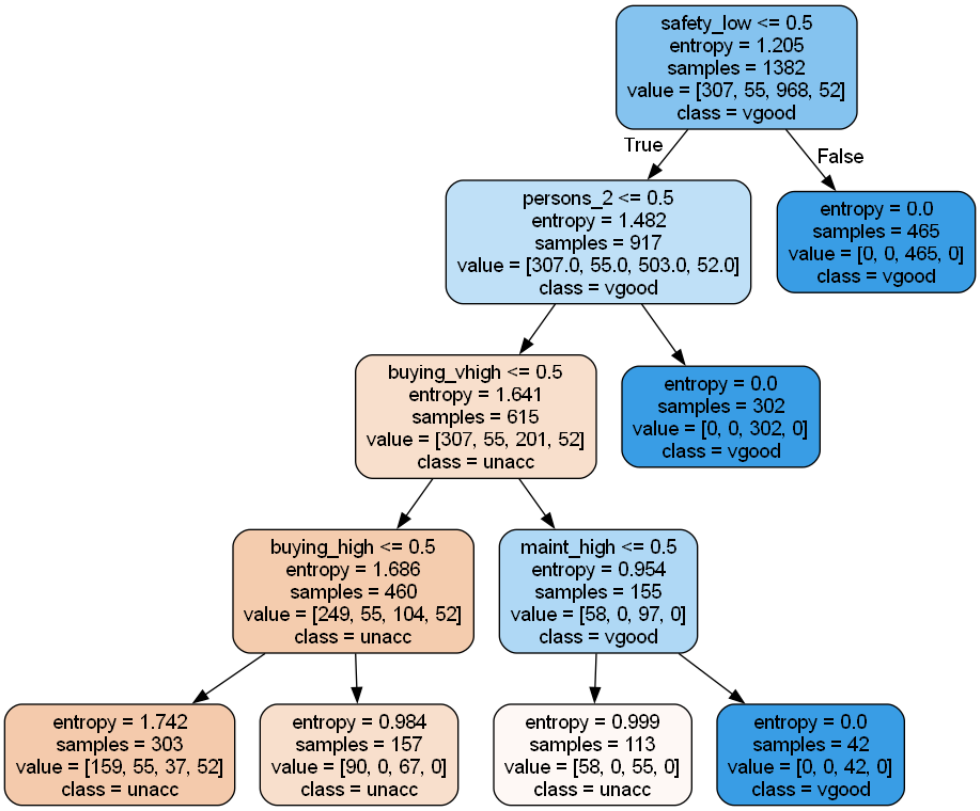

Figure 35: Car Evaluation: decision tree with `max_depth`=4 (80/20 split).

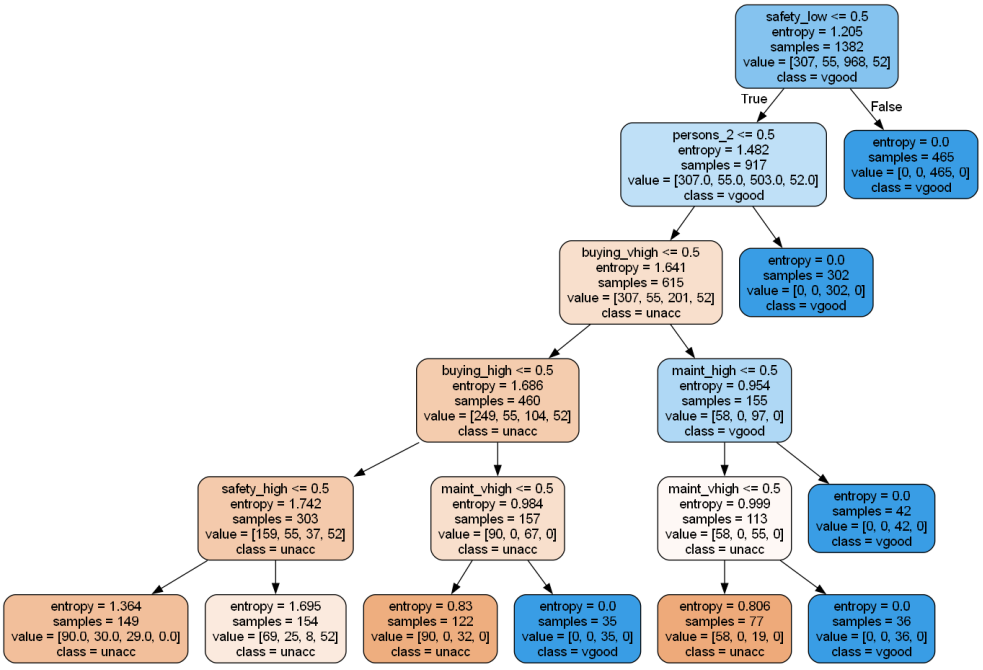

Figure 36: Car Evaluation: decision tree with `max_depth`=5 (80/20 split).

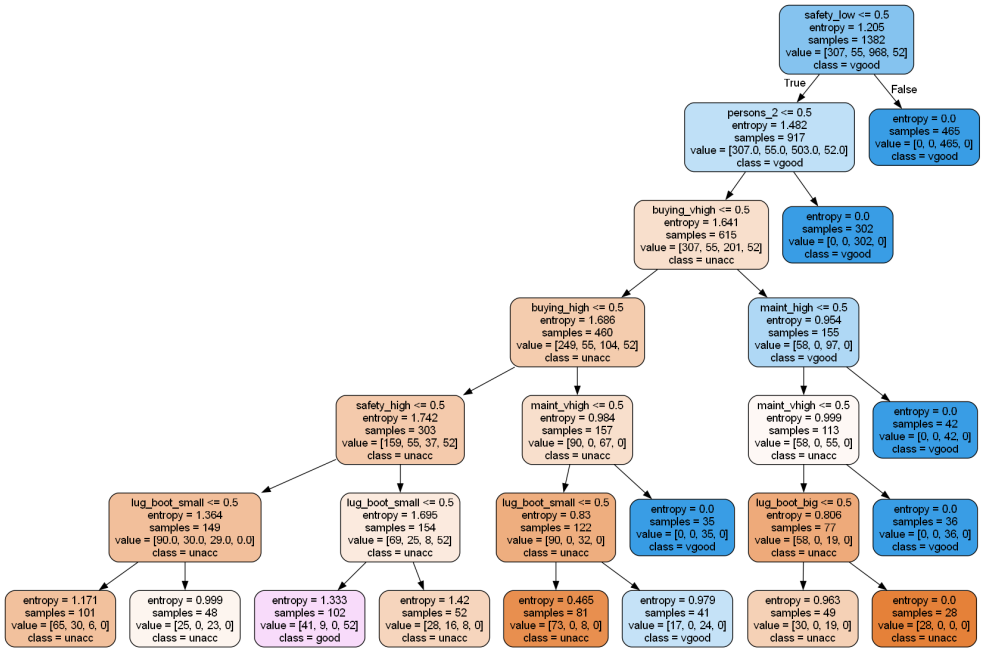

Figure 37: Car Evaluation: decision tree with `max_depth`=6 (80/20 split).

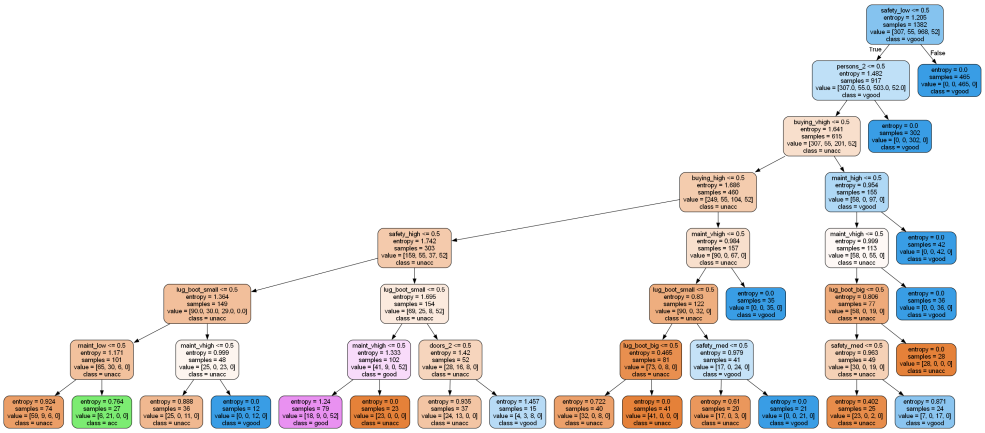

Figure 38: Car Evaluation: decision tree with `max_depth`=7 (80/20 split).

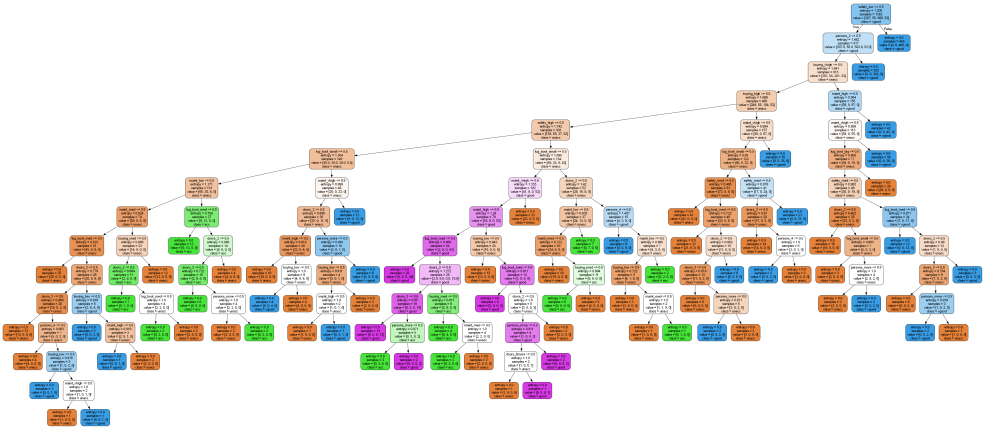

Figure 39: Car Evaluation: decision tree with `max_depth`=None (80/20 split).

6 Comparative Analysis

- **Objective:** Compare Decision Tree performance across:
  1. Breast Cancer (binary classification, continuous features)
  2. Wine Quality (multi-class, numerical features)
  3. Car Evaluation (multi-class, categorical features)
- **Comparison Criteria:**
  - Accuracy, Precision, Recall, F1-Score
  - Effect of feature type and count
  - Class distribution and balance
  - Impact of `max_depth` on overfitting
- **Observations:**
  - **Breast Cancer:** Highest accuracy; binary labels and well-separated numeric features helped model performance.
  - **Wine Quality:** Lower precision for middle-quality wines; overlapping features across quality groups reduced clarity.
  - **Car Evaluation:** Performed well despite 4 classes; decision tree easily handled categorical data. Slight overfitting observed at deep trees.
- **Conclusion:**
  - Decision Trees adapt well to both categorical and numerical data, but class imbalance and feature overlap affect performance.
  - Simpler datasets with clear boundaries (like Breast Cancer or Car Evaluation) yield higher accuracy.
  - Proper depth tuning is essential to maintain generalization.

7 References

1. DecisionTreeClassifier Documentation
2. Breast Cancer Dataset
3. Wine Quality Dataset