

University of Science and Technology of Hanoi



Machine Learning and Data Mining 2
Labwork 4 - Classification II

BA12-118 Nguyen Phuc Minh
BA12-066 Nguyen Hoang Ha

TABLE OF CONTENTS

I. Datasets.....	3
II. Decision Tree.....	3
Car evaluation.....	4
Drug classification.....	6
III. Random Forest.....	8
Car evaluation.....	9
Drug classification.....	10

I. Datasets

- Car evaluation
 - There are seven variables in the dataset. All the variables are of categorical data type.
 - These are given by “ buying, maint, doors, persons, lug_boot, safety and class ”.
 - “Class” is the target variable.
- Drug classification
 - There are six variables in the dataset.
 - These are given by “ Age, Sex, BP, Cholesterol, Na_to_K, Drug ”.
 - “Drug” is the target variable.

II. Decision Tree

- Check missing values and encode categorical variables.
- Split data into separate training (80%) and testing (20%) sets.
- Implement Decision Tree

The primary challenge in Decision Tree implementation is identifying the attributes to use as the root node at each level. This process, known as attribute selection, determines the splits that will best separate the data.

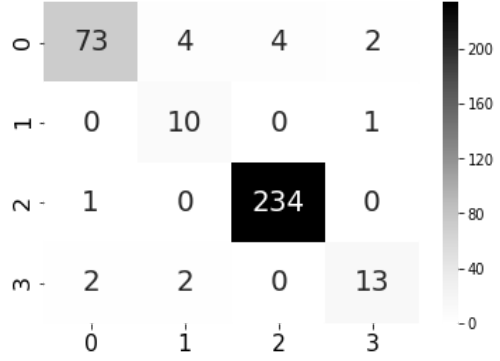
Common attribute selection measures include: Information gain (Entropy) and the Gini index

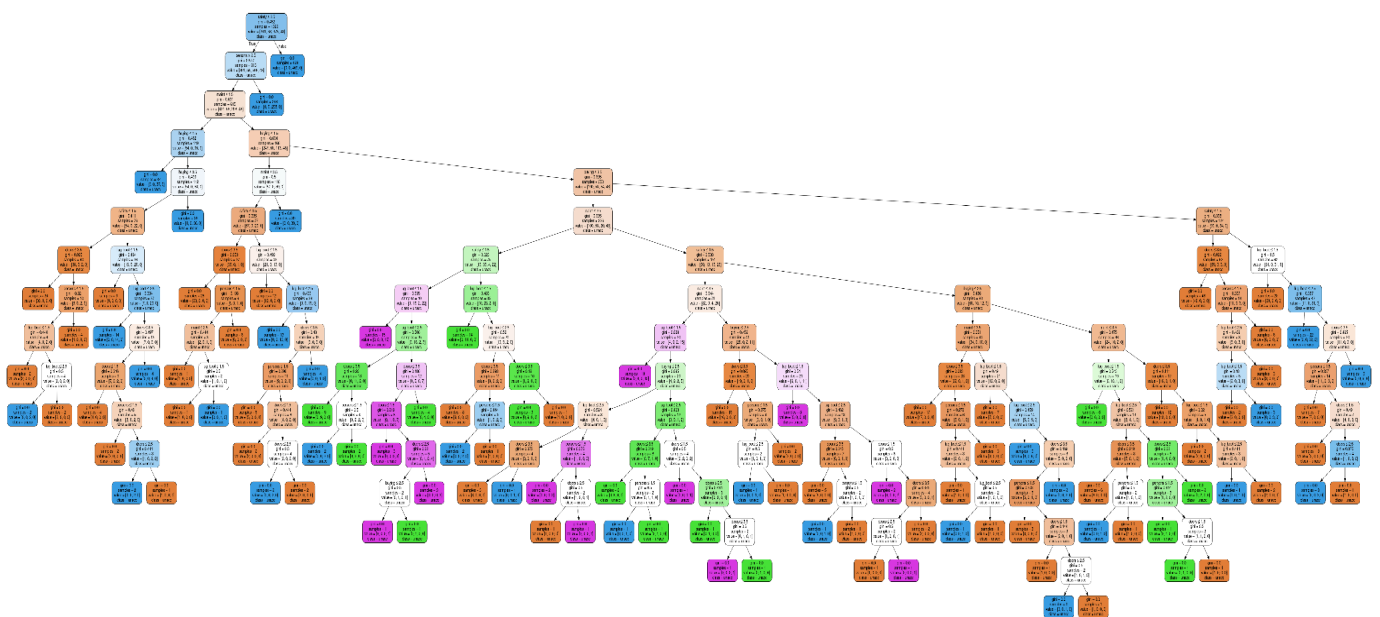
The Decision Tree algorithm uses entropy to calculate information gain. By measuring the decrease in entropy for each attribute, the information gain is determined. The attribute with the highest information gain is chosen as the splitting attribute at the node.

For discrete-valued attributes, the subset with the minimum Gini index is chosen as the splitting attribute. For continuous-valued attributes, each pair of adjacent values is evaluated as a split point, and the point with the smallest Gini index is selected. The attribute with the lowest Gini index overall is used for splitting.

Car evaluation

No max_depth defined

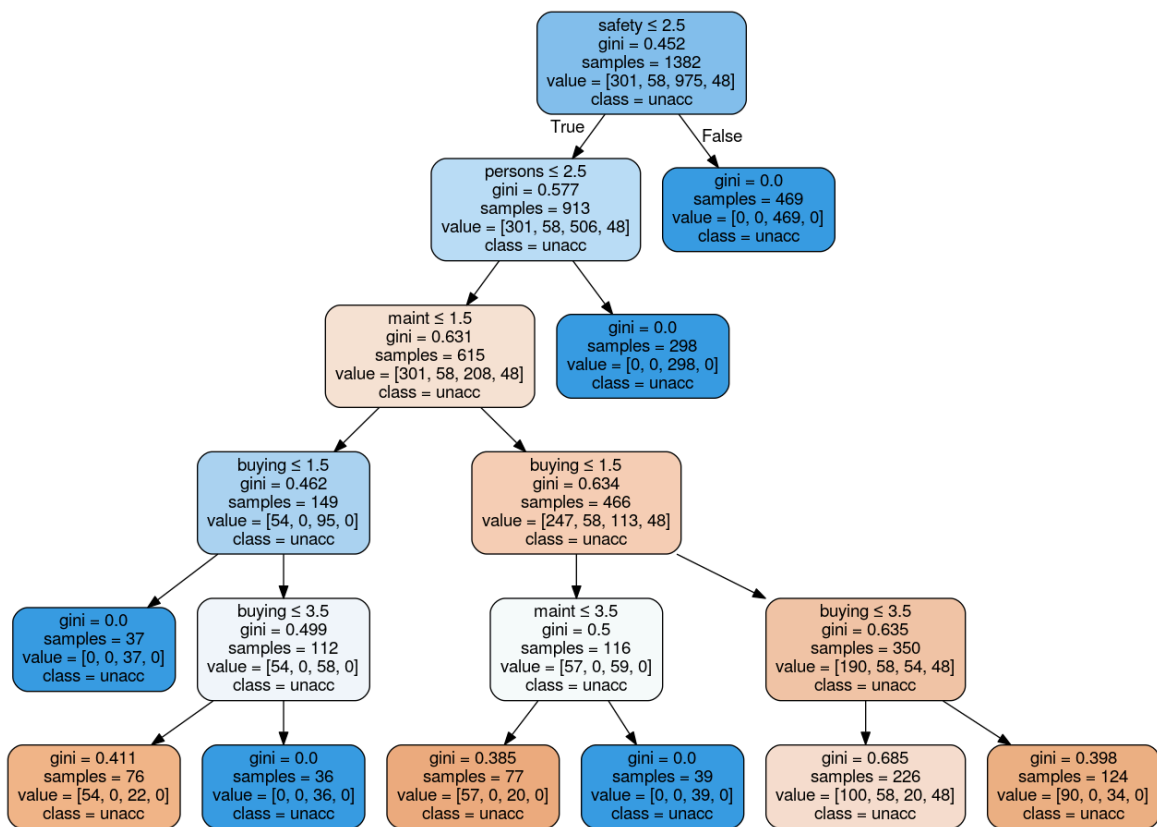
depth=None	Precision	Recall	F1-core	Support	Confusion Matrix				
acc	0.96	0.88	0.92	83					
good	0.62	0.91	0.74	11					
unacc	0.98	1.00	0.99	235					
vgood	0.81	0.76	0.79	17					
Accuracy			0.95	346					
Macro avg	0.85	0.89	0.86	346					
Weighted avg	0.96	0.95	0.95	346					



By limiting the depth, we can achieve a better balance between bias and variance, leading to improved generalization and manageable model complexity.

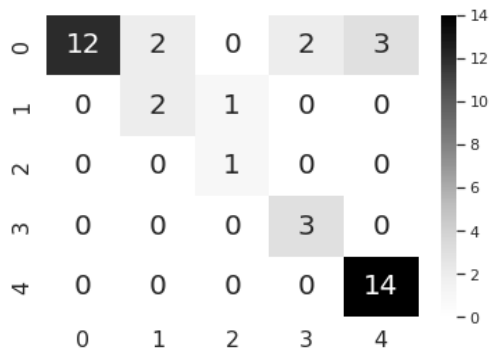
Max_depth = 5

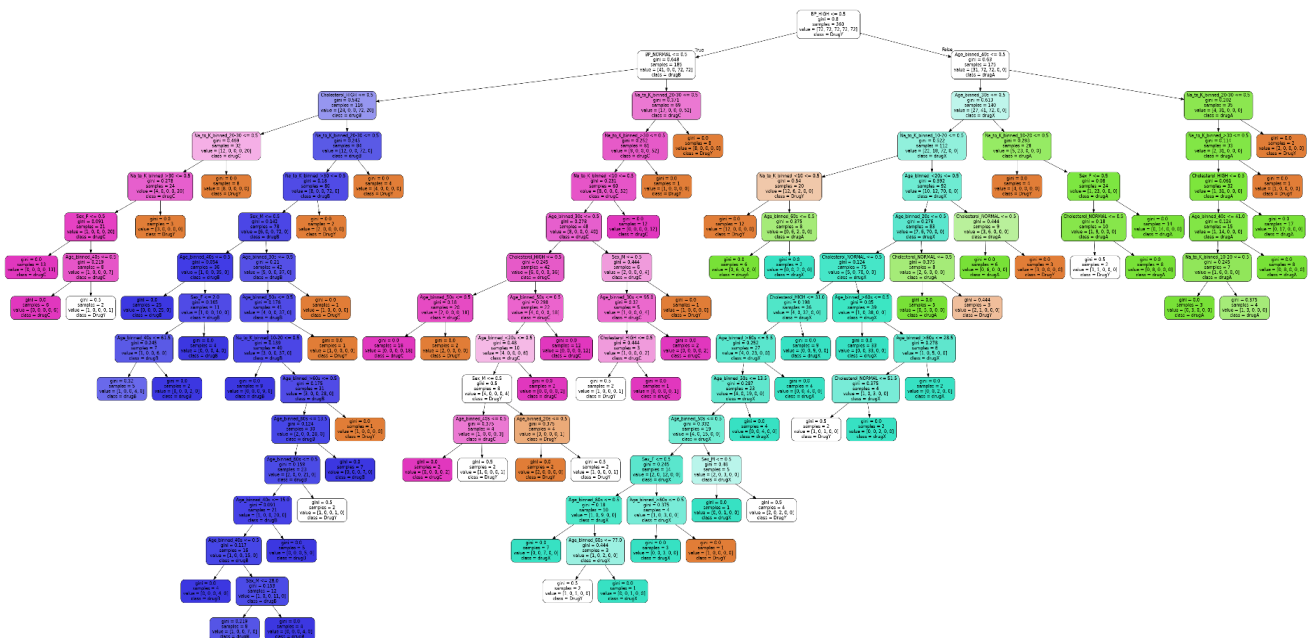
depth=5	Precision	Recall	F1-core	Support	Confusion Matrix
acc	0.69	1.00	0.81	83	
good	0.00	0.00	0.00	11	
unacc	1.00	0.96	0.98	235	
vgood	0.00	0.00	0.00	17	
Accuracy			0.89	346	
Macro avg	0.42	0.49	0.45	346	
Weighted avg	0.84	0.89	0.86	346	



Drug classification

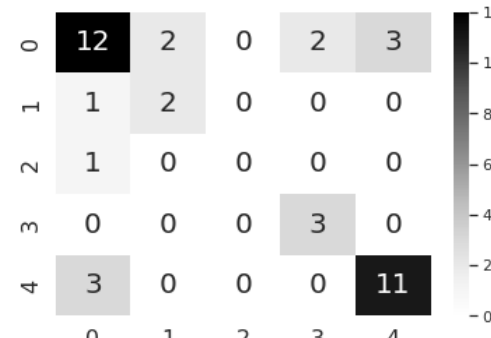
No max_leaf_nodes defined

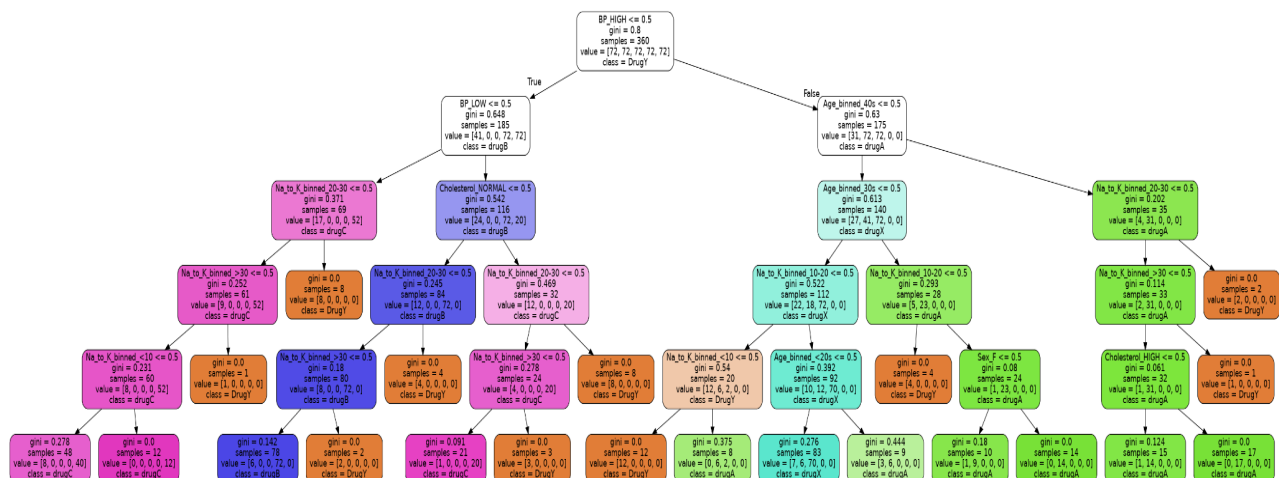
depth=5	Precision	Recall	F1-score	Support	Confusion Matrix					
DrugY	0.71	0.63	0.67	19						
drugA	0.50	0.67	0.57	3						
drugB	0.00	0.00	0.00	1						
drugC	0.60	1.00	0.75	3						
drugX	0.79	0.79	0.79	14						
Accuracy			0.70	40						
Macro avg	0.52	0.62	0.55	40						
Weighted avg	0.69	0.70	0.69	40						



By limiting the leaf nodes, we also can achieve a better balance between bias and variance, leading to improved generalization and manageable model complexity. In this case, it even improves the model's accuracy

max_leaf_nodes = 20

depth=5	Precision	Recall	F1-score	Support	Confusion Matrix					
DrugY	1.00	0.63	0.77	19						
drugA	0.50	0.91	0.57	3						
drugB	0.50	1.00	0.67	1						
drugC	0.60	1.00	0.75	3						
drugX	0.82	1.00	0.90	14						
Accuracy			0.80	40						
Macro avg	0.68	0.86	0.73	40						
Weighted avg	0.86	0.80	0.80	40						



III. Random Forest

- Check missing values and encode categorical variables.
- Create 100 training sets using the bagging technique and a single testing set.
- Implement Random Forest

The primary challenge in implementing a Random Forest is determining the attributes for splitting at each node within the individual decision trees. This process, known as attribute selection, identifies the splits that best separate the data. Standard attribute selection measures used in Random Forests include information gain (based on entropy) and the Gini index.

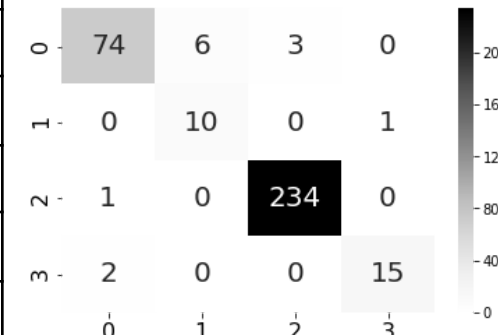
Random Forest algorithms use entropy to calculate information gain. By measuring the decrease in entropy for each attribute, the information gain is assessed. The attribute with the highest information gain is selected as the splitting attribute at each node within a tree.

For discrete-valued attributes, the subset with the minimum Gini index is selected as the splitting attribute. For continuous-valued attributes, each pair of adjacent values is evaluated as a potential split point, with the point that results in the smallest Gini index being chosen. The attribute with the lowest overall Gini index is then used for splitting the individual trees within the Random Forest.

In Part II, we applied the Decision Tree algorithm to both the original dataset and the dataset with limited leaf nodes. In this section, we will select the dataset that achieved the higher accuracy with the Decision Tree and apply the Random Forest model to it. Our goal is to observe changes in accuracy and other performance metrics when using the Random Forest algorithm.

Car evaluation

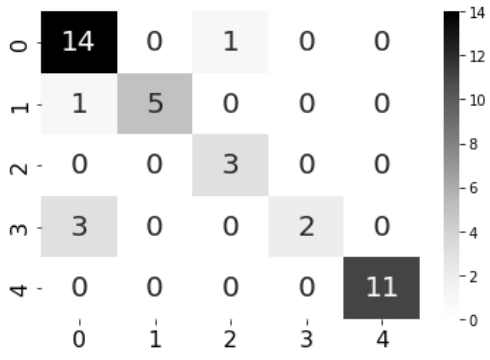
No max_depth defined

depth=none	Precision	Recall	F1-score	Support	Confusion Matrix																													
acc	0.96	0.89	0.92	83	 <table><tr><th></th><th>0</th><th>1</th><th>2</th><th>3</th></tr><tr><th>0</th><td>74</td><td>6</td><td>3</td><td>0</td></tr><tr><th>1</th><td>0</td><td>10</td><td>0</td><td>1</td></tr><tr><th>2</th><td>1</td><td>0</td><td>234</td><td>0</td></tr><tr><th>3</th><td>2</td><td>0</td><td>0</td><td>15</td></tr></table>						0	1	2	3	0	74	6	3	0	1	0	10	0	1	2	1	0	234	0	3	2	0	0	15
	0	1	2	3																														
0	74	6	3	0																														
1	0	10	0	1																														
2	1	0	234	0																														
3	2	0	0	15																														
good	0.62	0.91	0.74	11																														
unacc	0.99	1.00	0.99	235																														
vgood	0.94	0.88	0.91	17																														
Accuracy			0.96	346																														
Macro avg	0.88	0.92	0.89	346																														
Weighted avg	0.97	0.96	0.96	346																														

→ The Random Forest model achieved a slightly higher accuracy (0.96) compared to the Decision Tree model (0.95). In terms of precision, recall, and F1-scores, the Random Forest generally outperforms the Decision Tree, especially in the 'vgood' category. Both models show high performance in the 'unacc' category, but Random Forest maintains a slight edge in other categories. The higher macro-average and weighted average F1 scores for Random Forests show that it maintains better balance across all classes, while Decision Trees show a slightly lower average, giving less consistent handling of categories.

Drug classification

max_leaf_nodes = 20

depth=5	Precision	Recall	F1-score	Support	Confusion Matrix
DrugY	0.78	0.93	0.85	15	
drugA	1.00	0.83	0.91	6	
drugB	0.75	1.00	0.86	3	
drugC	1.00	0.40	0.57	5	
drugX	1.00	1.00	1.00	11	
Accuracy			0.88	40	
Macro avg	0.91	0.83	0.84	40	
Weighted avg	0.90	0.88	0.87	40	

→ The Random Forest model achieved higher accuracy (0.88) than the Decision Tree model (0.80). This shows that for this particular dataset and configuration, Random Forest is performing better in terms of overall classification accuracy. Additionally, Random Forest demonstrated higher precision and recall for most drugs, resulting in higher F1 scores. This indicates that Random Forests have a better balance between precision and recall than Decision Trees.