Reichman University

Introduction to Data Science

Homework Assignment 3 – Decision Trees

Dr. Gail Gilboa-Freedman
Dr. Naveh Eskinazi

Submission: 25/05/2022

GENERAL INSTRUCTIONS

In the current assignment you will classify used cars with the decision tree algorithm.

The work will be based on a CSV named **"car_evaluation.csv"** located on the course's Moodle site.

SUBMISSION:

Through the assignment box within the course Moodle, submit a **Jupyter Notebook file named HWA3_<student name>.ipynb** (e.g. HWA3_karin_tenne.ipynb)
Should include all the relevant code needed to perform the assignment's tasks along with the code's output.
(Recommendation: Add headers and sub-headers using the Markdown option)

Good Luck!

PART 1: PREREQUISITES

TASK 1: SETTING THE FOLDER

1. Create a Jupyter Notebook named **HWA3_<student name>.ipynb.**
2. Download from the CSV file named **"car_evaluation.csv"** from Moodle.
3. Upload the CSV file to Jupyter (Note: make sure the file is placed in the same location as your Jupyter Notebook)

TASK 2: IMPORT LIBRARIES & MODULES

4. Import the following libraries and modules within your notebook: **panda, numpy, DecisionTreeClassifier (from sklearn.tree), tree (from sklearn), and metrics (from sklearn)**

TASK 3: EXPLORE THE DATA

Use Python commands (e.g., head, columns, and shape) to plot the answers to the following questions:

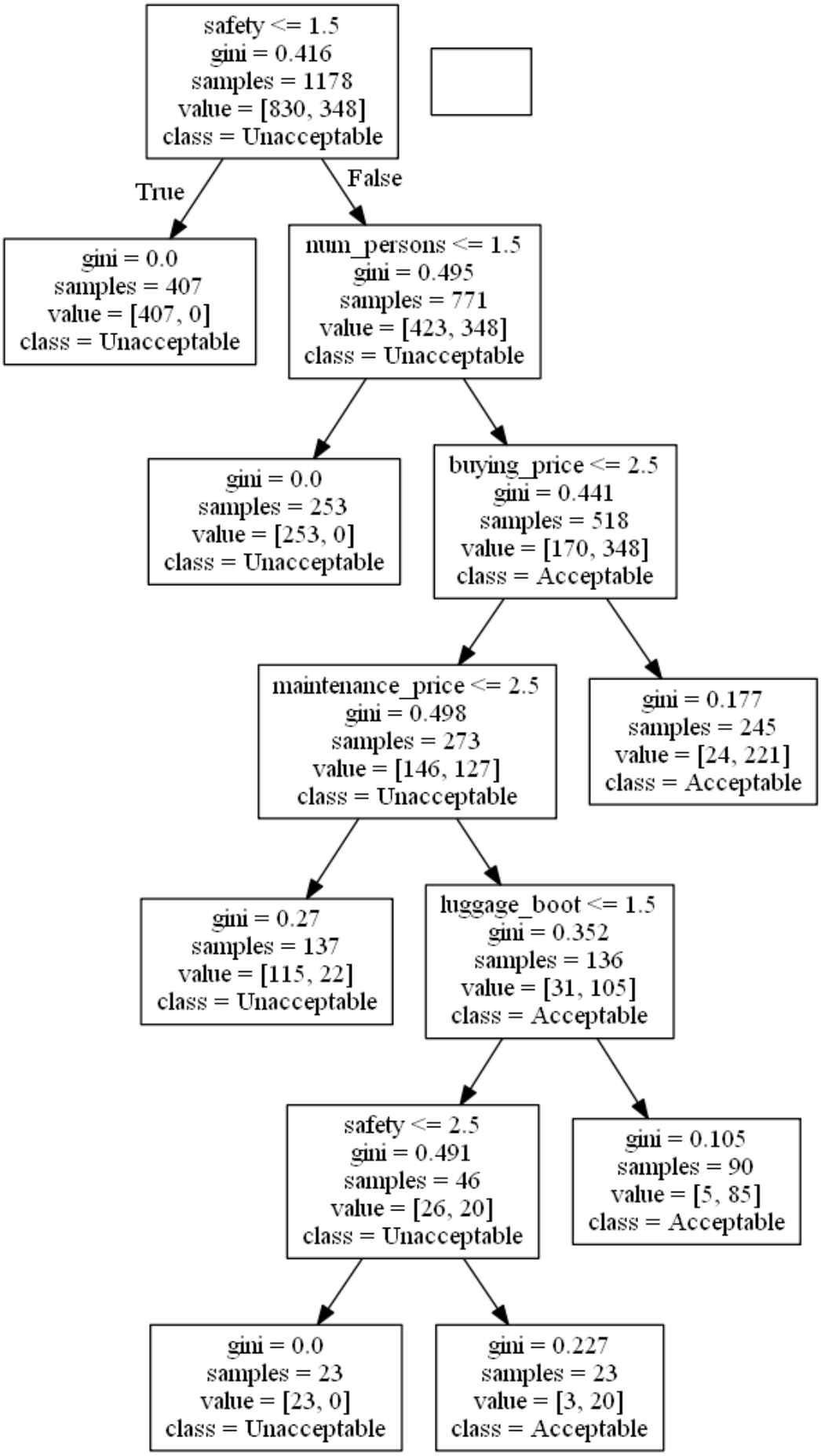
5. Based on how many **cars** the algorithm will be trained and tested? (Provide a numerical answer for both train and test)
6. How many **features** each car is represented by? (Provide a numerical answer)
7. Which features are categorized as **predictors**? (Provide a verbal answer)
8. Which features are categorized as **the label (target(**? (Provide a verbal answer)

PART 2: BUILDING A DECISION TREE MODEL

TASK 4: BUILDING THE MODEL

Use Python commands (i.e., DecisionTreeClassifier and fit) to build a Decision Tree model.

9. In next page you will see visual representation of the tree.


safety <= 1.5
gini = 0.416
samples = 1178
value = [830, 348]
class = Unacceptable
True
False
gini = 0.0
samples = 407
value = [407, 0]
class = Unacceptable
num_persons <= 1.5
gini = 0.495
samples = 771
value = [423, 348]
class = Unacceptable
gini = 0.0
samples = 253
value = [253, 0]
class = Unacceptable
buying_price <= 2.5
gini = 0.441
samples = 518
value = [170, 348]
class = Acceptable
maintenance_price <= 2.5
gini = 0.498
samples = 273
value = [146, 127]
class = Unacceptable
gini = 0.177
samples = 245
value = [24, 221]
class = Acceptable
gini = 0.27
samples = 137
value = [115, 22]
class = Unacceptable
luggage_boot <= 1.5
gini = 0.352
samples = 136
value = [31, 105]
class = Acceptable
safety <= 2.5
gini = 0.491
samples = 46
value = [26, 20]
class = Unacceptable
gini = 0.105
samples = 90
value = [5, 85]
class = Acceptable
gini = 0.0
samples = 23
value = [23, 0]
class = Unacceptable
gini = 0.227
samples = 23
value = [3, 20]
class = Acceptable

10. As can be seen from the decision tree, the attribute that was selected for the root of the tree is 'Safety'. The Gini index value of this attribute is 0.416. Provide a detailed calculation of this value (Hint: You can use Excel to perform the calculation).
11. Below are three used cars offered for sale. Based on the decision tree, determine for each of them whether it will be acceptable or not (Provide a verbal answer):
    - Car with a low buying price, very high maintenance price, 3 doors, 2 persons can be carried, medium luggage boot and low in safety.
    - Car with a high or very high buying price, medium maintenance price, 4 doors, 4 or more persons can be carried, big luggage boot and medium or high in safety.
    - Car with a medium buying, low maintenance price, 2 doors, 2 persons can be carried, small luggage boot and medium or high in safety.

TASK 5: EVALUATE THE MODEL

12. Use the Accuracy measure to evaluate the mode you have created in Task 4.
13. (Use your own words) Describe, in the context of the model you have built, **the meaning of the numerical result** you have received for the Accuracy measure (Provide a verbal answer using terms such as classification matrix, True Positive, etc.)
14. (Use your own words) Describe **one prominent disadvantage** of the Accuracy measure you have calculated.

Good Luck!