Topic 5: Decision Trees

Jonathan Ibifubara Pollyn

GitHub Repository: https://github.com/JonathanPollyn/Machine-Learning-for-Data-Science

College of Science, Engineering and Technology, Grand Canyon University

DSC-540: Machine Learning for Data Science

Dr. Aiman Darwiche

December 1, 2021

Introduction

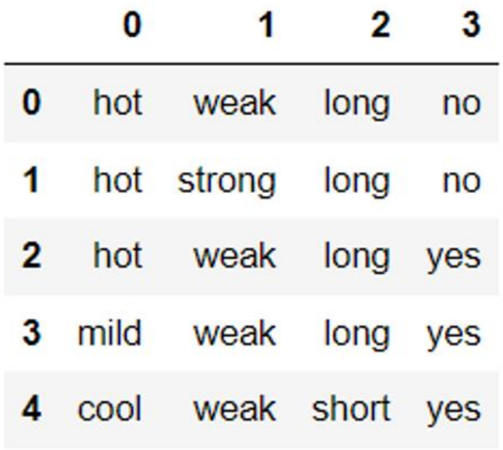
Decision trees are built using an algorithm that detects multiple segments of a data set based on certain conditions. It's one of the most popular and practical supervised learning approaches. Decision Trees are a non-parametric supervised learning method that can be used for classification and regression. Classification trees are tree models that work with a discrete range of values. In contrast, regression trees are decision trees where the target variable has a range of values (usually real numbers).

Decision tree techniques have already been demonstrated to be understandable, efficient, problem-independent, and capable of large-scale handling applications. However, they're also highly unstable classifiers regarding slight changes in the training data or approaches with many variances. Because of the elasticity of fuzzy sets' formalism, fuzzy logic improves these characteristics. A decision tree is applied to a dataset for people considering weather and traffic while deciding whether to commute by vehicle or public transportation for 14 days. The data are temperature, Wind, and Traffic-Jam as its attributes, and Car Driving Decision is the target variable.

Implementation

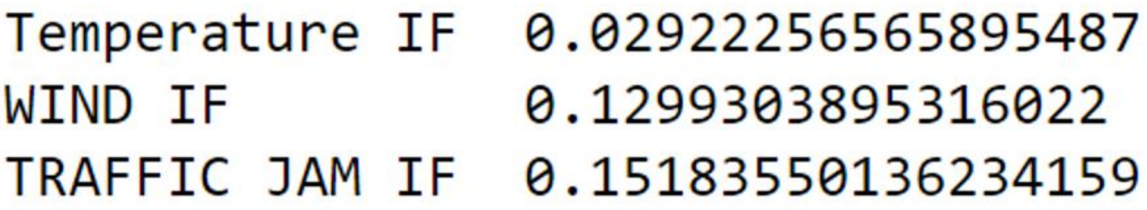
Figure 1 shows the data which have been extracted into a DataFrame for easy readability. The data have been transformed to present the columns temperatures, wind, traffic-jam, and car driving to 0,1,2, and 3, respectively. Figure 2 shows the gain calculation for the first three columns (temperature, wind, and traffic-Jam); the analysis indicates that the traffic jam came out approximately 0.15, the highest among the values calculated. From this information, the root node was selected to be the traffic. Then the ID3 algorithm from the sklearn was followed to get the Gini index which impurity in the data of that feature. The decision tree is presented in figure 1.3. The tree starts with the root node at samples and a Gini Index of 0.459.

In this node, the feature that best splits the different data classes is the traffic with a threshold value of 0.5. This results in two nodes, one with Gini 0.49 and one with Gini of 0.245. Figure 4 shows the confusion matrix when the entropy and Gini index are used. We see that they both have an accuracy of 60 and a weighted average of 0.36.



	0	1	2	3
0	hot	weak	long	no
1	hot	strong	long	no
2	hot	weak	long	yes
3	mild	weak	long	yes
4	cool	weak	short	yes

Figure 1: Commute data presented in Python DataFrame



```
Temperature IF  0.02922256565895487
WIND IF         0.1299303895316022
TRAFFIC JAM IF  0.15183550136234159
```

Figure 2: Gain calculation for Temperature, Wind and Traffic Jam

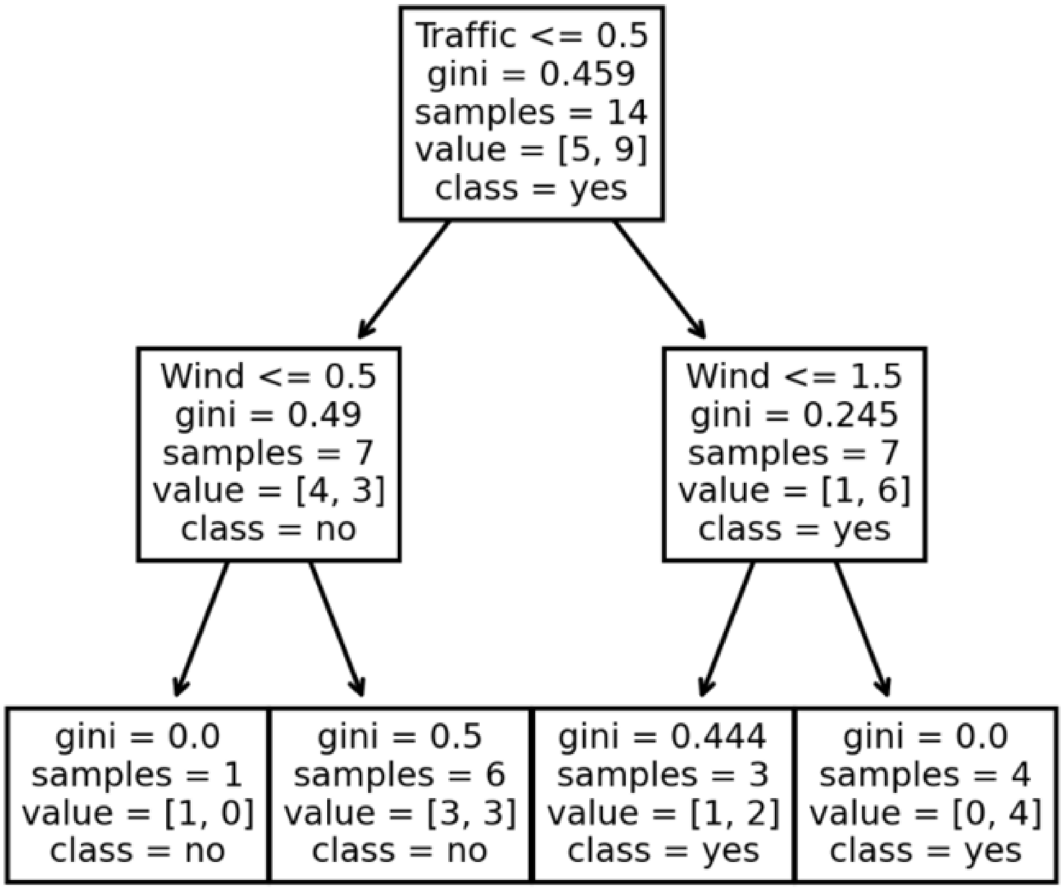
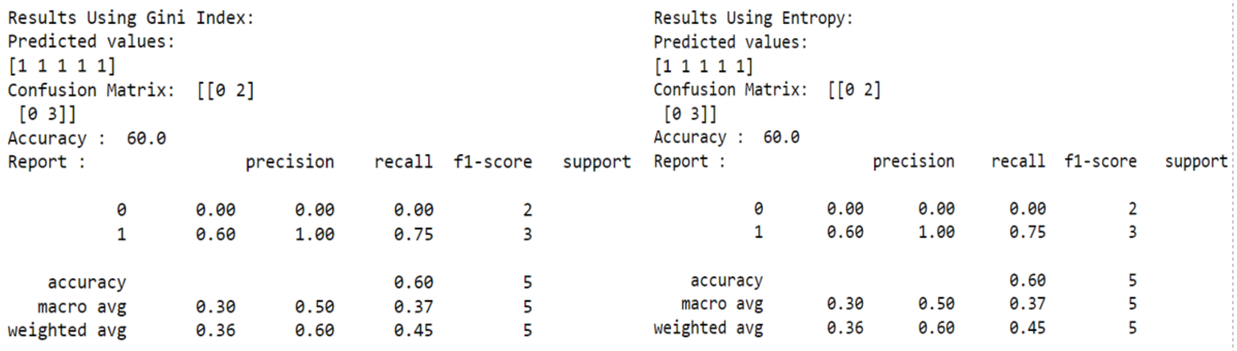


Figure 3: Decision Tree of Traffic Commute



```
Results Using Gini Index:
Predicted values:
[1 1 1 1 1]
Confusion Matrix:  [[0 2]
 [0 3]]
Accuracy :  60.0
Report :              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.60      1.00      0.75         3

    accuracy                           0.60         5
   macro avg       0.30      0.50      0.37         5
weighted avg       0.36      0.60      0.45         5
```

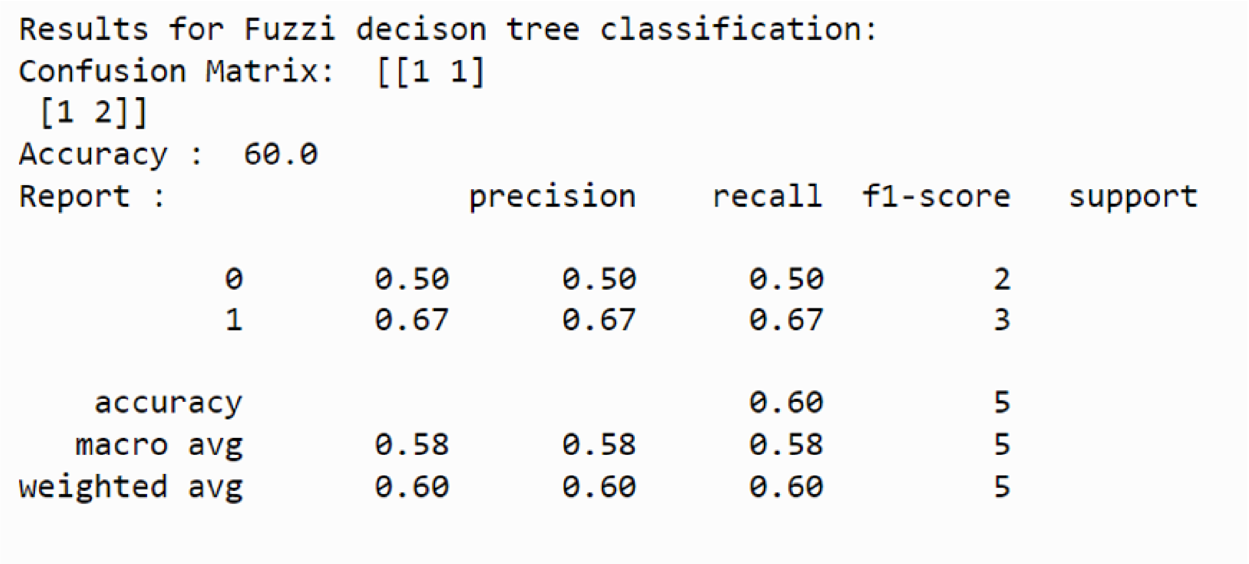
```
Results Using Entropy:
Predicted values:
[1 1 1 1 1]
Confusion Matrix:  [[0 2]
 [0 3]]
Accuracy :  60.0
Report :              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.60      1.00      0.75         3

    accuracy                           0.60         5
   macro avg       0.30      0.50      0.37         5
weighted avg       0.36      0.60      0.45         5
```

Figure 4: Entropy and Gini index

Fuzzy Decision Tree

Membership in a partitioned set (and consequently in a node) for a training example in ID3 is binary. Fuzzy sets and fuzzy logic are applied to fuzzify the attributes x1, x2, and x3.

The method for fuzzy sets is to define the fuzzy terms required to create the tree. Consider the fuzzy variables temperatures, wind, and traffic jams. Assume that a person can only decide to drive or not; this decision comes down to yes or no. Figure 5 shows the confusion matrix for the fuzzy decision tree. Here we have an accuracy of 60 with a 0.60 weighted average.



```
Results for Fuzzi decison tree classification:
Confusion Matrix:  [[1 1]
 [1 2]]
Accuracy :  60.0
Report :                precision    recall  f1-score   support

           0       0.50      0.50      0.50         2
           1       0.67      0.67      0.67         3

    accuracy                           0.60         5
   macro avg       0.58      0.58      0.58         5
weighted avg       0.60      0.60      0.60         5
```

Figure 5: Confusion Matrix for Fuzzy decision tree

References List

Banakar, A., Zareiforoush, H., Baigvand, M., Montazeri, M., Khodaei, J., & Behroozi-Khazaei, N. (2016). Combined application of decision tree and fuzzy logic techniques for intelligent grading of dried figs. *Journal of Food Process Engineering*, *40*(3), e12456. https://doi.org/10.1111/jfpe.12456

Gopal, M. (2019). *Applied machine learning* (1st ed.). McGraw-Hill Professional.