Homework – DSC 540 Week Three

Colton Proctor

Grand Canyon University

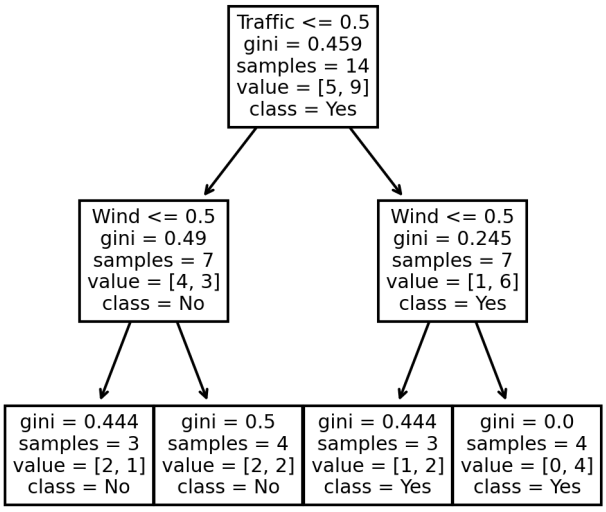
The given data set contains a small amount of data which is used to predict whether a person will be driving that day. The three variables that are used to determine the binary result are the temperature, amount of wind, and the density of traffic. The method chosen to predict this is a decision tree. To choose the optimal root node for the tree the information gain based off the Gini impurity was calculated for each of the three independent variables.

Information gain for temperature is: 0.02922256565895487

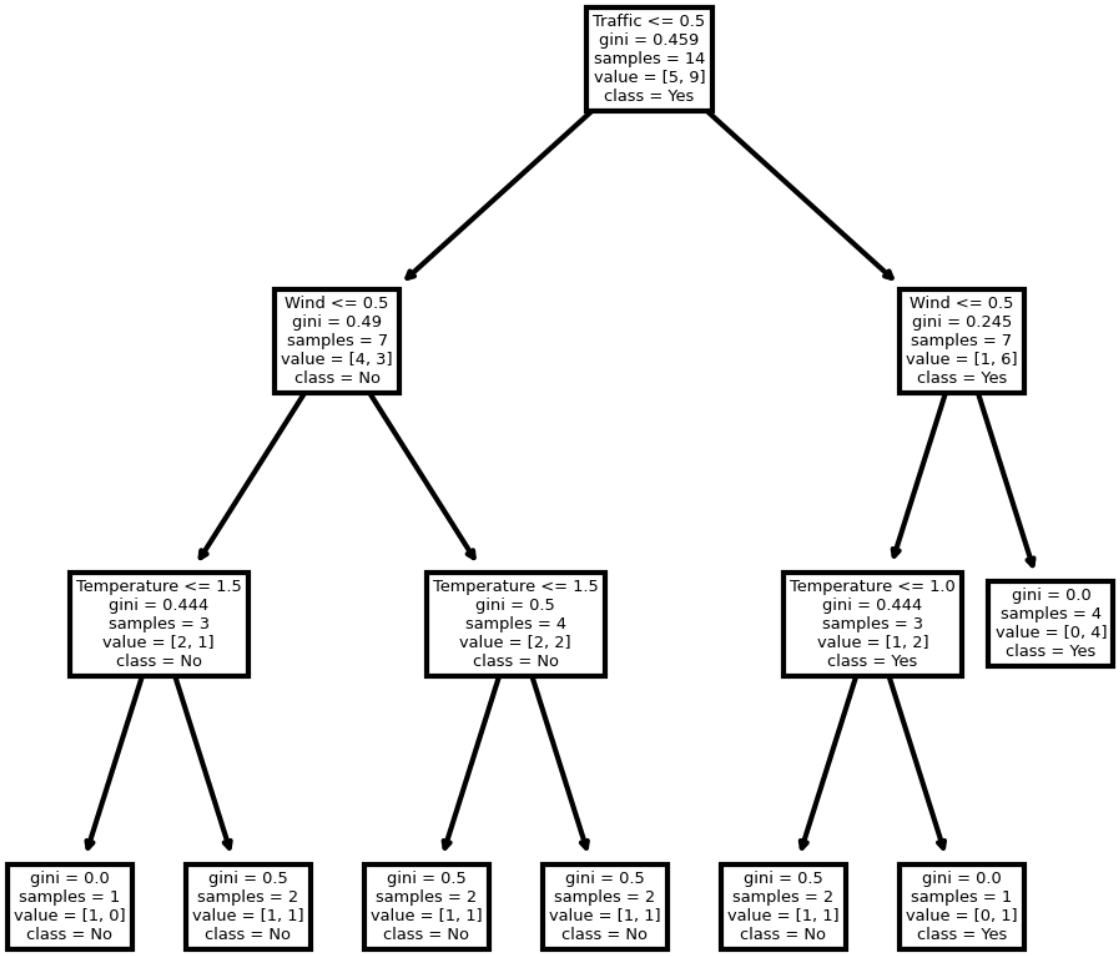
Information gain for wind is: 0.04812703040826949

Information gain for traffic is: 0.15183550136234159

Based on this calculation traffic was chosen as the variable to do the split at the root node, since it had the highest information gain. A partial decision tree was then created using this root node. It is shown below:



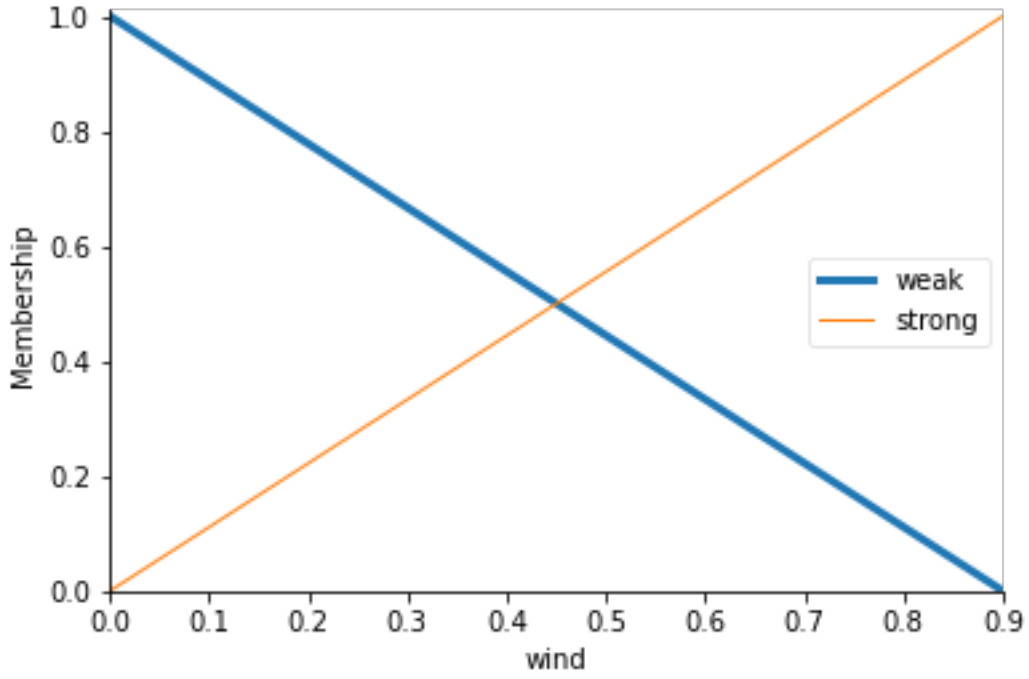
This tree was then further grown to determine whether it was possible to achieve a pure tree which contained only Gini values of 0. That tree is shown as follows:

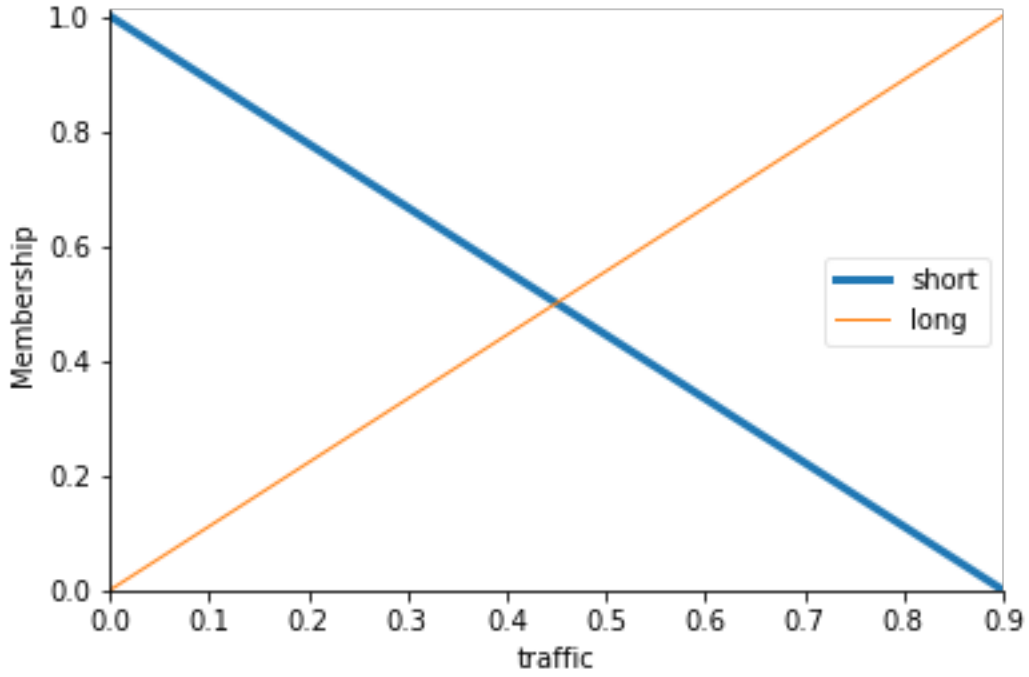


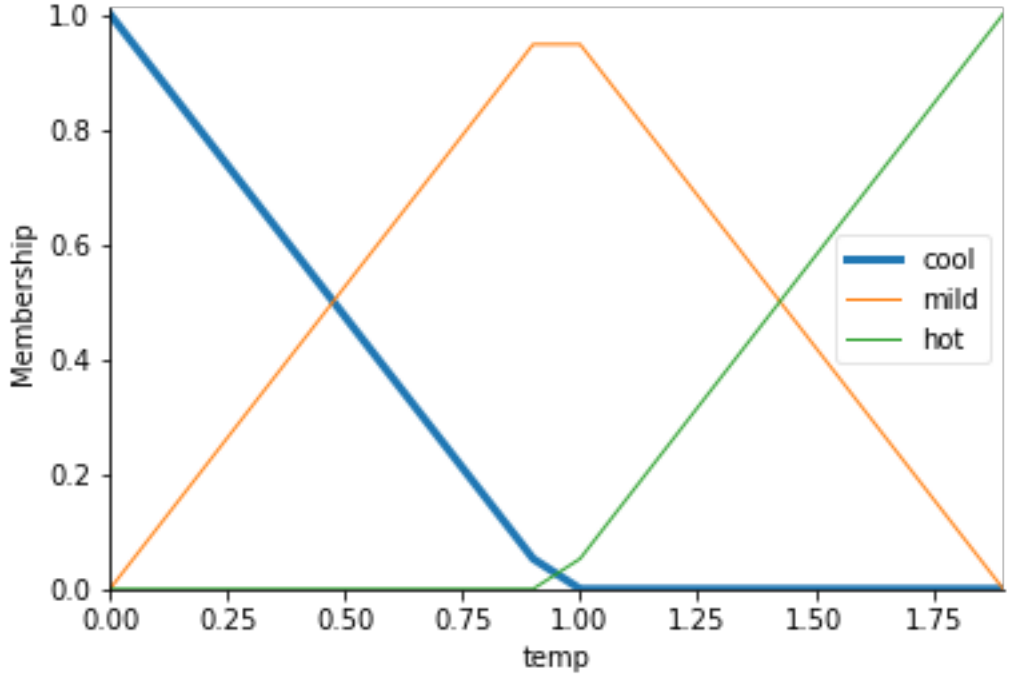
As shown in this tree it is not possible for this data set to be split into a tree that has no Gini

impurities. The tree is not able to grow larger than the one shown due to the number of variables that are available to split the tree on. It also has to do with the way that the variables are distributed. Since they contain values that are disparate for the other variables among each class that that is used to split, it causes groupings which cannot be perfectly balanced.

To fuzzify this this decision tree the decision was made to divide the three different independent variables into smaller subsets of themselves. They were encoded from as numerical values corresponding to their actual labels and then split into likelihood functions. These functions indicated their membership in the category of either driving or not driving. The graphs for the likelihood functions of each of the three variables are shown below.

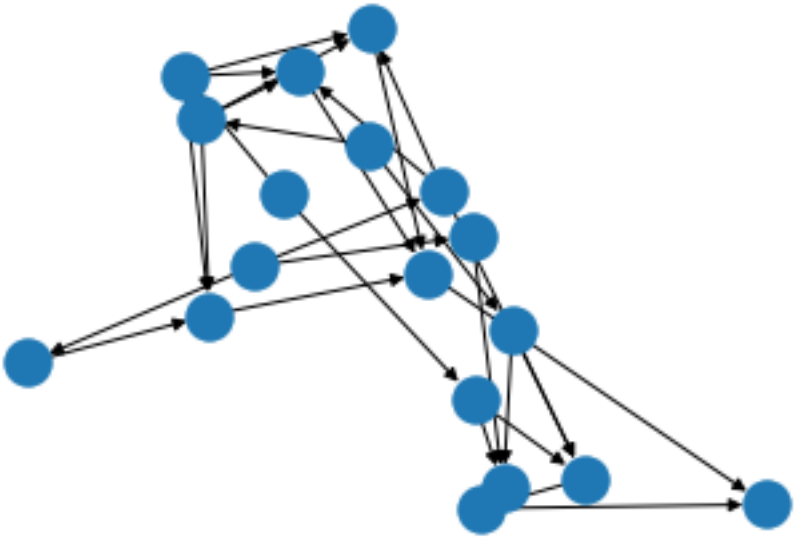




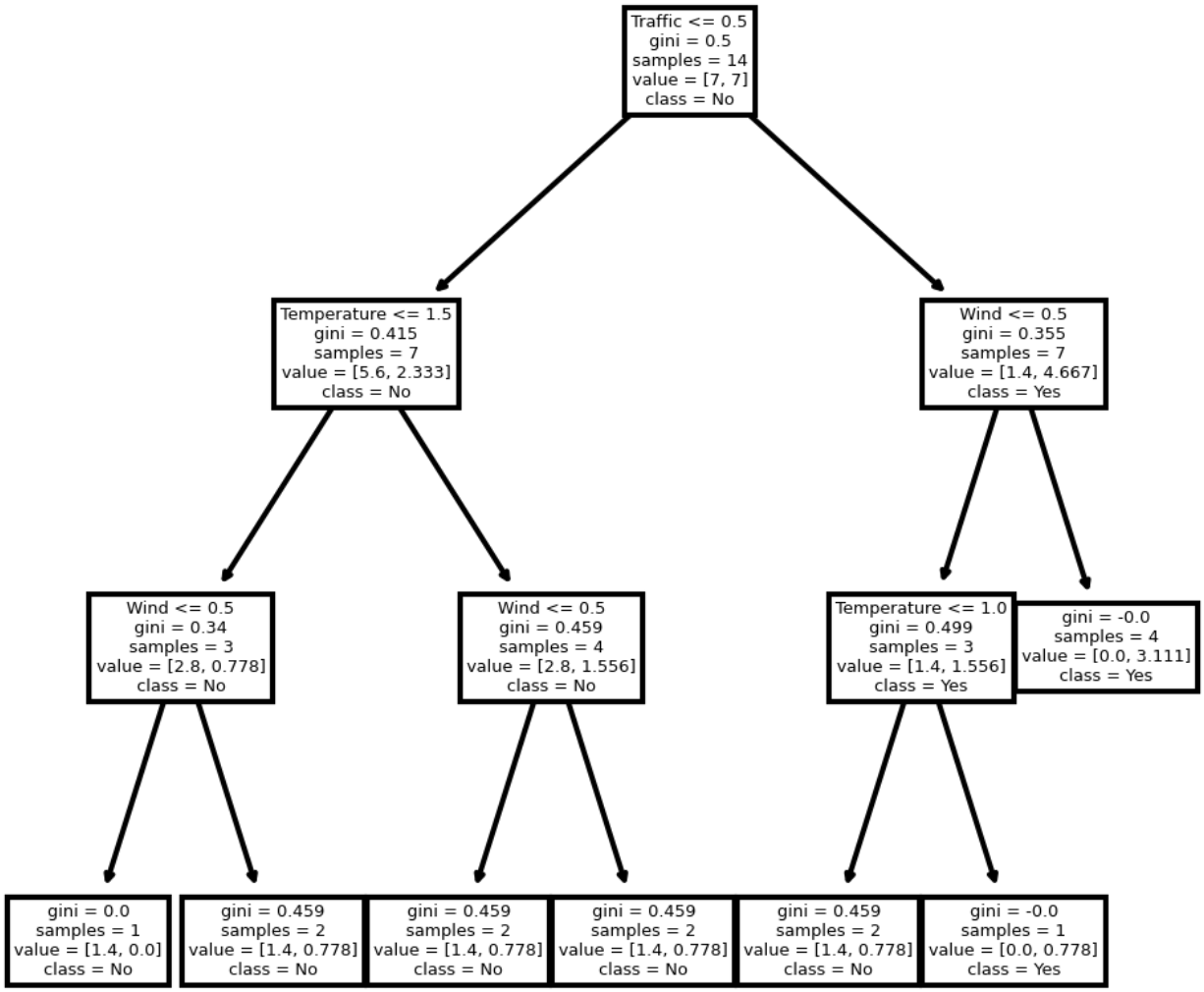


I then used these fuzzy classifications as sets of rules that corresponded to the 14 entries in the original data set. For each entry in the data set a rule was created, and then those rules were

combined into a control system which evaluated the result of whether the decision would be to drive or not. This control system is represented as a directed graph shown below:



The decision was made to still use Traffic as the root node for the fuzzy decision tree. This is because it ended up with the best performing tree out of all three of the choices. With fuzzification of the data the Gini values were able to be lowered slightly which optimized the tree over the standard version. A graph showing the tree created by the fuzzy model is shown below.



This tree utilized the concepts found in the article on the combined application of decision trees and fuzzy logic by implementing the antecedent, consequent ideology. Each of the

independent variables was encoded as an antecedent, which were used to determine the consequent of whether to drive. This was facilitated through the creation of rule sets which corresponded to the actual entries in the data set. For each data set the values for each antecedent was used in correlation with its antecedent value. This created the situation where the control system had access to each decision that was made in the data set.

Works Cited

Banakar, A., Zareiforoush, H., Baigvand, M., Montazeri, M., Khodaei, J., & Behroozi, K. N. (2017). Combined Application of Decision Tree and Fuzzy Logic Techniques for Intelligent Grading of Dried Figs. *Journal of Food Process Engineering*, *40*(3), n/a-N.PAG. https://doi-org.lopes.idm.oclc.org/10.1111/jfpe.12456

Gafa, C. (2020, May 13). *Fuzzy inference system implementation in Python*. Medium. Retrieved December 2, 2021, from https://towardsdatascience.com/fuzzy-inference-system-implementation-in-python-8af88d1f0a6e.