**Assignment 5: Technical Report**

**David D. Bui**

**Grand Canyon University**

**DSC-540**

**Professor Aiman Darwiche**

**December 1rd, 2021**

**GitHub:**

**https://github.com/DouglasBui/GCU/tree/main/DSC-540/Assignment5**

**Part 1: Decision Tree**

The dataset this project involves the decision whether an individual drives their car or takes the bus; the data contains 14 observations with categorical features such as Temperature, Wind, and Traffic-Jam. There is a process of creating a decision tree that utilizes information gain as the deciding factor for root placement. Calculating the information gain revealed that Traffic-Jam had the highest return, followed by wind. This resulted in Traffic-Jams being placed as the root node, and the initial criteria for splitting, and each additional split followed the descension of features by their IG. The goal behind this project was to isolate targeted “yes” instances into the last branching leaf node, and in doing so forming a pure node of just one classification. This Gini Index measurement is the probability of selecting an incorrect classification within a node. With this measurement it shows that there are only 3 leaf nodes that are pure, the remaining 4 have a probability of 50% to hold a misclassification.

When analyzing the data there is evidence that shows that these misclassifications are not the fault of the model. There are instances within data where the features hold identical elements that result in conflicting outcomes. This creates an uncertainty within the data, where mathematical logic cannot make a distinction. This pulls our data into the analysis realm of probability. The issue with the current model is that it is structured for logical distinction and not probability, making the goal of only pure leaf nodes impossible.

Part 2: Fuzzy Decision Tree

In handling the shortcomings of the previous decision tree, a fuzzy decision tree can be implemented in attempt to improve the accuracy of classification of the model. It does this by making use of Fuzzy Logic to smear the classification boundary. A rather interesting approach was implemented to deal with the strict categorization of the features. Each feature had its data embellished in a manner that accommodated Fuzzy Rules. To explain, the temperature attribute was broken down into a numerical fashion that reflected a human’s understanding of the terms “hot, mild, and cold.” All while keeping the initial element terms the same. This process utilizes randomization to allow for logical distinction to take place. This was performed on each attribute before placing them though membership functions in the act of fuzzification. The output of these membership functions resulted in a scaled interpretation of the embellished data that mimicked the data’s original state. These resulting arrays of data were joined and placed into the same decision tree function that was used in part 1. In doing so created a fuzzy decision tree, in which comparative evaluation can take place between old and new.

An accuracy scoring function compared the two trees and revealed that the previous decision tree yielded 71.43% accuracy, while the fuzzy tree showed an improvement with an accuracy of 92.86%. The fuzzy tree model was run dozens of times and its accuracy score never changed, which shows that the randomly embellished data when pushed through the membership functions with fuzzy ruling in place didn’t have on impact on the initial terms and elements, showing a crisp output that was angled towards finding unity. These outputs were then tested to find a new root node that held the highest Information gain, which showed Temperature as the new leading node.

A visualization of the fuzzy tree shows the process of how splitting took place within the tree function. It focused on the feature that held the highest Information gain (IG), splitting it continuously until that feature reached unity and resulting in a pure leaf node, and then descended the feature IG order. It should be noted that the last bottom right node always resulted in a Gini Index of .05. This is because the last feature, even though reaching unity, still contained conflicting results. There were a multitude of contradictions in the data, but now only one remained. This could be changed by altering the ruling of the membership functions, but I wasn’t able to accomplish this.

After the tree visualization I wanted to show a more descriptive transformation of the decision tree into a fuzzy decision tree and its membership function. For the membership function I attempted to model its unity classification process, but I couldn’t show a trapezoid function, instead I just displayed a triangular model. To explain this triangular model into a trapezoid one the ‘Average’ line should contain a plateau at its peak where it reaches unity. The next set of models show the effects of fuzzification a decision tree. I used Principal Component Analysis to shrink the data’s feature space into just 2 features to execute this comparison. I created a decision tree made from the embellished data that reflects a model of perfect classification, and then the same model with fuzzification applied to it. A complicated explanation is that the original decision tree is not exactly illustrated here as it is in our evaluation, this is because this method of comparison required PCA, which enacts its own form of data embellishment that allows for logical distinction, resulting in this displayed form of a perfected sklearn model. What we capture here is an understanding that our fuzzy tree model is expanding its classification boundary of the original decision tree model to reflect closer to a perfected model.

In conclusion, there is evidence behind the utility of applying Fuzzy Logic to a decision tree that contains uncertainty. Without the need to entirely change the model for one that measures for probability. The results show a clear improvement in its accuracy of classification. There is still room for improvement, and this improvement could come in the form of altering the Fuzzy Rules of each set membership. Nevertheless, the results came closer to a perfected tree.

References

Es-sabery, F., & Hair, A., (2020), A MapReduce C4.5 Decision Tree Algorithm Based on Fuzzy Rule-Base System, Taylor & Francis, <https://doi.org/10.1080/16168658.2020.1756099>

Fuzzy Decision Trees (Artificial Intelligence), what-when-how, <http://what-when-how.com/artificial-intelligence/fuzzy-decision-trees-artificial-intelligence/>

Lutes, J., (2020), Entropy and Information Gain in Decision Trees, towardsdatascience, <https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293>