Rachael Keyes

PRO Points Project

1. **Describing the data**
   The outcome variable for this dataset would be whether or not an individual has diabetes, the value 1 meaning they are likely to have it and 0 meaning not.
   The features for prediction are glucose, BMI, blood pressure, diabetes pedigree function, and age.
   How the outcome variable relates to the features and what insights can be gained from analyzing the data.
2. **Finding the best value for minimum split**
   After analyzing the changes in accuracy via the confusion matrix, I've settled on 50 to be the minimum split value for this tree. Using this value gave me an correct classification rate of approx. 75%, and good visibility of most leaf nodes in the tree. Although 74% isn't ideal, other minimum split values like 25 and 40 gave me correct classification rates of 59% and 62%, and significantly impaired my ability to read the tree because of overcrowding of the leaf nodes. While working on past decision trees from this semester I haven't used a value as high as 50, but it seems to be working better than the lower values for this particular data set.
3. **Find the Node with the lowest and highest probability**
   The leaf node with the highest probability is node #27, which has a value of (0.086, 0.914),and the leaf node with the lowest probability is node #5, which has a value of (1.0, 0.0). Node #27 indicates that individuals with glucose levels greater than 143.5, diabetes pedigree function less than 0.31, and who are younger than 31.5 are more likely to have diabetes. Node #5 indicates that individuals who are younger than 28.5 years old, have glucose levels less than 127.5, and have a BMI less than 31.4 are not likely to have diabetes.
4. **Provide at least four examples of data points and use the tree to predict the outcomes**
   Example 1:
   This person
   -has glucose of 130
   -is 26 years old
   -has a BMI of 34
   -Diabetes Pedigree Function is 0.431
   Outcome:
   Node #10
   Based on the values [0.65, 0.35], there is a 35% chance that this individual has diabetes

   Example 2:
   This person
   -has glucose of 145
   -is 50 years old
   -has a BMI of 37
   -Diabetes Pedigree Function is 0.567

-Blood pressure of 94.0
Outcome:
Node #28
Based on the values [0.667, 0.333], there is a 33% chance that this individual has diabetes

Example 3:
This person
-has glucose of 155
-is 32 years old
-has a BMI of 37
-Diabetes Pedigree Function is 0.592
-Blood pressure of 91.0
Outcome:
Node #27
Based on the values [0.086, 0.914], there is a 91.4% chance that this individual has diabetes

Example 4:
This person
-has glucose of 128
-is 65 years old
-has a BMI of 26
-Diabetes Pedigree Function is 0.298
-Blood pressure of 90.7
Outcome:
Node #12
Based on the values [0.933, 0.067], there is a 6.7% chance that this individual has diabetes