### HOMEWORK 2 DECISION TREE



### MY PROBLEM



# CLASSIFY WHETHER OR NOT A GIVEN PERSON HAS A HEART DISEASE

### **ATTRIBUTES**



- 1. Age
- 2. Sex
- 3. Weight
- 4. Activity level (exercise)
- 5. Smoking
- 6. Family heart disease history

### **MY RULES**

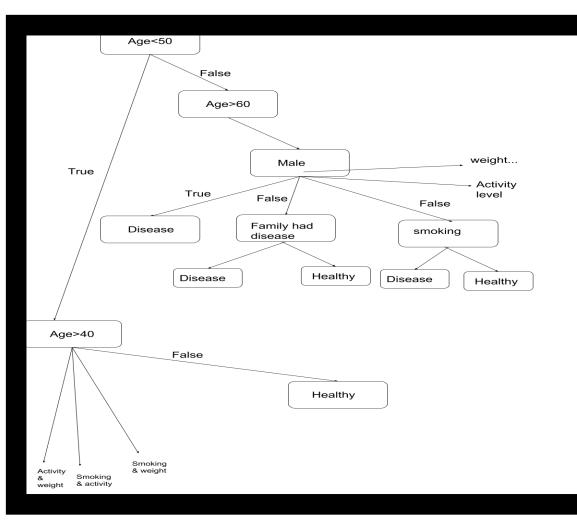
#### Disease:

- 1. If age > 60 AND Male
- 2. If 50 < age < 60 \*If Family heart disease history
  - If Smoking
  - If Weight > 100
  - If Activity level = 0
- 3. If 40 < age < 50
  - If Activity level = 0 AND Weight > 100
  - If Smoking AND Activity level = 0
  - If Smoking AND Weight > 100
- 4. If Family heart disease history AND Male

Healthy:

1. All other possibilities





### **MY TREE**

#### $sex \le 0.5$ gini = 0.499samples = 500value = [239, 261] class = disease False True family $hd \le 0.5$ age ≤ 40.5 gini = 0.354gini = 0.319samples = 244samples = 256value = [51, 205]value = [188, 56]class = healthy class = disease age ≤ 59.5 age $\leq 41.0$ gini = 0.0gini = 0.0gini = 0.489gini = 0.489samples = 112samples = 136samples = 132samples = 120value = [112, 0]value = [0, 136]value = [76, 56]value = [51, 69]class = healthy class = disease class = healthy class = disease weight $\leq 99.5$ age $\leq 60.5$ gini = 0.0gini = 0.0gini = 0.257gini = 0.186samples = 66samples = 43samples = 66samples = 77value = [66, 0]value = [43, 0]value = [10, 56]value = [8, 69]class = healthy class = healthy class = disease class = disease Unnamed: 0 ≤ 115.0 age ≤ 59.5 smoking $\leq 0.5$ gini = 0.0gini = 0.408gini = 0.083gini = 0.48samples = 49samples = 46samples = 28samples = 20value = [0, 49]value = [8, 12]value = [2, 44]value = [8, 20]class = disease class = disease class = disease class = disease gini = 0.444gini = 0.408gini = 0.32gini = 0.0gini = 0.0gini = 0.0samples = 12samples = 8samples = 7samples = 39samples = 25samples = 3value = [8, 4]value = [0, 8]value = [2, 5]value = [0, 39]value = [5, 20]value = [3, 0]class = healthy class = disease class = disease class = disease class = disease class = healthy

## PRODUCED TREE



### COMPARISON

Of the hand made, and generated trees



### **PRIORITIES**

#### Hand-generated tree priorities

Age > Sex > Weight/Activity level/Smoking/Family history

### Code-generated tree priorities

Sex > Age/Family history > Age > Weight/Age > Smoking/Age



- I tried to put emphasis on the attributes such as sex, age, and family heart disease history, and it was reflected in the code-generated tree, by being on top levels of the tree, and having large impact on the result.
- The secondary attributes smoking/weight/activity level, were used for creating rules for more specific classification. By creating these statistics, I hoped to reduce underfitting, and I believe I largely succeeded.
- The code-generated tree also put the same amount of priority to these secondary statistics as I did, since they appear at 4th+ level of the tree.

### CONCLUSION



94.4%

It seems that the rules that I was thinking about during the creation of 'right' data, were very similar to the rules which code-generated tree used.

code-generated decision tree was successful at identifying the trends in data, and creating the rules that dictate how each person should be classified

# Thanks!

