**CSC 177: Data Analytics and Mining**

**Linear Regression Project & Classification Tree Homework**

Team Challengers (23):

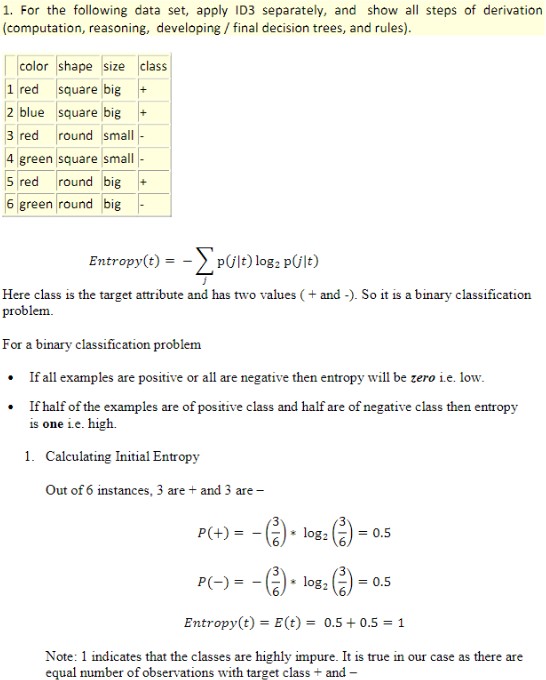
1. Srujay Reddy Vangoor

2. Vaibhav Jain

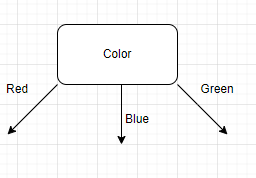
3. Bashar Allwza

4. Varun Bailapudi

5. Uddayankith Chodagam



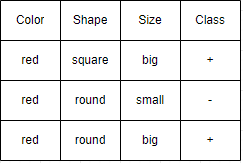
Color contains the highest Entropy Gain, so we use this as our initial splitting attribute. This results in the following decision tree:



Next, lets calculate the Entropy of red:

Gain = Entropy (Set) - I(Attribute)

Entropy Binary Equation: **E(a) = -p (+) \*log2(p (+)) – p(-)\*log2(p(-))**

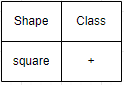
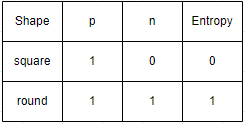
* a represents the attribute.
* p represents the probability of + or - of the class attribute.
* I represents Average information

E(red) = -2/3 log2(⅔) -(⅓) log2(⅓) = 0.91829583

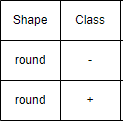
E(red) ≈ 0.92

Next, lets calculate the entropy of each following attributes:

# (Shape)

E (red, shape=square) = -(1) log2(1/1) - 0 = 0 E(red, shape=round) = 1(½)log(½) - (½)log2(½) = 1

I(shape) = (⅓) \* 0 + (⅔) \* 1 ≈ 0.66

Gain = E(red) - I(shape) Gain = 0.92 - 0.66 = .26

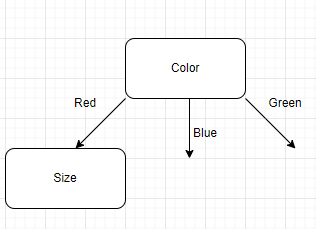
# (Size)

I(Size) = (⅔) \* 0 + (⅓) \* 0 = 0

The Entropy of both big and small in this case is 0, so the total Information I is 0 as well.

Gain = Entropy (Set) - I(size) Gain = 0.92 - 0 = 0.92

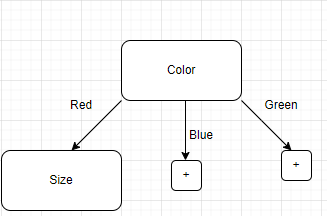
After calculating the Information Gain for both size and shape, we find that size has a higher Gain. Therefore, the size attribute will be the next node in our Classification tree.



If we take a look at the dataset again, there are a few important observations we must make:

* The entropy of blue is 0 and has no differentiation. Thus, the class attribute will always be the same regardless of the color, shape, or size based on the data we are provided for this color.
* The entropy of green is 0 and has no differentiation. Thus, the class attribute will always be the same regardless of the color, shape, or size based on the data we are provided for this color.

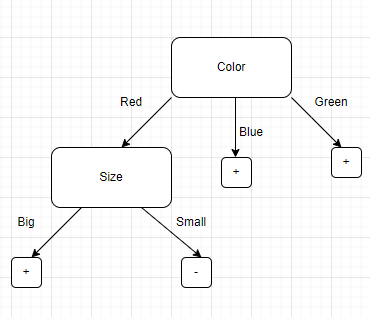
Because of the observations listed above, we can arrive at the following tree:



When we drill down into Size, then Shape, we can also make another important observation:

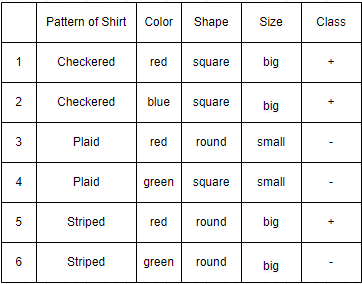
* The Entropy of (red, big) and (red, small) is also 0. These combinations also have no differentiation according to the dataset provided.

The observation listed above helps us to arrive at the final classification tree shown below:



Part 2)

If we Were to add another attribute, then it is possible that the above tree is different. Adding an attribute (‘Pattern of Shirt’) with a higher Gain than color, then it would take the place of color as the root node. In other cases, the ‘Pattern of Shirt’ Attribute may take the spot of any of the nodes in the tree depending on the Gain.

Let's work through the following example: The

Entropy(t) = 0.5 + 0.5 = 1

# Pattern of Shirt

E (Pattern = Checkered) = -(1) log2(1) - (0) log2(0) = 0 E(Pattern = Plaid) = -(0)log2(0) - (1)log2(1) = 0 E(Pattern = Striped) = -(1/2)log2(1/2) - (1/2)log2(1/2) = 1

Avg Entropy = (2/6) (0) + (2/6) (0) + 2/6(1) Gain (Outlook) = 1- 0.3333 = 0.67

# Color

E (Color = red) = -(⅔) log2(⅔) - (⅓)log2(⅓) = .92 E (Color = blue) = 0

E (Color = green) = 0

Avg Entropy = (½)0.92 +(⅙) (0) + (2/6) (0) = 0.46

Gain (Outlook) = 1 - 0.46 = 0.54

# Shape

E (Shape = square) = -(⅔) log2(⅔) - (⅓)log2(⅓) = 0.92

E (Shape = round) = -(⅓) log2(⅓) - (⅔)log2(⅔) = 0.92

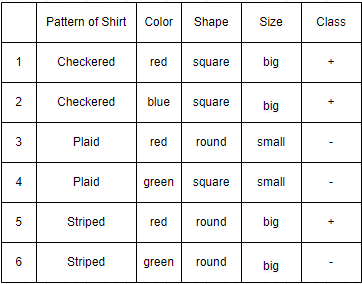
Avg Entropy = (½) (0.92) + (½) (0.92) Gain (Outlook) = 1 - 0.92 = 0.08

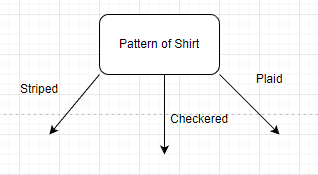
# Size

E (Size = big) = -(¾) log2(¾) - (¼)log2(¼) = 0.81

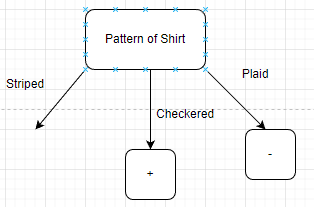
E (Size = small) = -(0 - 1) log2(1) = 0

Avg Entropy = (⅔) (0.81) - (⅓)(0) = 0.54 Gain(Outlook) = 1 - 0.54 = 0.46

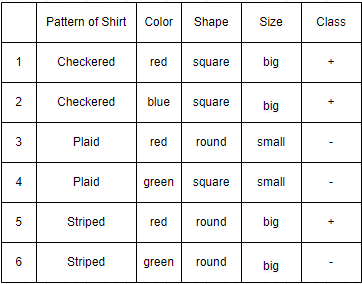
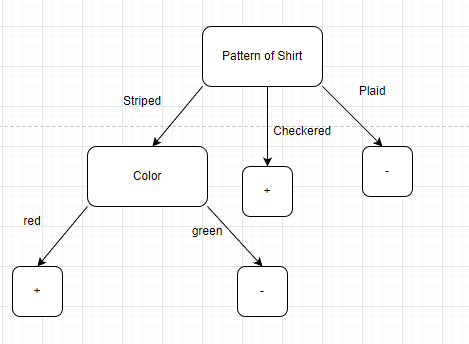
In this case, ‘Pattern of Shirt’ will be our first split.



Right away, we can observe that both Checkered and Plaid have no differentiation. Checkered will always be + and Plaid will always be - according to our dataset. We can show this off in the tree below:



If we continue with the IDM algorithm on the only path left available, Striped, we can see that color is the only other attribute with differentiation, and thus would be used as our next splitting attribute. Furthermore, this category can only be split into (Striped, red) or (Striped, green) both options which also have no differentiation. We can deduce the resulting tree as shown below:



This is one of many ways that adding a missing attribute could change the entire classification tree.

Yes, a scientist will make an impact on the manager and CEO in case they discover the new attribute and its influence in getting more reliable results valuable to the company.