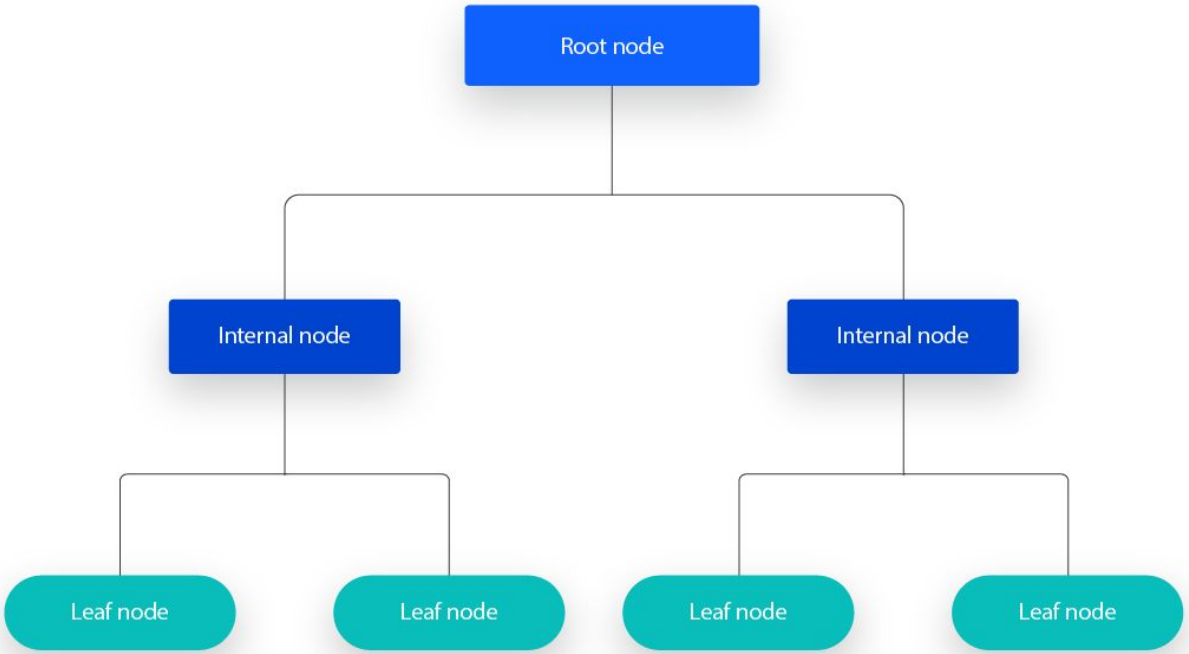
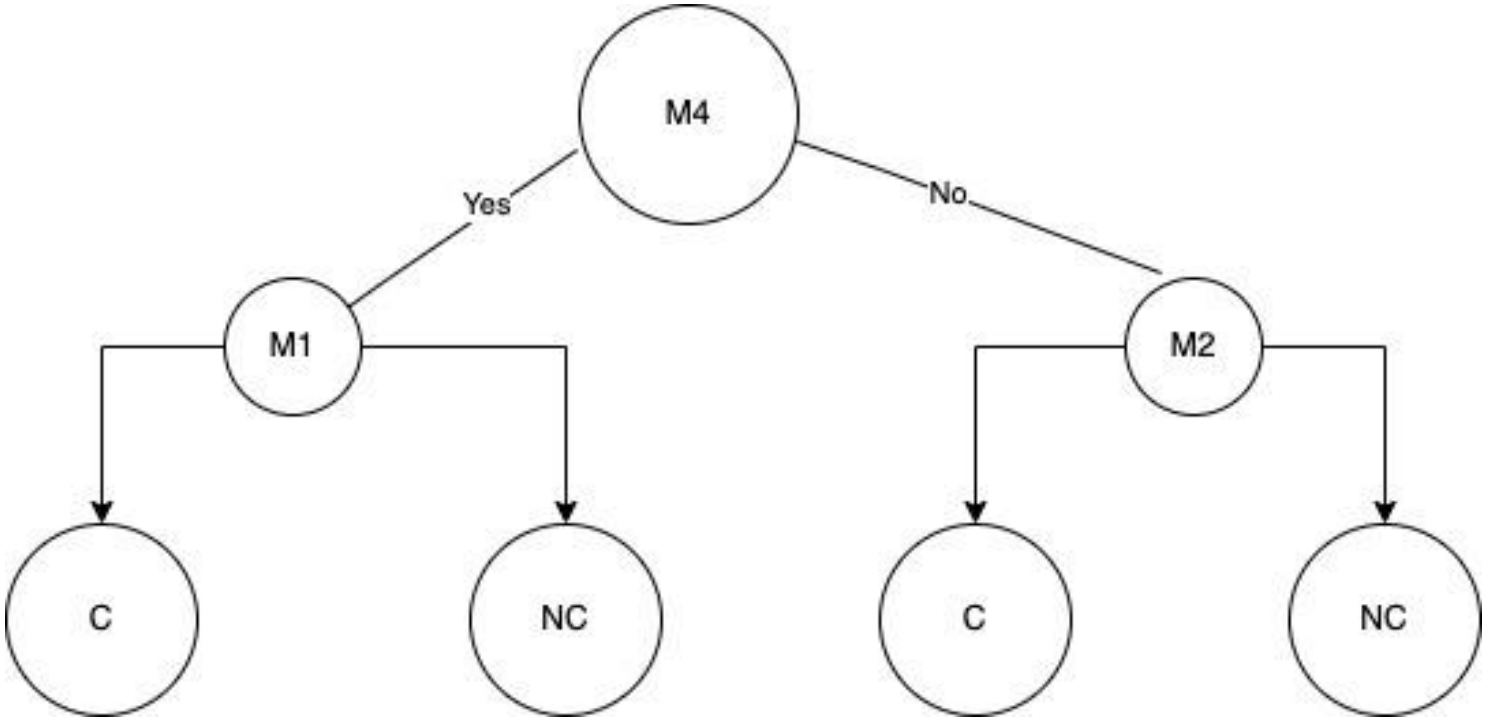
# Decision Trees

Anmol Mishra & Jessica Torrey

# Structural Overview



# Decisions



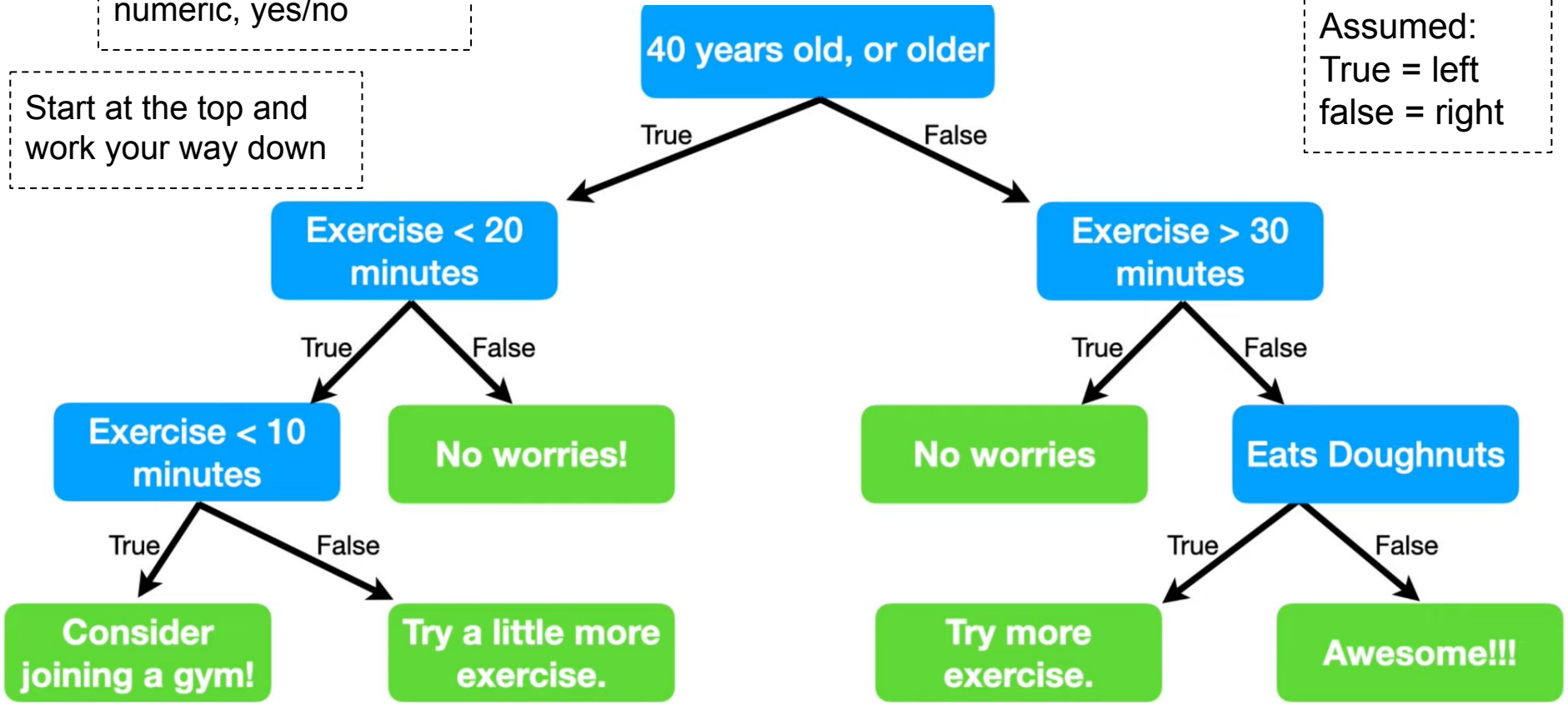
# Classification Tree Example

Classification into category = classification tree
Predicts numeric values = regression tree

Combines data types:
numeric, yes/no

Start at the top and
work your way down

Assumed:
True = left
false = right



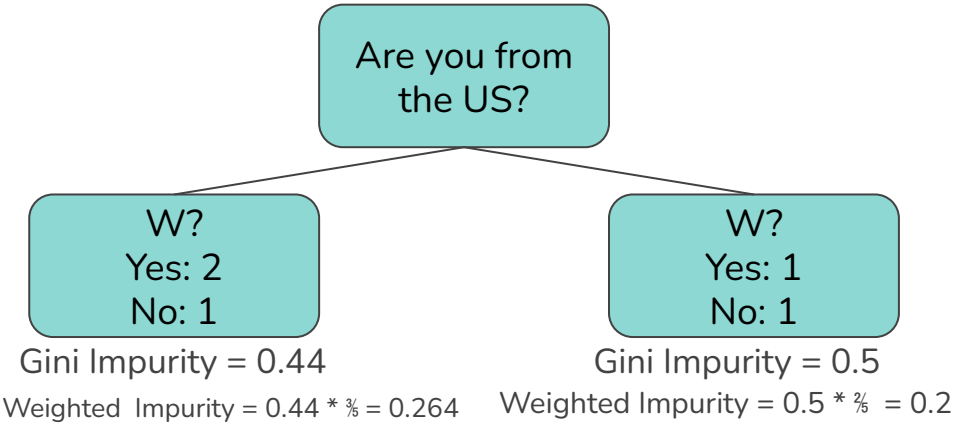
# Are you a girl in this class?

| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

**Building a decision tree given data**

# What should be at the top of the tree?

Category Choice : Country



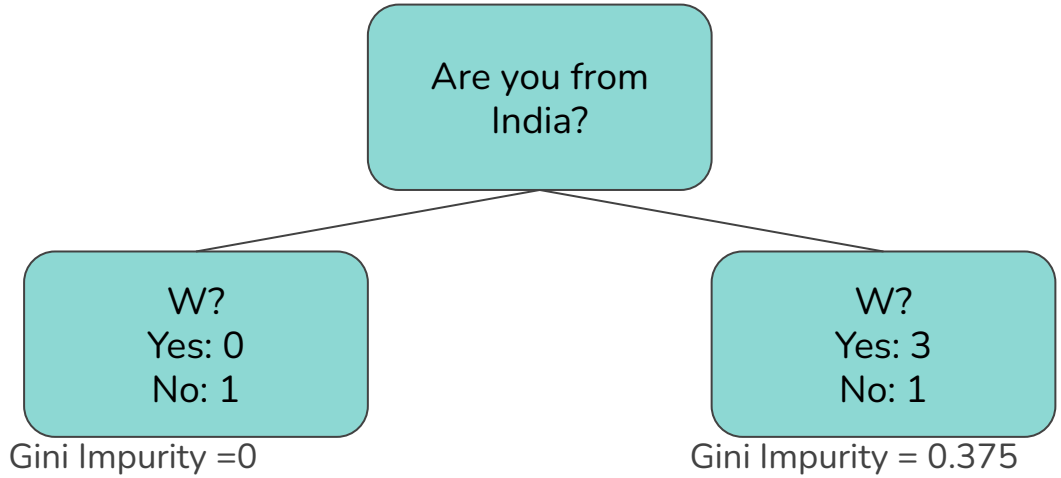

Total Impurity = 0.464

Look at each parameter to determine how well it predicts if you are girl in the class

| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**



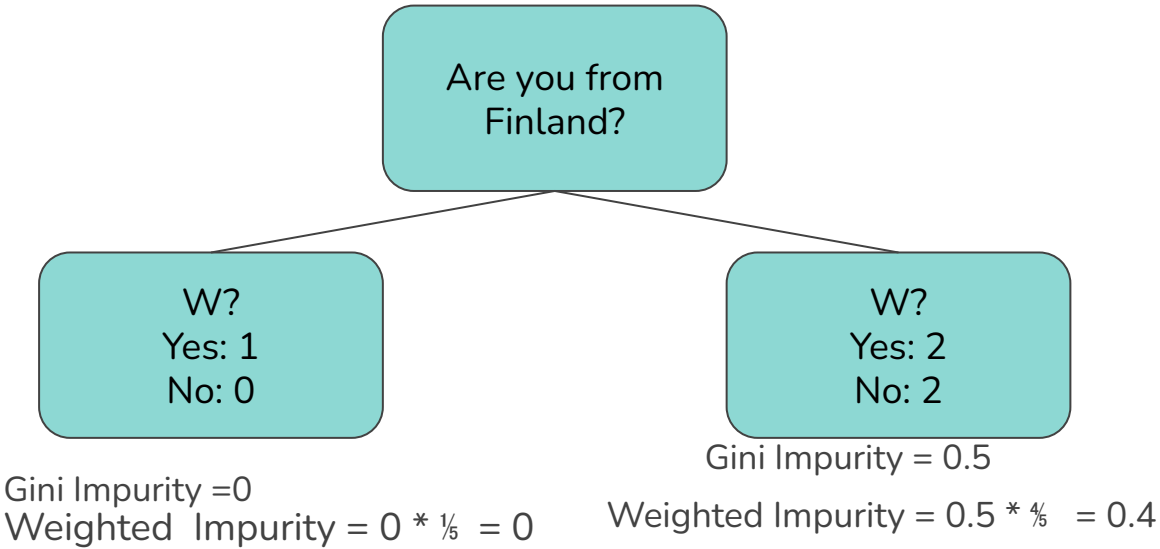

Weighted Impurity = 0 * ⅕ = 0

Weighted Impurity = 0.375 * ⅘ = 0.3

Total Impurity = 0 + 0.3 = 0.3

Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**




Total Impurity = 0 + 0.4 = 0.4

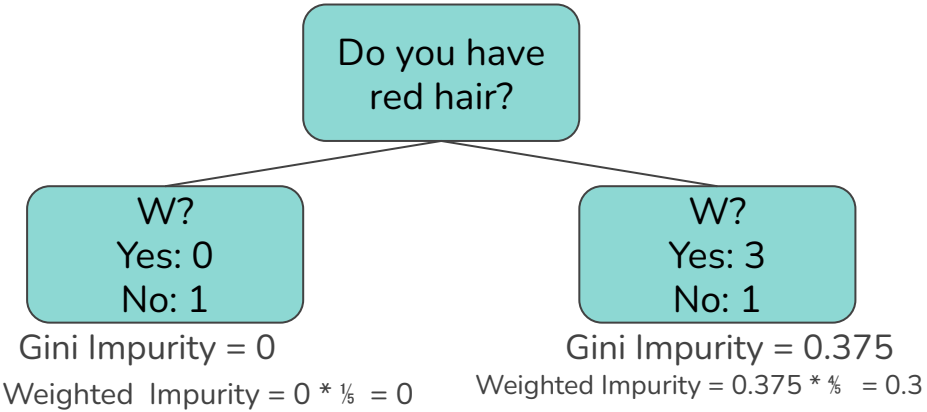
Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

# What should be at the top of the tree?

Category Choice:

Hair Color



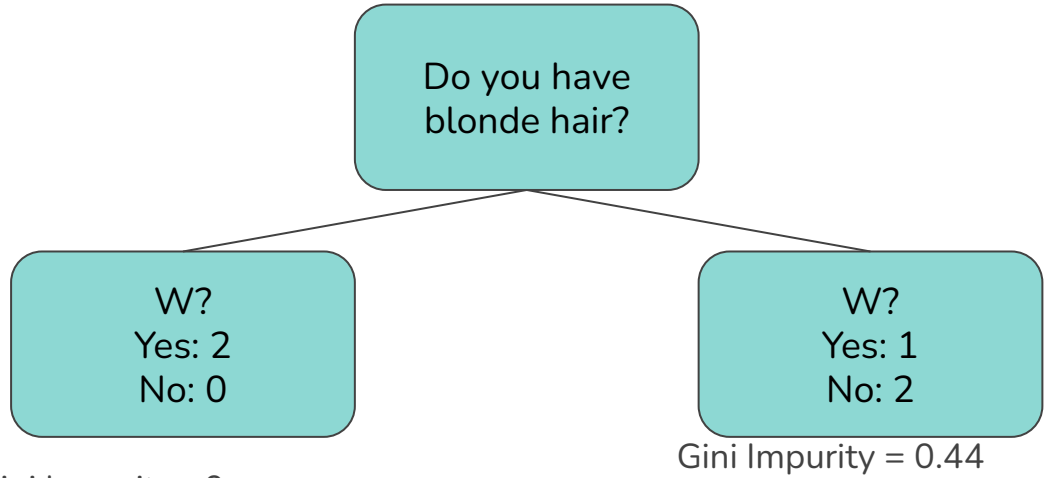

Total Impurity = 0.3

Look at each parameter to determine how well it predicts if you are girl in the class

| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

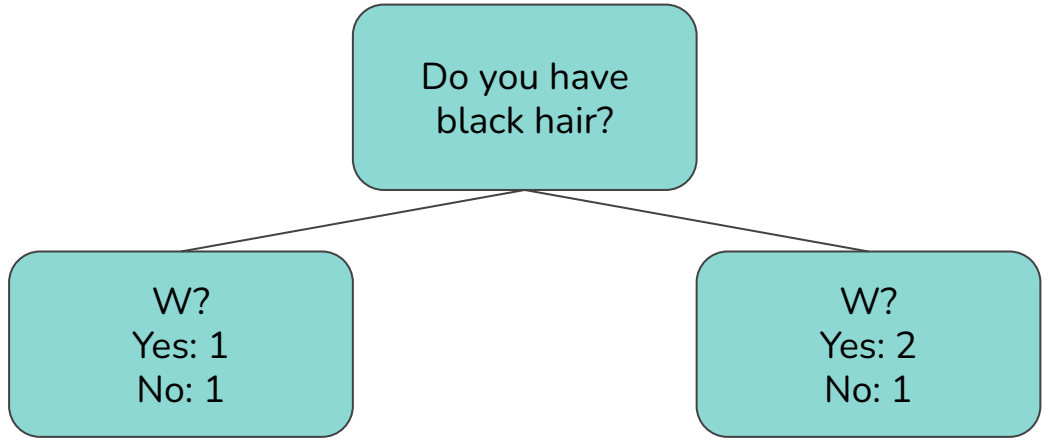
Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**




Total Impurity = 0 + 0.264 = 0.264

Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**




Gini Impurity =0.5

Weighted Impurity = 0.5 * 2/5 = 0.2

Gini Impurity = 0.44

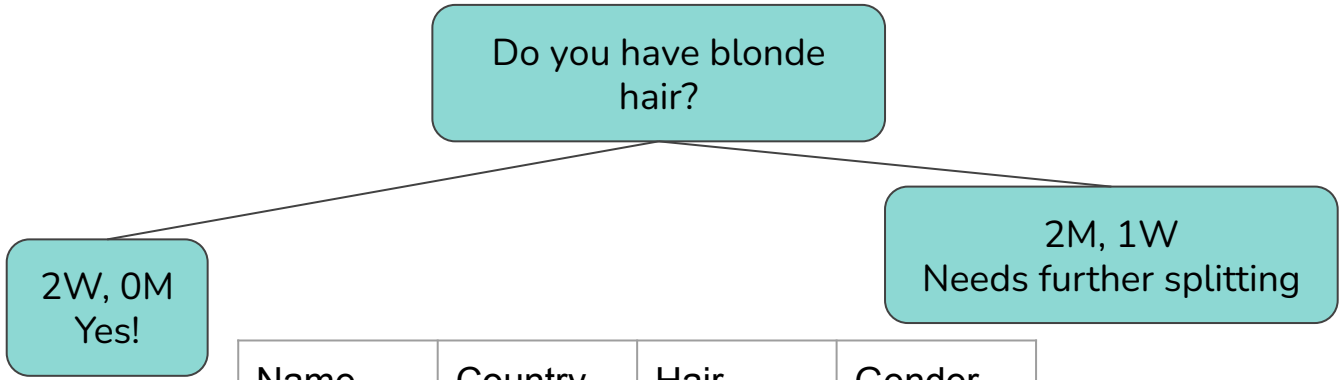
Weighted Impurity = 0.44 * 3/5 = 0.264

Total Impurity = 0.2 + 0.264 = 0.464

Gini Impurity for a Leaf = 1 - (the probability of "Yes")$^2$ - (the probability of "No")$^2$

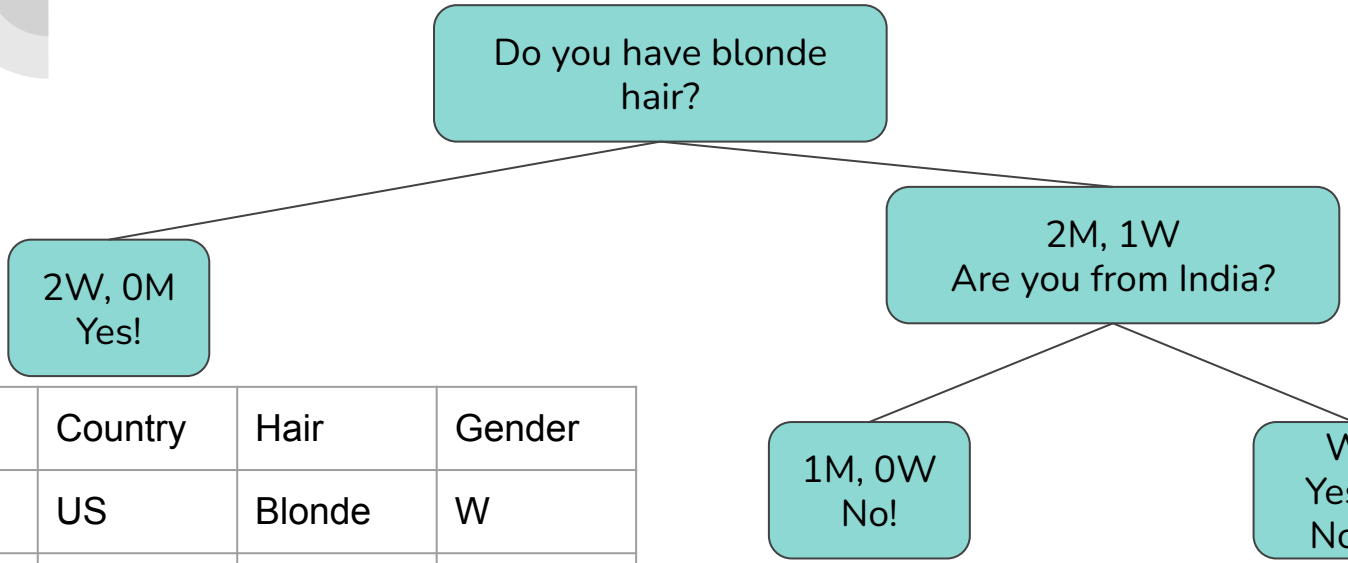
Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

# Are you a girl in this class?




| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

# Are you a girl in this class?




| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

4 questions possible at this stage -

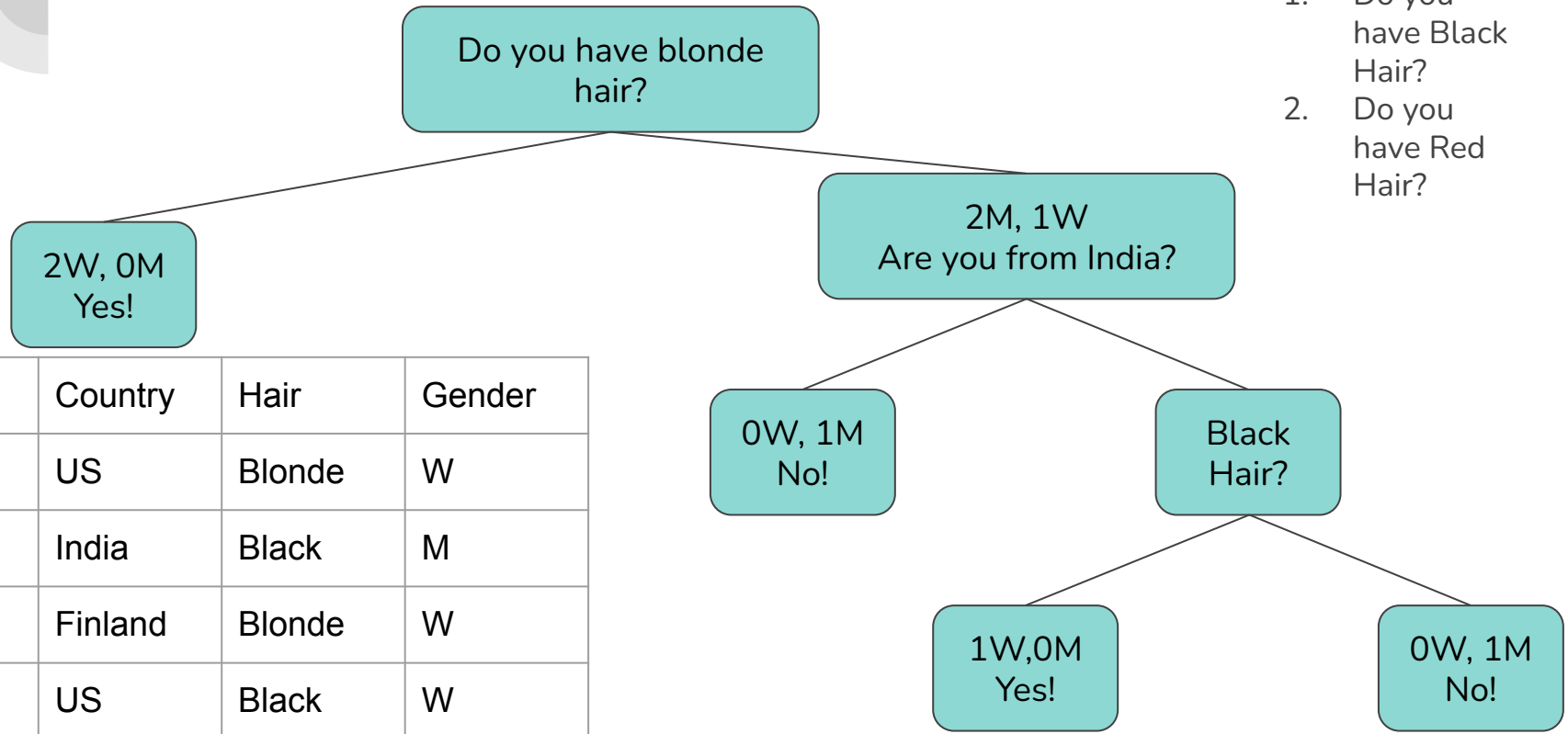
1. Are you from India?
2. Are you from US?
3. Do you have black hair?
4. Do you have red hair?

All 4 questions lead to same Gini Impurity.
Why?

# Are you a girl in this class?

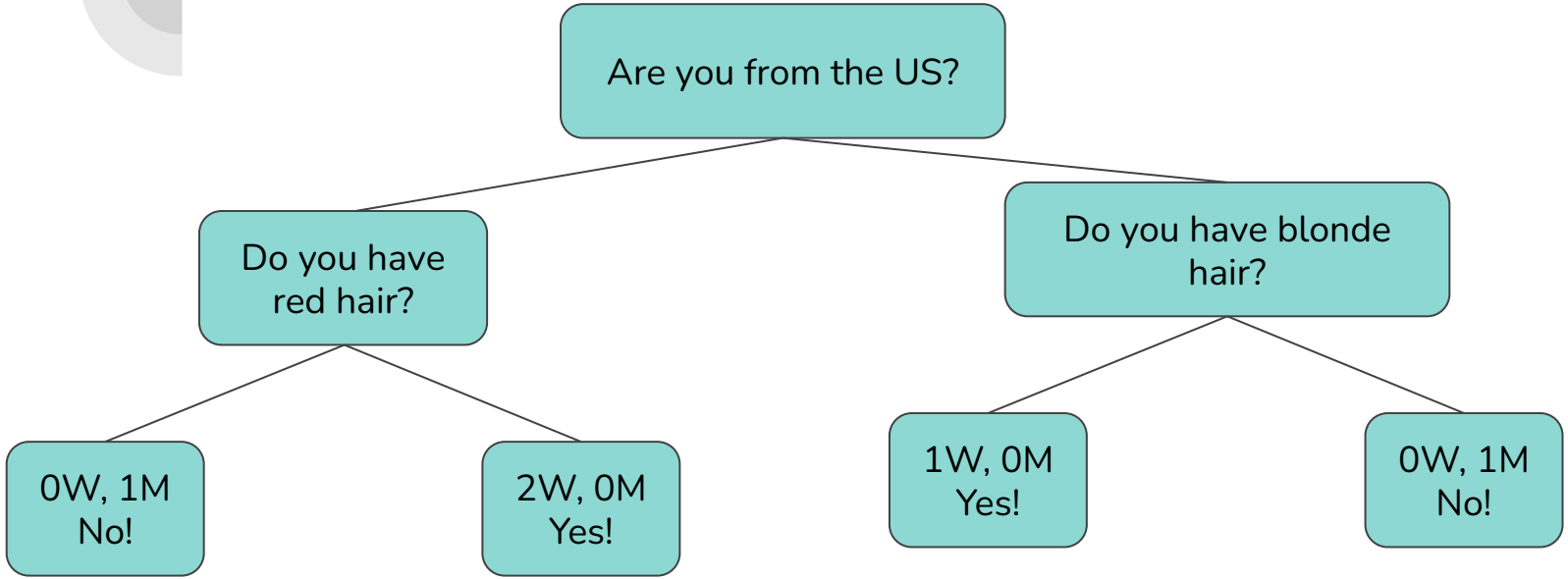
2 questions possible at this stage -

1. Do you have Black Hair?
2. Do you have Red Hair?




| Name | Country | Hair | Gender |
|---|---|---|---|
| Jessica | US | Blonde | W |
| Anmol | India | Black | M |
| Suvi | Finland | Blonde | W |
| Isabelle | US | Black | W |
| Jed | US | Red | M |

# Alternative Decision Tree




Which tree is better?

# Comments

There are two major metrics used to split Decision Trees -

1. Gini Index - Isolates classes earlier, but may lead to more depth
2. Entropy - More balanced trees, hence less depth

Gini Index is the more commonly used one, also the default metric used in the Scikit Learn Package on Python.

Question. Which metric would you think is more likely to generate the second tree amongst the trees we've shown?

# Decision Trees: Pros and Cons

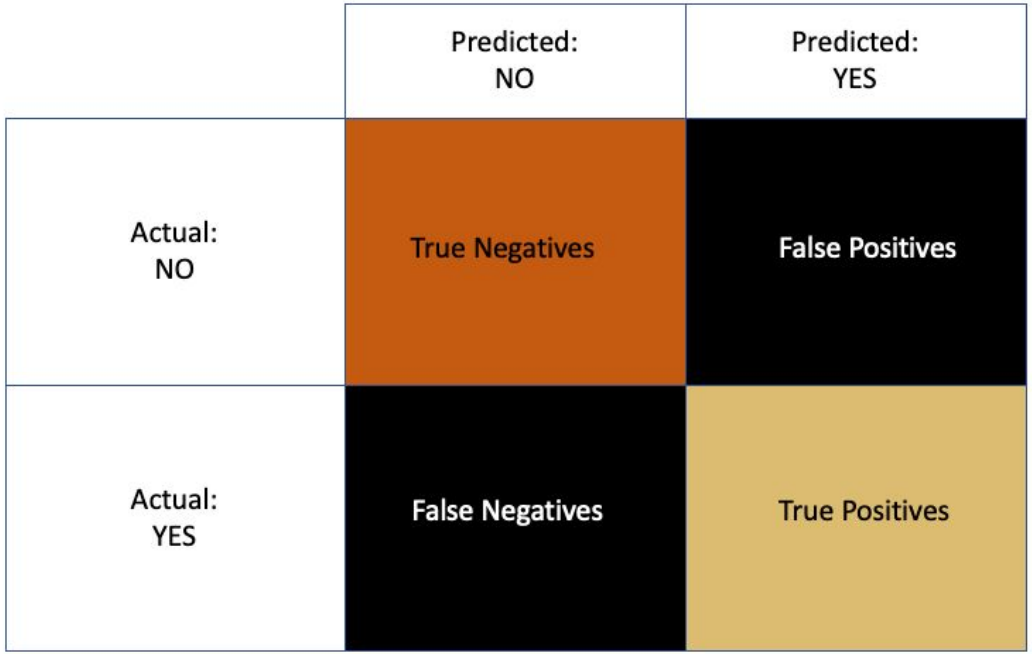
Pros:

- Easy to interpret and understand
- Easy data preparation
- Few assumptions
- Versatile

Cons:

- Overfitting
- Expensive training phase
- Optimization at current node rather than forward thinking
- Unstable (A change in data can have a drastic impact on the tree)

# Confusion Matrix



# Spotify Genre Classification

- “Spotify makes guesses with the use of decision trees and inputs of your playlist data”
- Suggest a song to a particular person.
  - Who would like song X?
- Assign a genre to a song.
  - Ask questions about the song to classify it
- “Using machines help us do this more objectively, and deal with way more parameters”
  - Spotipy library

Github - https://github.com/dhunstack/asmc-spotify-analysis

| acousticness | danceability | energy | instrumentalness | key | liveness | loudness | mode | speechiness | tempo | valence | song_title |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.399 | 0.748 | 0.465 | 0 | 5 | 0.294 | -8.163 | 1 | 0.0869 | 119.872 | 0.412 | like i need u |

# Sources


https://www.ibm.com/topics/decision-trees#:~:text=data%20mining%20solutions-,Decision%20Trees,internal%20nodes%20and%20leaf%20nodes.

https://jinkim0804.medium.com/spotify-decision-trees-with-music-taste-4c11a660ddc0

https://medium.com/@ashitaboyina/music-genre-classification-70ae70469403

https://en.wikipedia.org/wiki/Decision_tree

https://www.youtube.com/watch?v=_L39rN6gz7Y

https://insidelearningmachines.com/advantages_and_disadvantages_of_decision_trees/

https://www.analytixlabs.co.in/blog/decision-tree-algorithm/#Disadvantages_of_Decision_Tree