Classification and Regression Trees

Josiah Davis

Objectives

- Describe the output of a decision tree to someone without a data science background
- Describe how the algorithm creates the decision tree
- Predict the likelihood of a binary event using the decision tree algorithm in scikit-learn
- Create a decision tree visualization
- Determine the optimal tree size using a tune grid and the AUC metric in Python
- Describe the strengths and weaknesses of a decision tree

Objectives

- **Describe the output of a decision tree to someone without a data science background**
- Describe how the algorithm creates the decision tree
- Predict the likelihood of a binary event using the decision tree algorithm in scikit-learn
- Create a decision tree visualization
- Determine the optimal tree size using a tune grid and the AUC metric in Python
- Describe the strengths and weaknesses of a decision tree

What are decision trees?

What are decision trees?

- A supervised learning technique that can be used for discrete and continuous output.

What are decision trees?

- A supervised learning technique that can be used for discrete and continuous output.

- Visually engaging and easy to interpret.

What are decision trees?

- A supervised learning technique that can be used for discrete and continuous output.
- Visually engaging and easy to interpret.
- Excellent model for someone transitioning into the world of data science.

What are decision trees?

- A supervised learning technique that can be used for discrete and continuous output.
- Visually engaging and easy to interpret.
- Excellent model for someone transitioning into the world of data science.
- Foundational to learning some very powerful techniques.

What are decision trees?

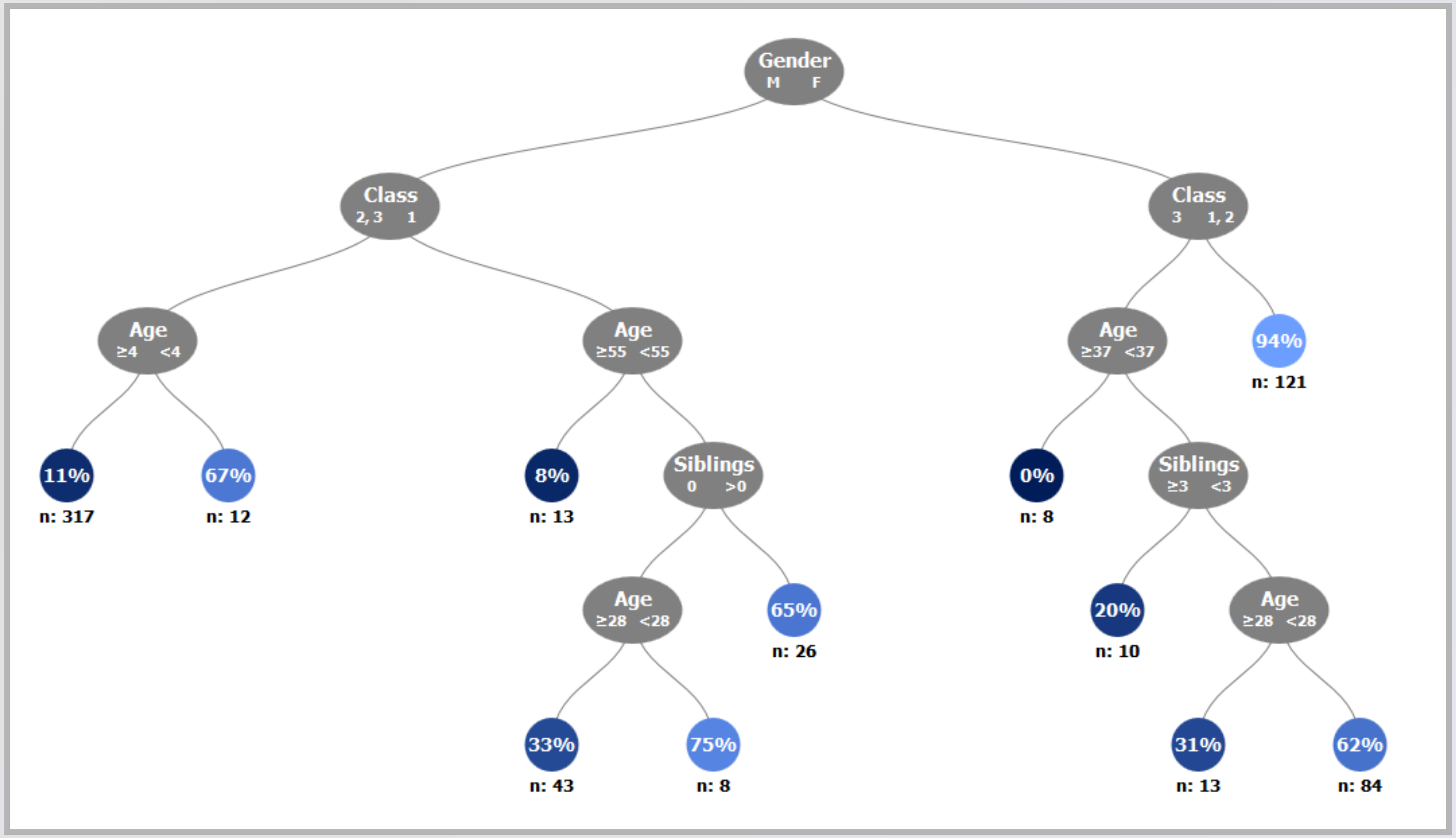
- A supervised learning technique that can be used for discrete and continuous output.
- Visually engaging and easy to interpret.
- Excellent model for someone transitioning into the world of data science.
- Foundational to learning some very powerful techniques.
- Are prone towards high-variance.

What are decision trees?

- A supervised learning technique that can be used for discrete and continuous output.
- Visually engaging and easy to interpret.
- Excellent model for someone transitioning into the world of data science.
- Foundational to learning some very powerful techniques.
- Are prone towards high-variance.
- We will focus on the CART algorithm.

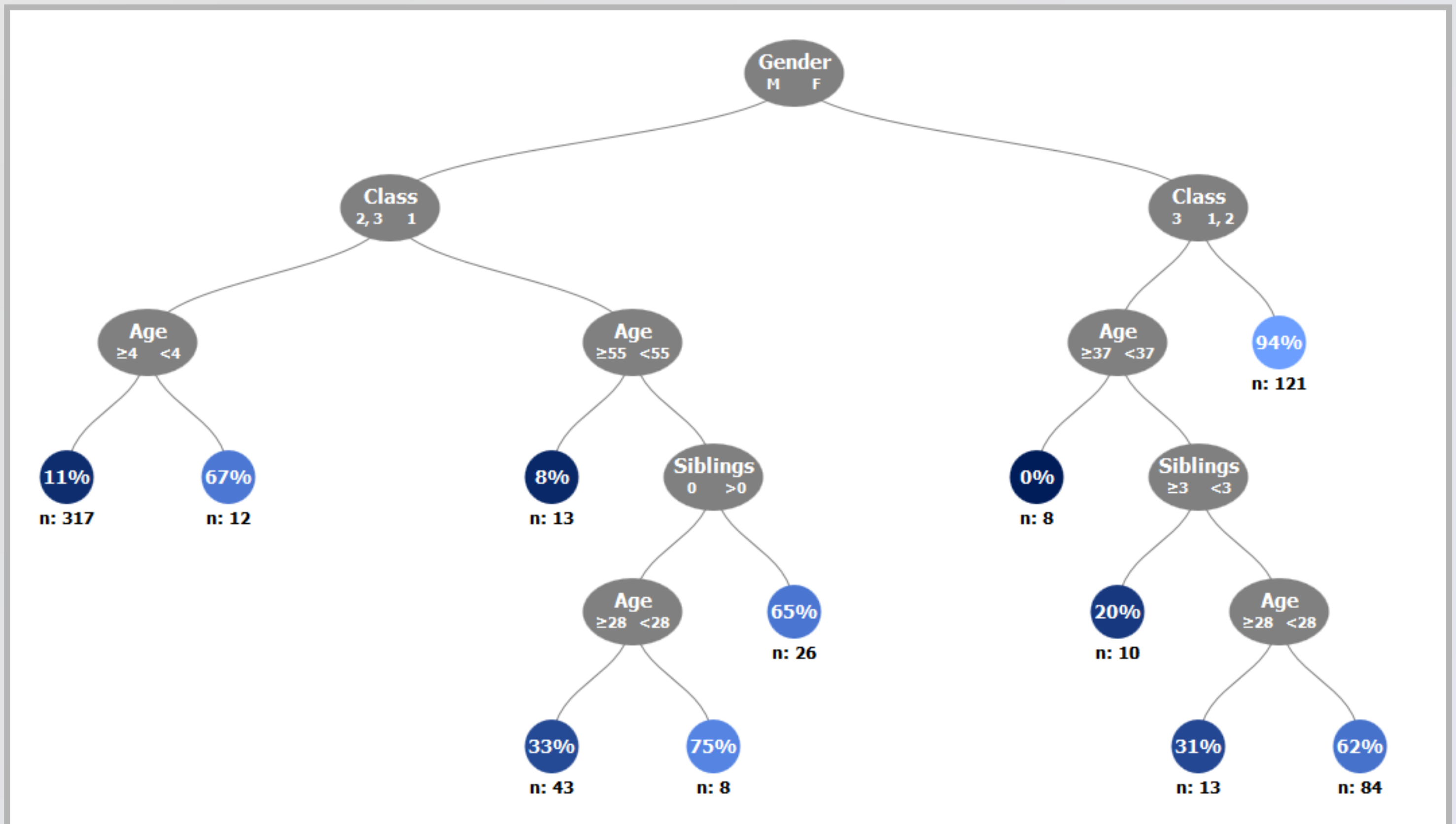
Description

Best place to start understanding decision trees is to look at one of them. The diagram below shows a decision tree trained on the titanic data set.



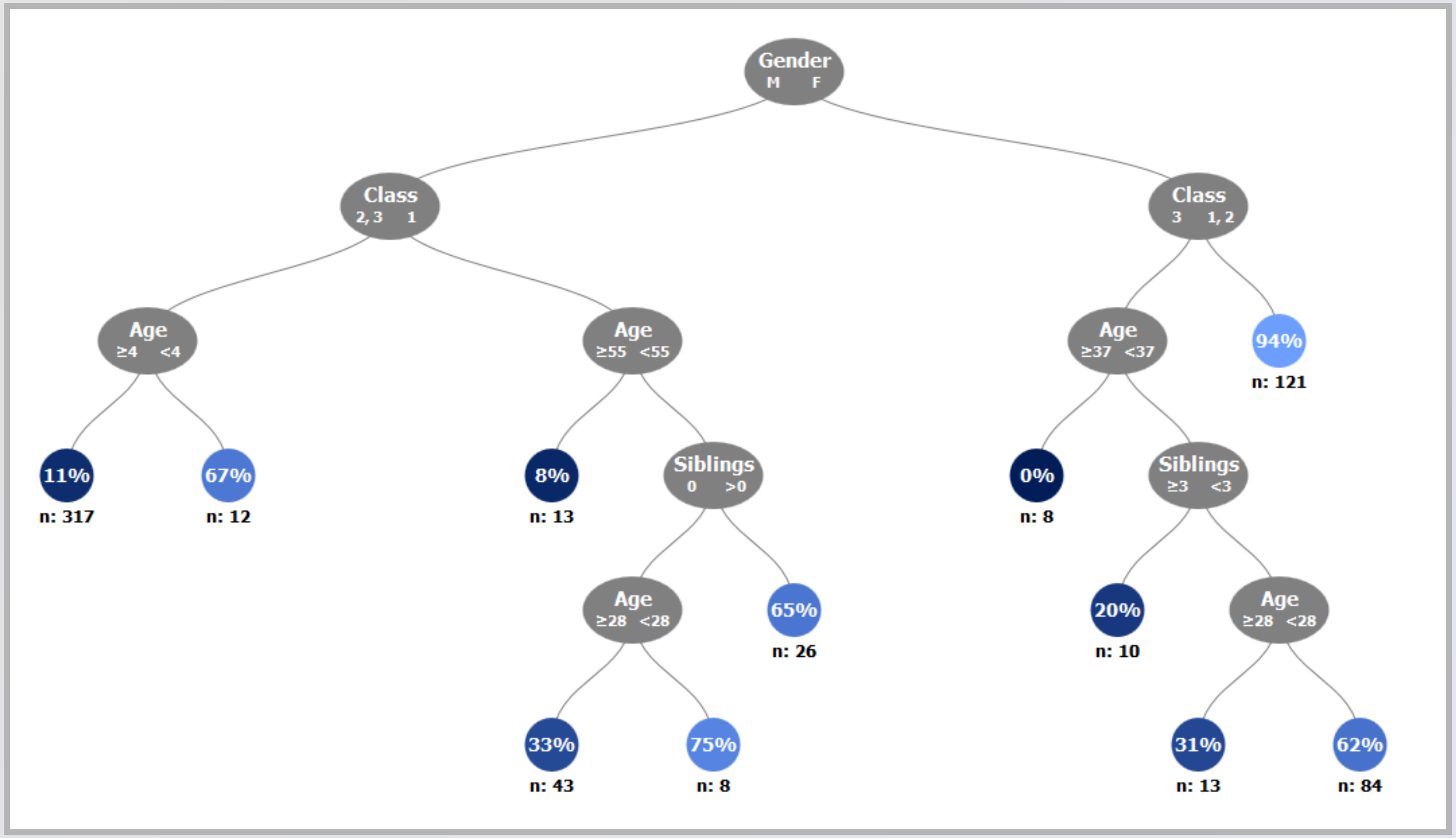
Description

Decision Trees are made up of interconnected nodes, which act as a series of questions / test conditions (e.g., is the passenger male or female?)



Description

Terminal nodes show the output metric, in this case the percentage of titanic survivors for a given combination of variables.



This raises questions

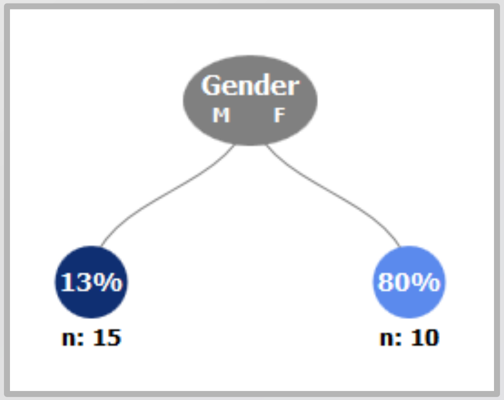
- How does the algorithm choose which variables to include on the tree?
- How does the algorithm choose where variables should be located on the tree?
- How does the algorithm choose when to stop the tree?

Objectives

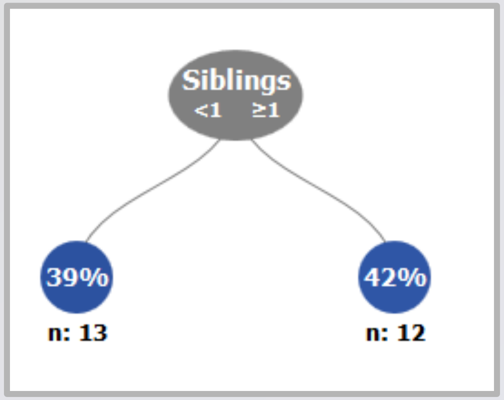
- Describe the output of a decision tree to someone without a data science background
- **Describe how the algorithm creates the decision tree**
- Predict the likelihood of a binary event using the decision tree algorithm in Python
- Create a decision tree visualization
- Determine the optimal tree size using a tune grid and the AUC metric in Python
- Describe the strengths and weaknesses of a decision tree

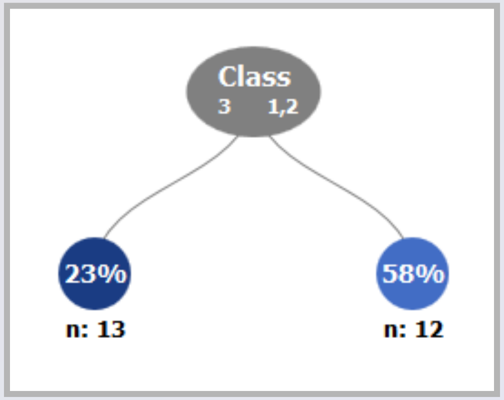
The Algorithm, Introduced

Different variables and split options are evaluated to determine which split will provide the greatest separation between classes.



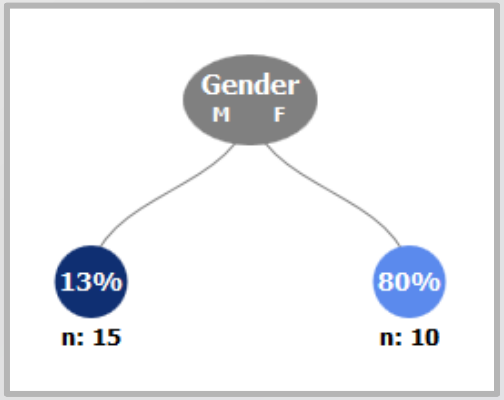
Which split option would you select?



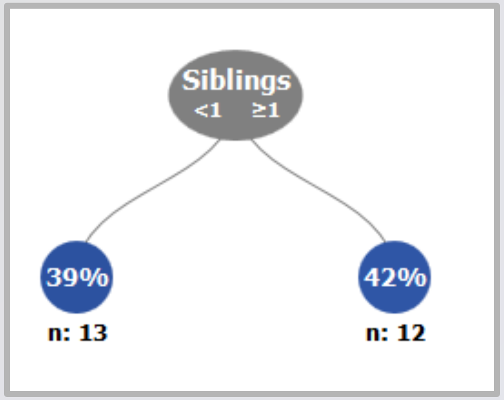


The Algorithm, Introduced

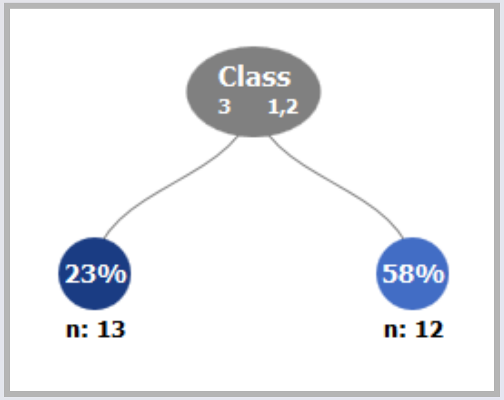
Different variables and split options are evaluated to determine which split will provide the greatest separation between classes.



Which split option would you select?



How can we determine the best split analytically?



Measuring Purity

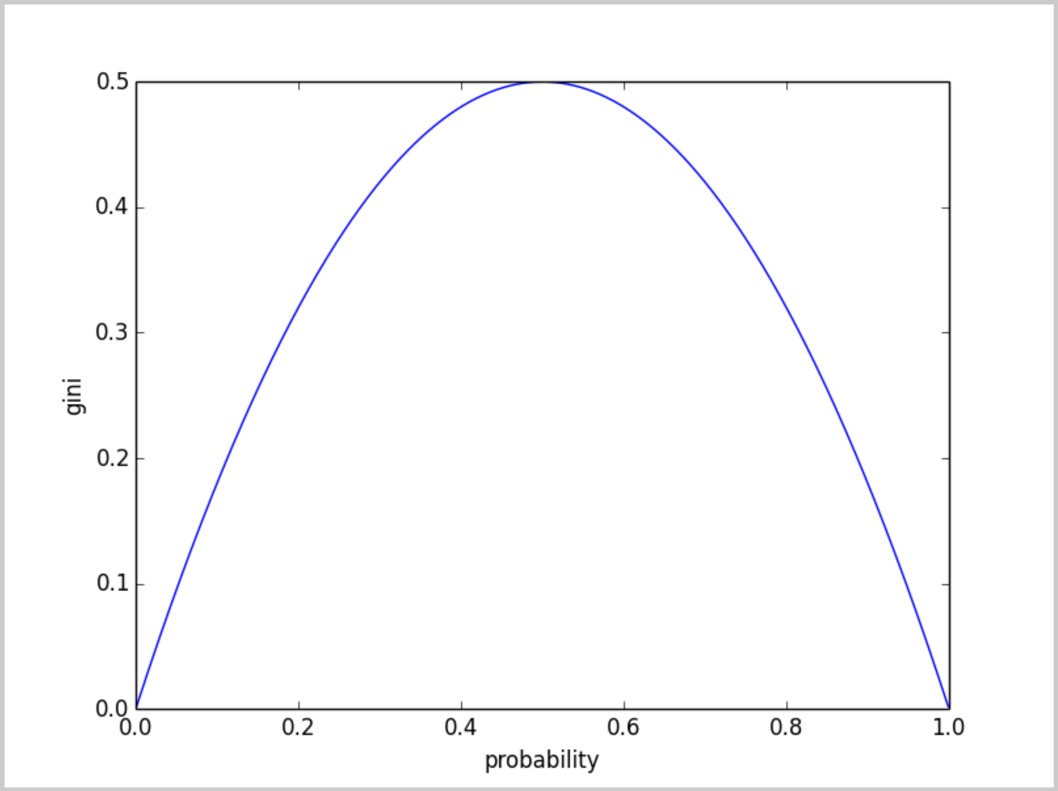
For decision trees, splits are chosen by measuring the class purity before and after the split. Purity can be calculated by the Gini index.

$$Gini = 1 - \sum_{i=1}^{class} proportion_i^2$$

Josiah Davis

Measuring Purity

A Gini index of 0.5 indicates equal representation between both classes, and a Gini index of 0 indicates perfect separation between classes (i.e., perfect purity)



The Algorithm – Overview

- Calculate the purity of the data

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split
- Calculate the purity of the data after the split

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split
- Calculate the purity of the data after the split
- Repeat for all variables

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split
- Calculate the purity of the data after the split
- Repeat for all variables
- Choose the variable with the greatest increase in purity

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split
- Calculate the purity of the data after the split
- Repeat for all variables
- Choose the variable with the greatest increase in purity
- Repeat for each split until some stop criteria is met

The Algorithm – Overview

- Calculate the purity of the data
- Select a candidate split
- Calculate the purity of the data after the split
- Repeat for all variables
- Choose the variable with the greatest increase in purity
- Repeat for each split until some stop criteria is met

Let’s go through an example

Before the Split

Before Split	All
Survived	10
Died	15

Before the Split

Before Split	All
Survived	10
Died	15

What is the gini coefficient?

Before the Split

Before Split	All
Survived	10
Died	15

$$1 - \sum \left(\frac{class_i}{total} \right)^2$$

Before the Split

Before Split	All
Survived	10
Died	15

$$1 - \left(\frac{survived}{total}\right)^2 - \left(\frac{died}{total}\right)^2$$

Before the Split

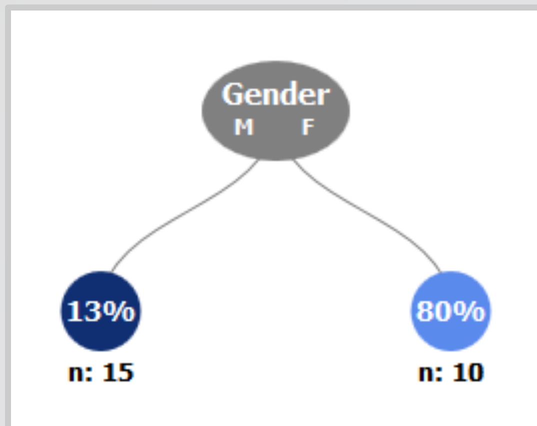
Before Split	All
Survived	10
Died	15

$$1 - \left(\frac{survived}{total}\right)^2 - \left(\frac{died}{total}\right)^2$$

$$1 - \left(\frac{10}{25}\right)^2 - \left(\frac{15}{25}\right)^2 = 0.48$$

After the Split

$Gini_O$
$=0.48$

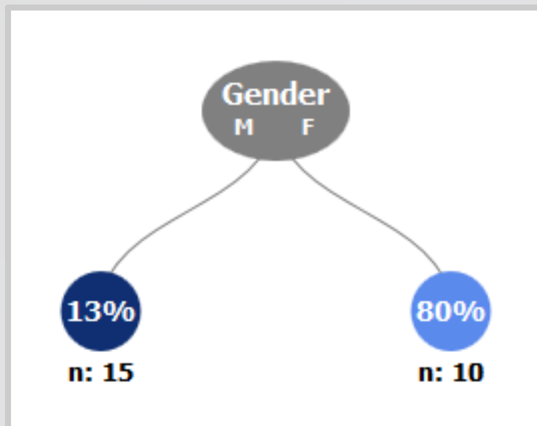


Gender	M
Survived	2
Died	13

Gender	F
Survived	8
Died	2

After the Split

$Gini_O$
$=0.48$



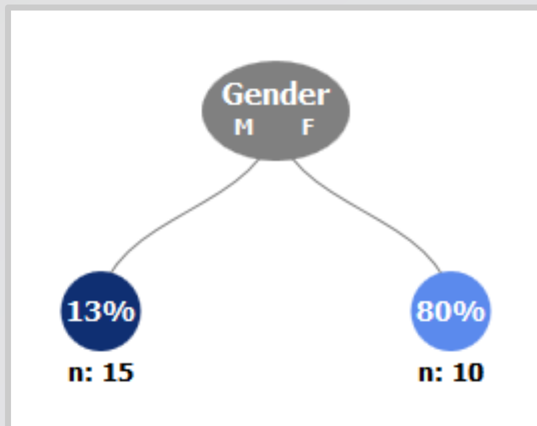
Gender	M
Survived	2
Died	13

Gender	F
Survived	8
Died	2

What is the gini coefficient?

After the Split

$Gini_O$
$=0.48$



Gender	M
Survived	2
Died	13

Gender	F
Survived	8
Died	2

$Gini_M$
$1-(2/15)^2-(13/15)^2=0.23$

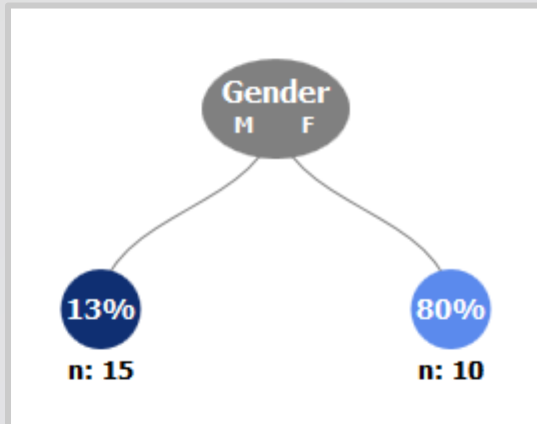
$Gini_F$
$1-(8/10)^2-(2/10)^2=0.32$

Josiah Davis

After the Split (continued)

$Gini_O$

$=0.48$



$Gini_M$

$=0.23$

$Gini_F$

$=0.32$

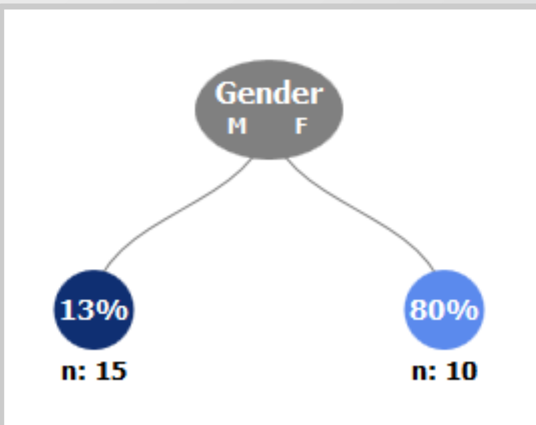
$Gini_C$

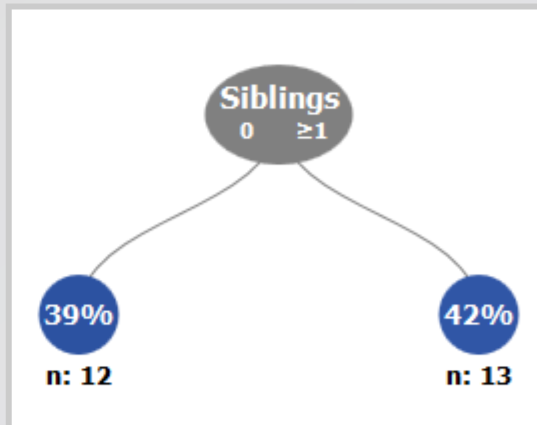
$$Gini_M \left(\frac{M}{M+F}\right)+Gini_F \left(\frac{F}{M+F}\right)=$$

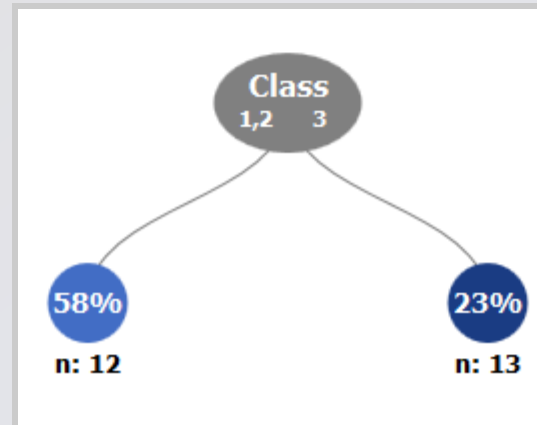
$$0.23\left(\frac{15}{10+15}\right)+0.32\left(\frac{10}{10+15}\right)=0.27$$

Choosing the Split

How does the gini coefficient compare for the Siblings and class variables?







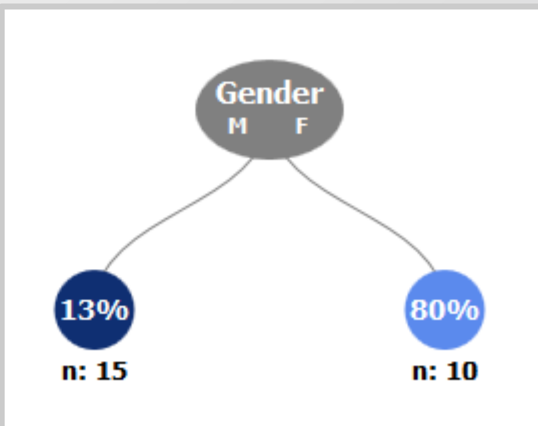
Gender	M	F
Survived	2	8
Died	13	2
$Gini_C$	0.27	

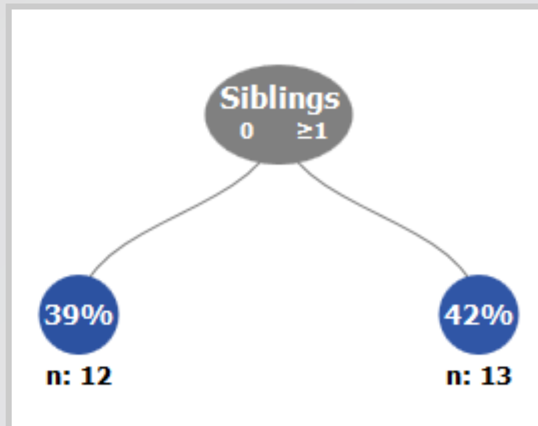
Siblings	0	≥1
Survived	5	5
Died	7	8
$Gini_C$	??	

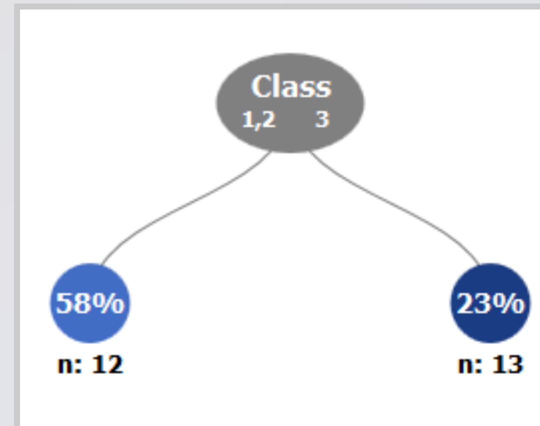
Class	1,2	3
Survived	7	3
Died	5	10
$Gini_C$	??	

Choosing the Split

How does the gini coefficient compare for the Siblings and class variables?







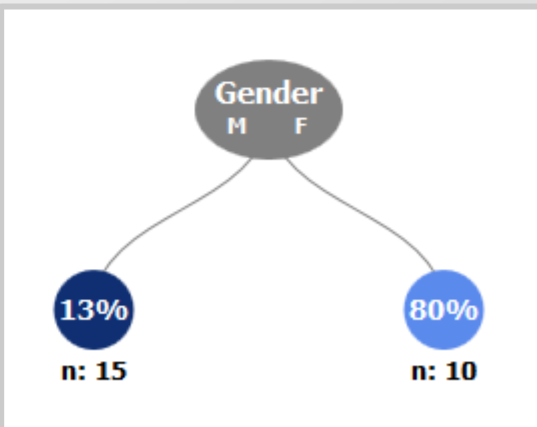
Gender	M	F
Survived	2	8
Died	13	2
Gini$_C$	0.27	

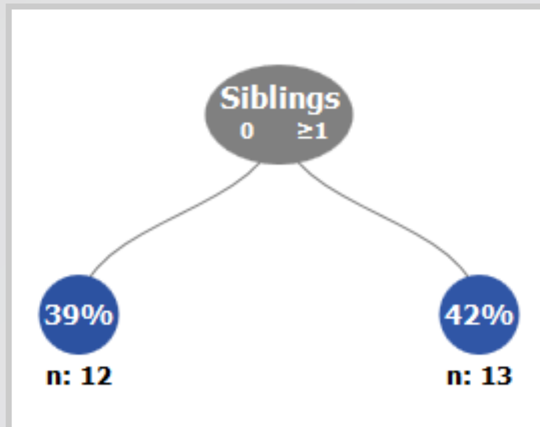
Siblings	0	≥1
Survived	5	5
Died	7	8
Gini$_C$	0.48	

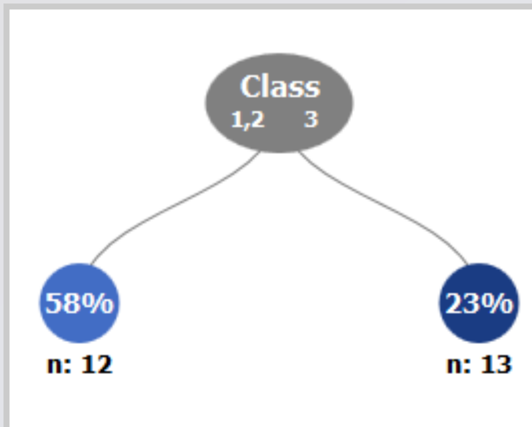
Class	1,2	3
Survived	7	3
Died	5	10
Gini$_C$	0.42	

Choosing the Split

In this example, the algorithm will select the Gender variable since it provides the greatest increase in purity.







Gender	M	F
Survived	2	8
Died	13	2
$Gini_C$	0.27	

Siblings	0	≥1
Survived	5	5
Died	7	8
$Gini_C$	0.48	

Class	1,2	3
Survived	7	3
Died	5	10
$Gini_C$	0.42	

Objectives

- Describe the output of a decision tree to someone without a data science background
- Describe how the algorithm creates the decision tree
- **Predict the likelihood of a binary event using the decision tree algorithm in scikit-learn**
- **Create a decision tree visualization**
- **Determine the optimal tree size using a tune grid and the AUC metric in Python**
- Describe the strengths and weaknesses of a decision tree

Objectives

- Describe the output of a decision tree to someone without a data science background
- Describe how the algorithm creates the decision tree
- Predict the likelihood of a binary event using the decision tree algorithm in scikit-learn
- Create a decision tree visualization
- Determine the optimal tree size using a tune grid and the AUC metric in Python
- **Describe the strengths and weaknesses of a decision tree**

Meta-Evaluation

What do you think?

Meta-Evaluation

Advantages:

- The nature of its output provides the decision tree algorithm with a degree of interpretability that other more complex algorithms don’t provide

Meta-Evaluation

Advantages:

- The nature of its output provides the decision tree algorithm with a degree of interpretability that other more complex algorithms don't provide
- Understanding how the decision Tree works is foundational to understanding how more complex, and widely-used models work, such as random forests and boosted trees.

Meta-Evaluation

Advantages:

- The nature of its output provides the decision tree algorithm with a degree of interpretability that other more complex algorithms don’t provide
- Understanding how the decision Tree works is foundational to understanding how more complex, and widely-used models work, such as random forests and boosted trees.

Disadvantages:

- The decision Tree tends to perform worse than more sophisticated modeling techniques due to their **high** variance