Decision Trees

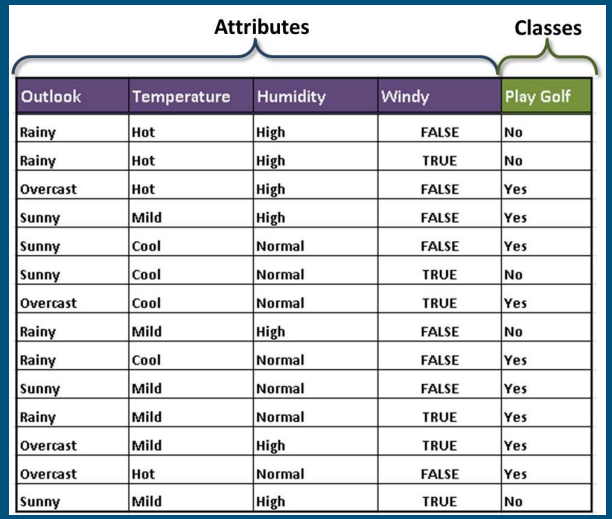
Carol Lopez, Finn McSweeney, Zeeshan Pervaiz, & Rachel Walter

Background on Decision Trees

- Also known as Classification and Regression Trees (CART)
- The first regression tree algorithm was seen in 1963 when Morgan and Sonquist published their work on Automatic Interaction Detection (AID)
- The first classification tree algorithm was seen in 1972 when Messenger and Mandel published their work on Theta Automatic Interaction Detection (THAID)
- CART was developed in 1983 by Breiman et al, they used ideas from AID and THAID to create CART
- CART refers to all types of decision trees, but it also refers to the specific algorithm developed by Breiman et al
- Other popular CART algorithms are ID3, Chi-Square, Reduction in Variance, and C4.5 (which is an extension of ID3)

When is CART typically used?

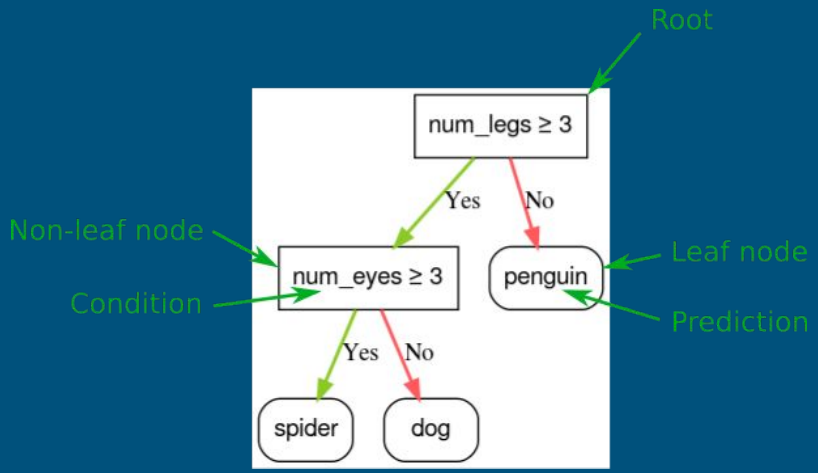
- CART is used when you are trying to predict an outcome
- Regression is used when the outcome variable (what you are trying to predict) is continuous
- Classification is used when the outcome variable is categorical
- To use CART, the dataset should have columns that contain features, which are essentially attributes or predictors, and one column that contains what you are trying to predict (the response or the outcome)



Attributes				Classes
Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

How it Works

- A decision tree is a machine learning model used for classification and regression.
- It splits the training data into subsets: each getting smaller all based on certain features.
- The algorithm creates a decision node which splits the data into two groups and follows the pattern as it moves down the tree.
- When new data is added, the data follows down the decision tree based off of certain characteristics.



Advantages and Disadvantages

Pros:

- Less data preparation compared to other ML algorithms.
- Normalization of data is not required.
- Missing data does not affect the process of building a decision tree
- Intuitive
- Decision Trees can capture and classify non-linear relationships

Cons:

- High variance
- Small changes in the data can alter the structure of the tree significantly.
- Not well suited for continuous variables, they work best with categorical or binary data

Data Processing Steps

- Little to no need for data preprocessing: Decision Trees looks at and evaluates each feature individually instead of the dataset as a whole.
  - Gini Index vs Entropy
- We can however make the data easier to work with by: handling missing values and null values and converting categorical values into numerical form.
  - LabelEncoder (sklearn.preprocessing)
- Selecting test size for the model and adjusting the parameters could improve the accuracy.

Hyperparameters

Parameter	Description	Effect
Max_depth	Parameter controlling the number of layers where we split the nodes.	Lower: Model is faster but not as accurate. Higher: higher accuracy but prone to overfitting and slow.
Min_samples_split	The minimum number of samples required to split a node.	Setting to higher values can help mitigate overfitting.
min_samples_leaf	The minimum number of samples required to be at a leaf node.	It reduces overfitting
Criterion	Determines how the impurity of a split will be measured	Gini: probability of impurity at each node. Entropy: calculated measure of impurity at each node.

GINI INDEX: What is the best split?

- Gini index is a metric for classification tasks in CART. It stores sum of squared probabilities of each class.
- Compute the Gini Index for each attribute/feature values: Gini (Attribute='value')

$$Gini(Attribute = value) = Gini(Av) = 1 - \sum_j p_j^2$$

- Compute the weighted sum of Gini Indexes for attribute/feature Gini (Attribute)

$$Gini(Attribute) = \sum_v p_v * Gini(Av)$$

- Pick the Lowest Gini Index Attribute

Outlook	Temperature	Humidity	Wind	PlayGolf
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Yes = 9

No = 5

Total = 14

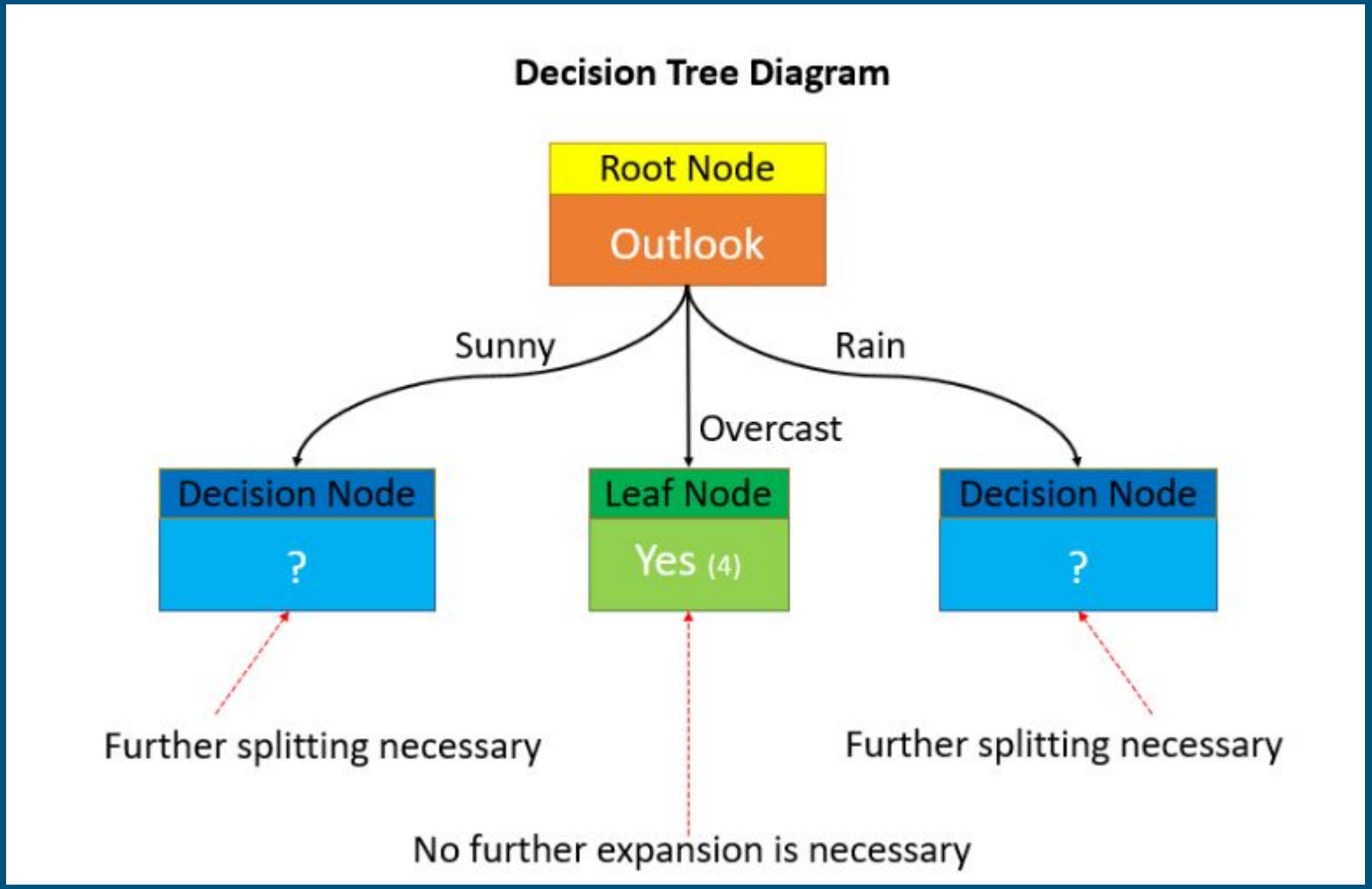
Outlook	Yes	No	Total
Sunny	3	2	5
Overcast	4	0	4
Rainy	3	2	5
Total	10	4	14

$$Gini(Outlook = Sunny) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$Gini(Outlook = Overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0$$

$$Gini(Outlook = Rainy) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$Gini(Outlook) = \frac{5}{14} * (0.48) + \frac{4}{14} * (0) + \frac{5}{14} * (0.48) = 0.3429$$


Decision Tree Diagram
Root Node
Outlook
Sunny
Overcast
Rain
Decision Node
?
Leaf Node
Yes (4)
Decision Node
?
Further splitting necessary
No further expansion is necessary
Further splitting necessary

Works Cited

Google. (2022, July 18). *Decision Trees*. Google. Retrieved January 4, 2023, from https://developers.google.com/machine-learning/decision-forests/decision-trees

Loh, Wei-Lin. (n.d.). *A Brief History of Classification and Regression Trees* [PowerPoint Slides]. Department of Statistics, University of Wisconsin-Madison. https://washstat.org/presentations/20150604/loh_slides.pdf

milaan9. (n.d.). PYTHON_DECISION_TREE_AND_RANDOM_FOREST/002_Decision_Tree_PlayGolf_CART.ipynb at main · milaan9/python_decision_tree_and_random_forest. GitHub. Retrieved January 5, 2023, from https://github.com/milaan9/Python_Decision_Tree_and_Random_Forest/blob/main/002_Decision_Tree_PlayGolf_CART.ipynb

Seif, George. "A Guide to Decision Trees for Machine Learning and Data Science." *Medium*, Towards Data Science, 11 Feb. 2022, https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956.

YouTube. (2017). *Let's Write a Decision Tree Classifier from Scratch*. *YouTube*. Retrieved January 4, 2023, from https://www.youtube.com/watch?v=LDRbO9a6XPU.