**Decision Trees – Information Sheet – CS 5100**

**Definition:** A decision tree is a tree-like structure used for decision-making, where each node represents a decision point, and leaf nodes provide the outcomes.

**Key Components:**

* Root Node: The starting point of the tree.
* Internal Nodes: Represent decision points based on features.
* Leaf Nodes: Indicate final predictions or outcomes.

A black background with white text

Description automatically generated**Key formulas:**

* Gini Impurity – measures how good a split is

**Process of Building a Decision Tree From Data**

* Feature Selection – Identify the best feature to split the data using Gini Impurity .
* Splitting the Data – Divide the dataset into subsets based on the selected feature.
* Recursive Partitioning – Repeatedly apply feature selection and splitting for each subset
* Stopping Criteria – condition or level based

Algorithm Pseudocode

Function BuildTree(data, depth):

If stopping criteria met (depth >= max\_depth OR |data| < min\_samples\_split OR all targets are the same):

Create a leaf node with the predicted value.

Return leaf node.

BestSplit = Find the best split with lowest Gini Impurity

If BestSplit is None:

Create a leaf node with the predicted value.

Return leaf node.

Split data into left and right subsets using BestSplit.

LeftSubtree = BuildTree(left subset, depth + 1)

RightSubtree = BuildTree(right subset, depth + 1)

Create a decision node with BestSplit, LeftSubtree, and RightSubtree.

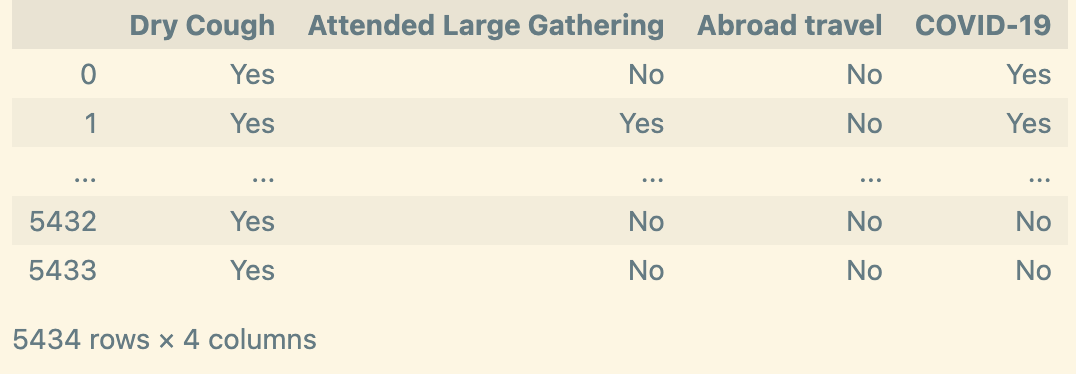
Return decision node.

**Why are decision trees important?**

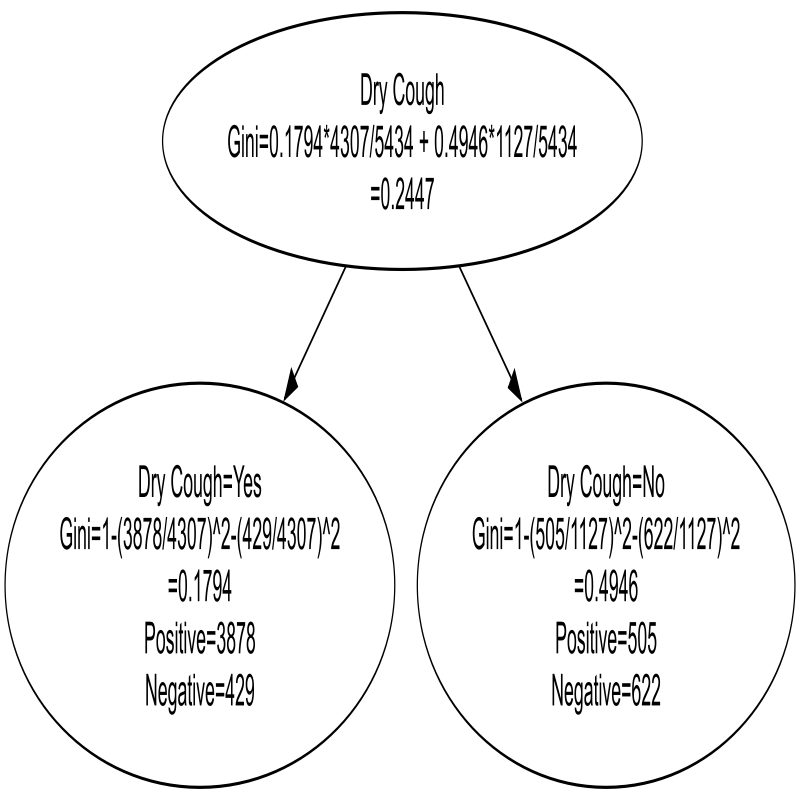
Decision trees are invaluable due to their balance between simplicity, interpretability, and functionality across a wide range of applications. They can be used in a classification or regressive approach and provide the foundation for many more advanced machine learning techniques, algorithms, and capabilities.

**Example problem:**

Say that we have a dataset like below, we can apply the algorithm above to build a decision tree to predict if someone is COVID positive.



An example of calculating the Gini Impurity



The final tree

