DECISION TREE CLASSIFICATION AND REGRESSION

Sourav Karmakar

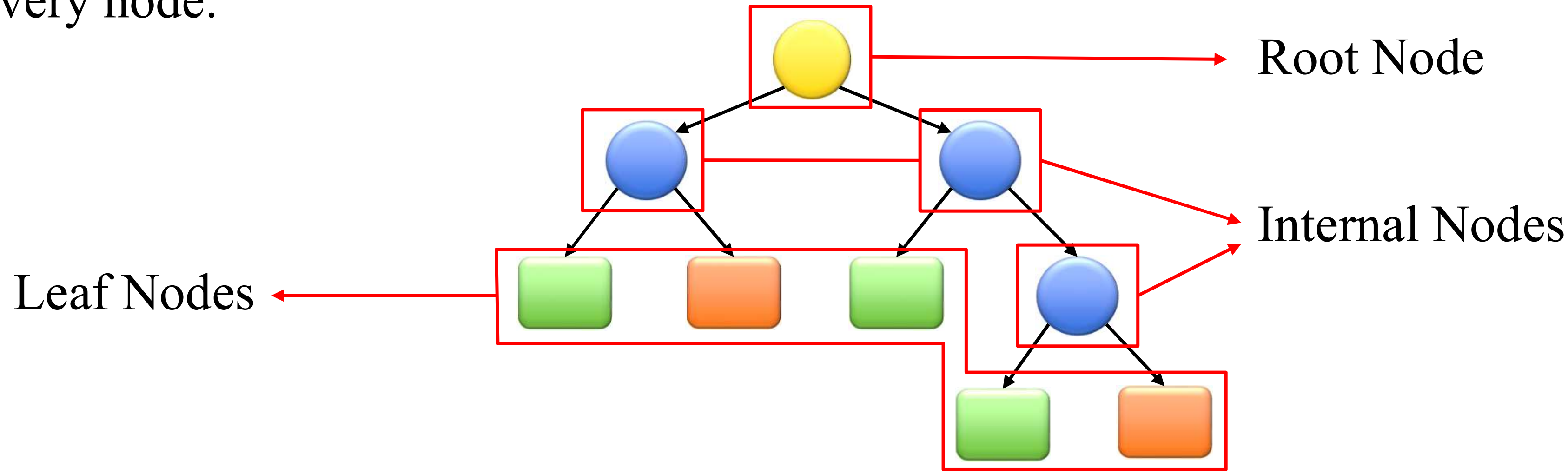
souravkarmakar29@gmail.com

Outline

- What is a Decision Tree Classifier
- Example of a Decision Tree Classifier
- Building a Decision Tree from a dataset
- Specifying test conditions
- Determining best split
- Calculating node impurity
- Determining node impurity and best split
- Decision Tree Regression
- When to stop splitting
- Merits and Demerits of Decision Tree

Decision Tree Classifier

- A **Decision Tree** splits the dataset using the structure of a tree and it makes a decision at every node.



What does each component describe?

Root Node and Internal Nodes → Test on a attribute / feature

Branches → Outcomes of the test on attributes

Leaf Nodes → Target Class Label

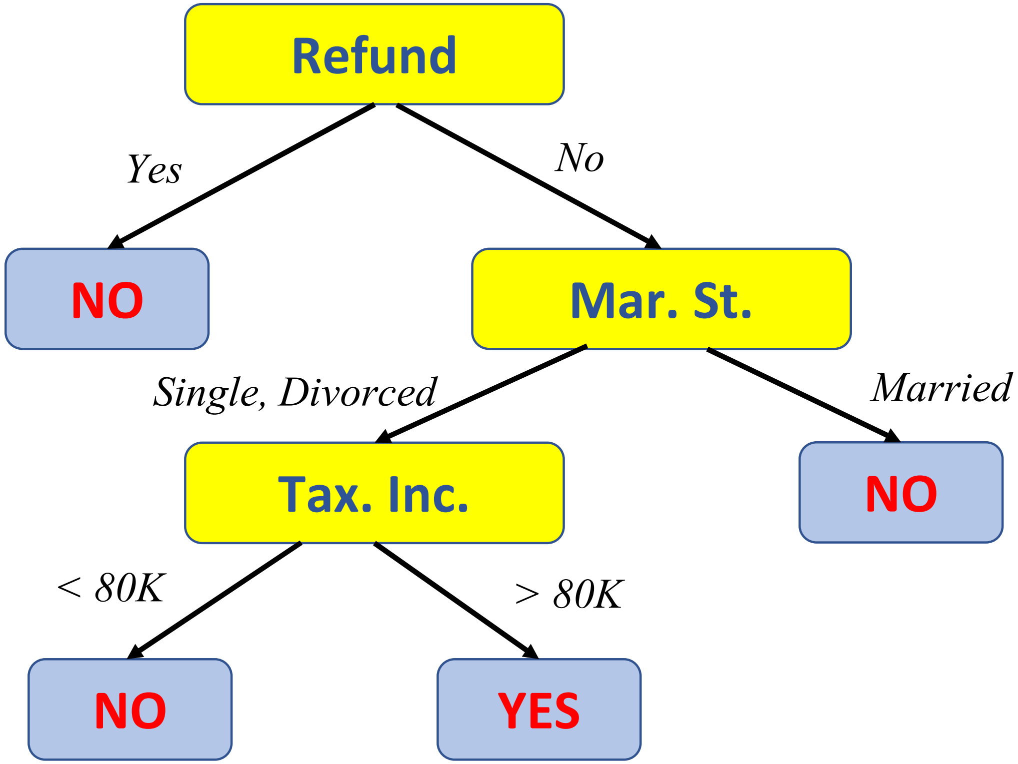
Example Of A DT Classifier

- **Creating a Model:**

ID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

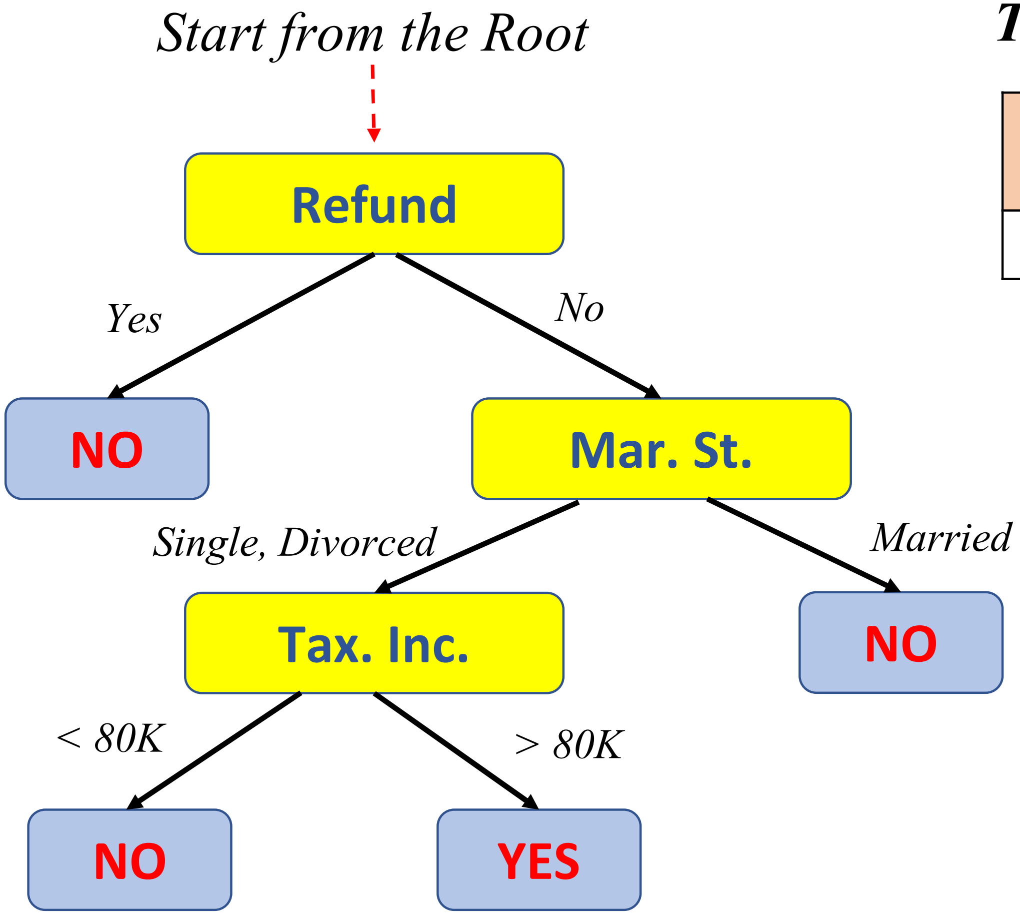
Categorical Categorical Continuous Class

We can fit a Decision Tree like following for the given training data.



Example Of A DT Classifier

- **Applying the model on a test data:**

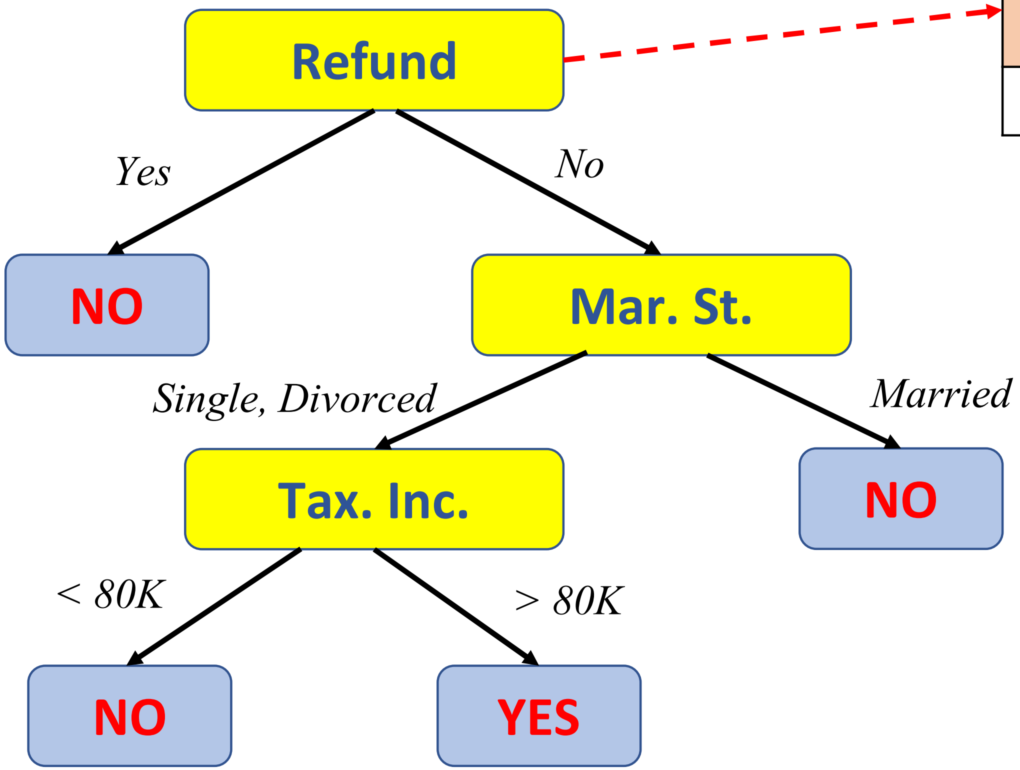


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Example Of A DT Classifier

- **Applying the model on a test data:**

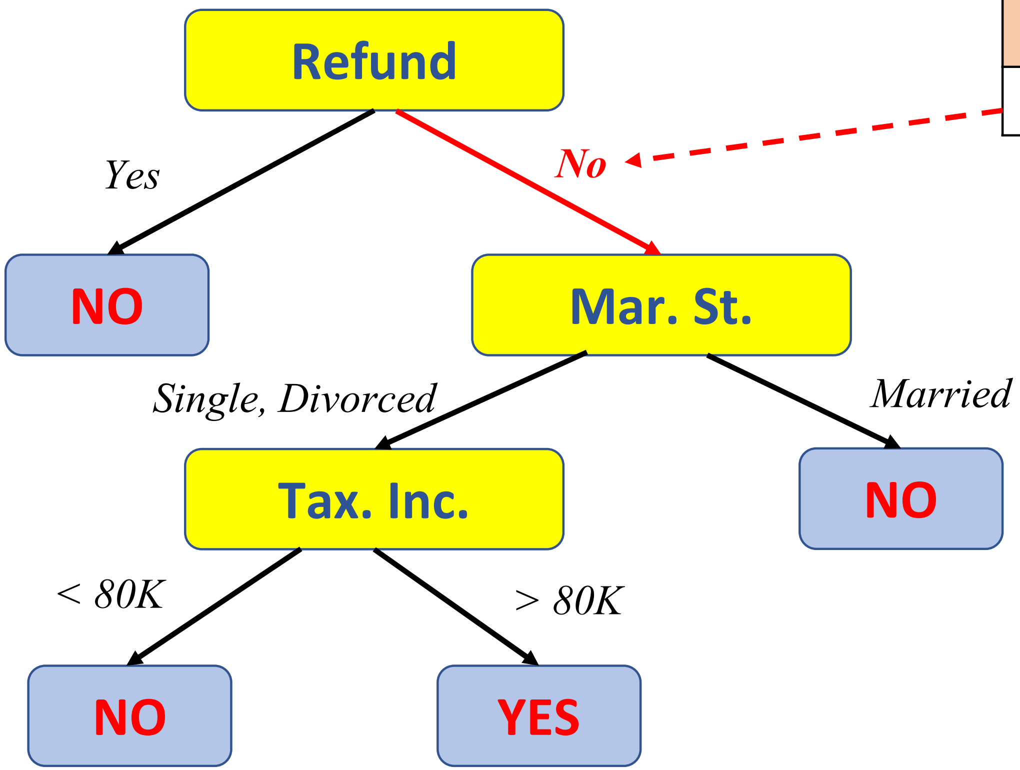


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Example Of A DT Classifier

- **Applying the model on a test data:**

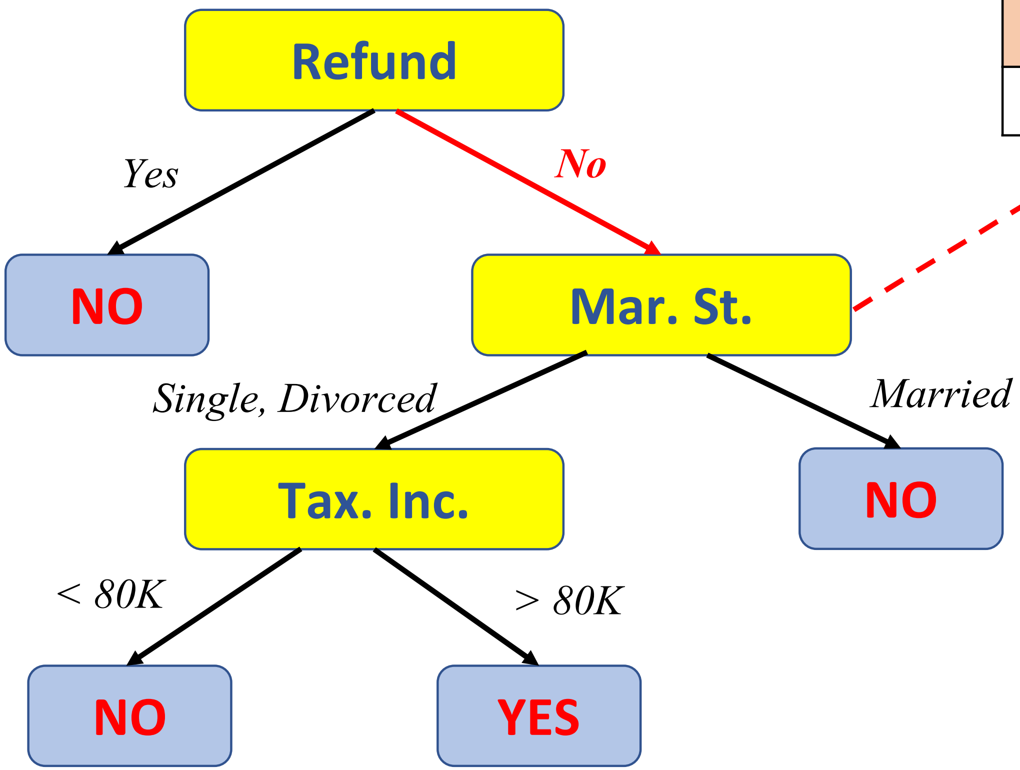


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Example Of A DT Classifier

- **Applying the model on a test data:**

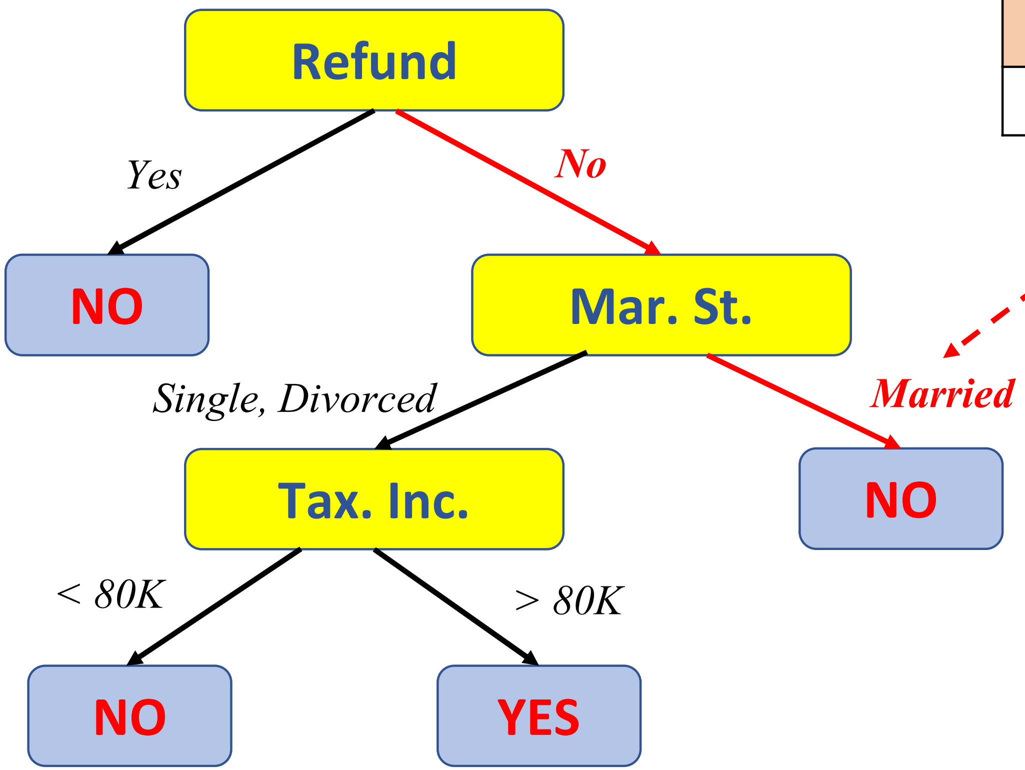


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Example Of A DT Classifier

- **Applying the model on a test data:**



Test Data

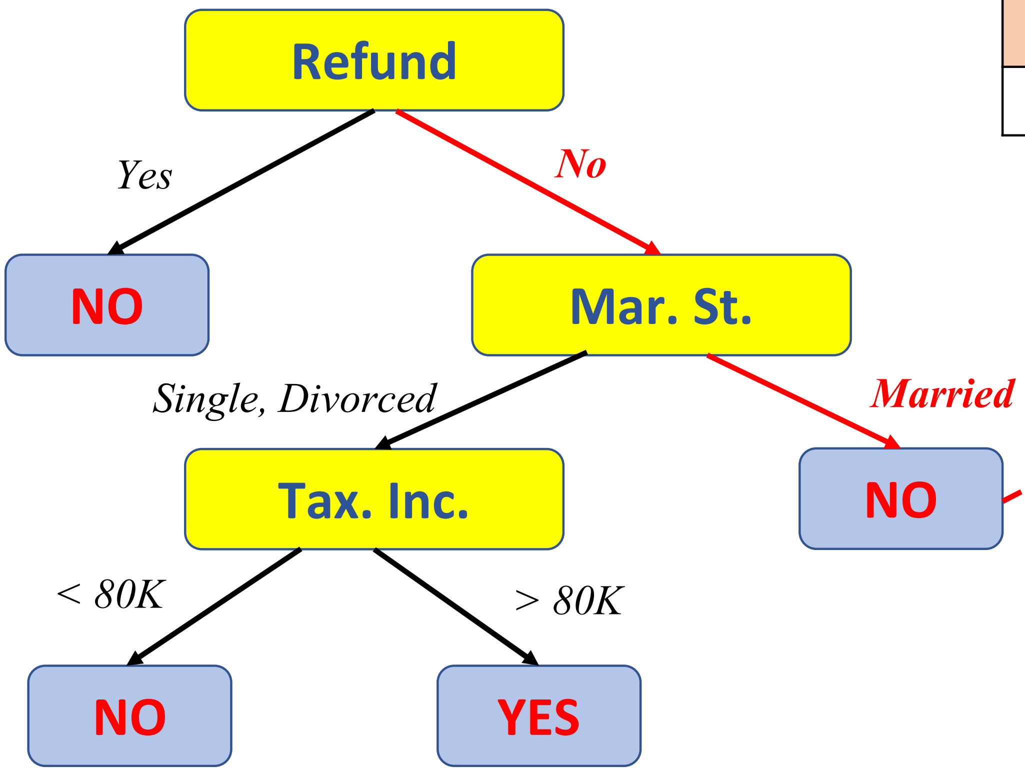
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Example Of A DT Classifier

- **Applying the model on a test data:**

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to “No”

Decision Tree Building (Induction)

- There are many algorithms to construct a Decision Tree from labelled training dataset.
  - Hunt's Algorithm
  - CART (Short form of **C**lassification **A**nd **R**egression **T**ree)
  - ID3 (ID stands for **I**terative **D**ichotomiser)
  - C4.5
  - SLIQ (**S**upervised **L**earning **I**n **Q**uest)
  - SPRINT (**S**calable **P**a**R**allelizable **IN**duction of decision **T**rees)

Scikit Learn implements optimized version of CART algorithm.

sklearn decision tree documentation: https://scikit-learn.org/stable/modules/tree.html

Decision Tree Building (Induction)

- **Greedy Strategy:**
  - Split the records based on an attribute test that optimizes certain criteria.

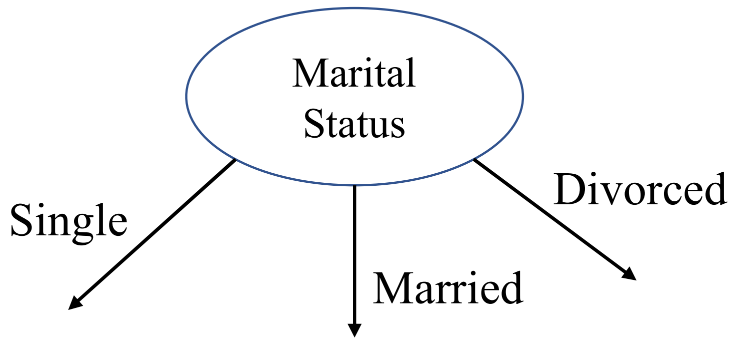
- **Issues:**
  - Determine how to split the records.
    - How to specify attribute test conditions?
    - How to determine best split?

  - Determine when to stop splitting.

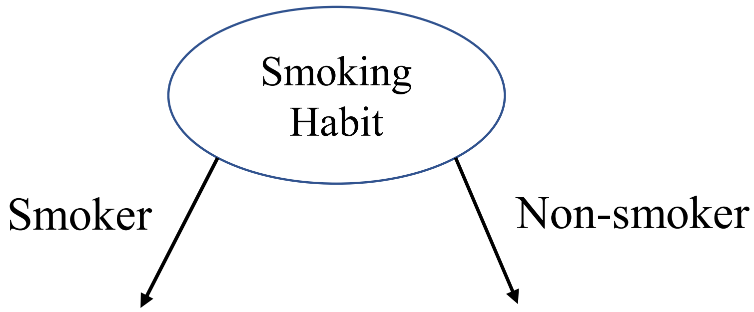
How To Specify Test Conditions

- Depends on attributes types
  - Categorical
  - Continuous
- Depends on number of ways to split
  - 2-way split
  - Multiway Split

- **Splitting based on Categorical Attribute:**



Multiway Split



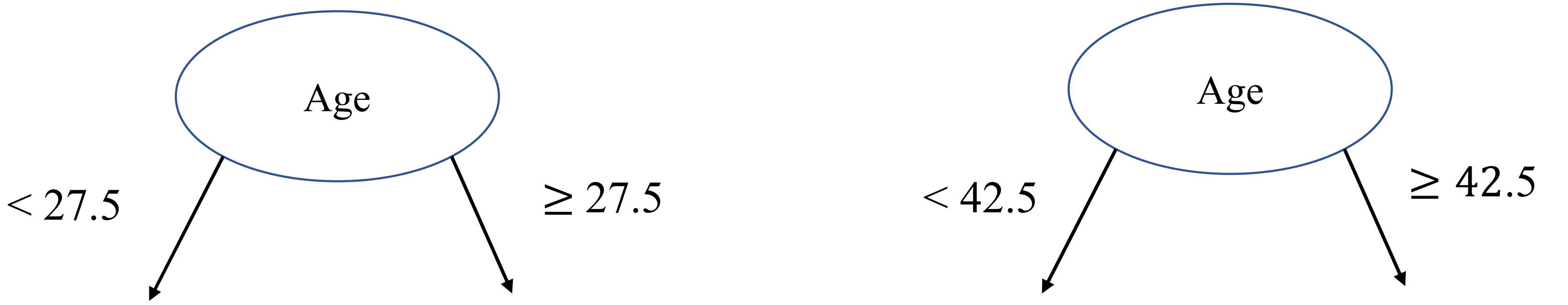
2-way Split

How To Specify Test Conditions

- **Splitting based on Continuous Attribute:**

  Decision trees handle continuous attributes by finding the **optimal split point** that best divides the data. They don't create a separate branch for every unique value. Instead, they treat the continuous attribute as a set of potential binary splits.

  For a continuous attribute, the algorithm first considers all unique values of the attribute in the dataset. It then sorts these values in ascending order. The potential split points are typically the **midpoints** between each pair of adjacent unique values. For example, if the attribute (age) values are 20, 35, and 50, the potential split points would be 27.5 and 42.5.



How To Determine Best Split

- Splitting a node creates child nodes. We need to find the best way to split a node.
- We need to find which attribute split the dataset best at a given node. If the best attribute type is continuous, we need to find the optimum value to split.
- Child nodes with *homogeneous* class distribution are preferred.
- Need a measure of Node impurity which will help to assess the splits numerically.
- **Intuition:**
  - Consider a binary classification problem.
  - Following scenarios are observed in two different nodes.

Class-1 : 10
Class-2 : 10

- Non-homogenous Node
- High degree of impurity
- Needs further split

Class-1 : 18
Class-2 : 2

- More Homogenous Node
- Low degree of impurity

How To Determine Best Split

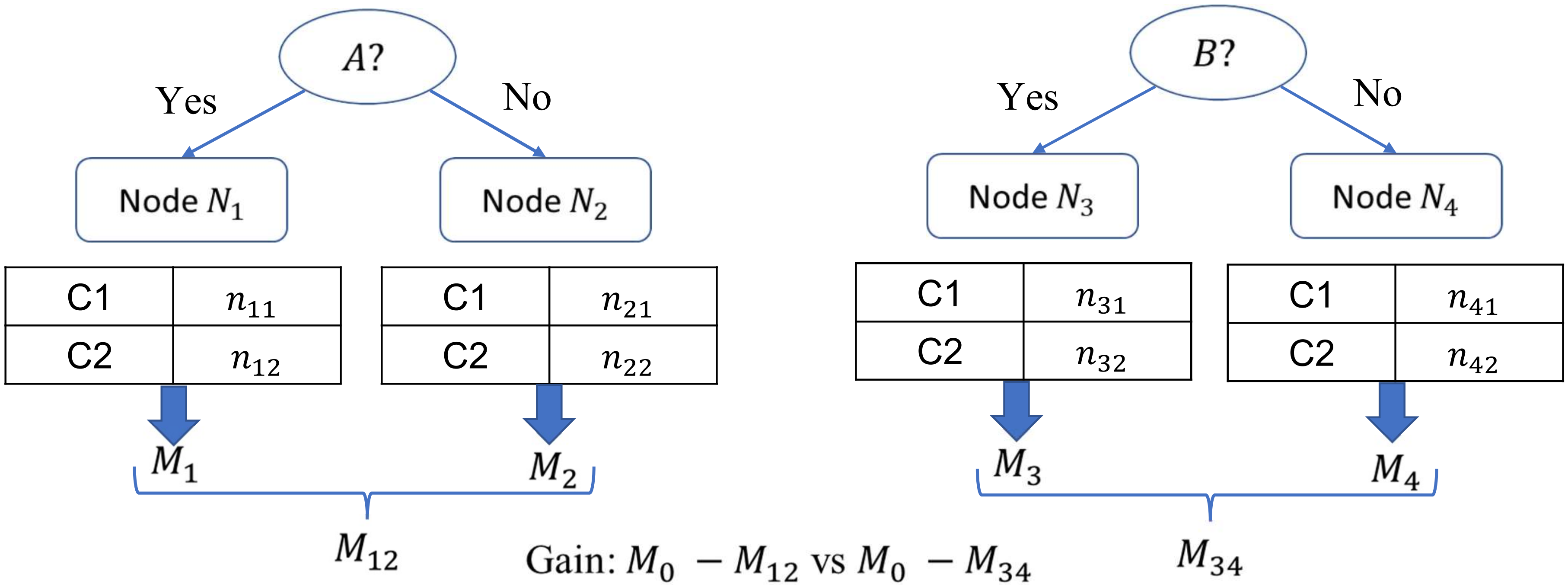
- Let's consider Binary Classification (Class- C_1 and C_2). Let M denote the measure of Node impurity. We shall discuss about measures of node impurity shortly.

Before splitting:

C1	n_{01}
C2	n_{02}

$\rightarrow M_0$

- Let there are two attributes A and B.



Gain: $M_0 - M_{12}$ vs $M_0 - M_{34}$

Measures Of Node Impurity In Classification

- There are various ways to measure Node impurity
  - **GINI Index:** Used in CART, SLIQ, SPRINT.
  - **Information Gain:** Used in ID3 and C4.5
  - **Misclassification Error**
  - **Log-loss**

And Many More…

- In this discussion we shall discuss GINI index for computation of impurities.

Measure Of Impurity : GINI

- GINI index for a given node t is calculated as:

$$GINI(t) = 1 - \sum_j \{p(j|t)\}^2$$

Where $p(j|t)$ is the relative frequency of class-j at node t.

- Maximum Value = $\left(1 - \frac{1}{n_c}\right)$, ($n_c$ = Number of Classes) occurs when all the samples at the node are equally distributed among classes. Implies high degree of impurity.

- Minimum Value = 0, occurs when all the samples at the node belong to one class. Implies homogeneity.

Measure Of Impurity : GINI

- **Examples of computing GINI index:**

$$GINI(t) = 1 - \sum_j \{p(j|t)\}^2$$

C1	0
C2	6

$$P(C_1) = \frac{0}{6} = 0, \qquad P(C_2) = \frac{6}{6} = 1,$$

$$\therefore\ GINI = 1 - [P(C_1)^2 + P(C_2)^2] = 1 - [0 + 1] = 0$$

C1	1
C2	5

$$P(C_1) = \frac{1}{6}, \qquad P(C_2) = \frac{5}{6},$$

$$\therefore\ GINI = 1 - [P(C_1)^2 + P(C_2)^2] = 0.278$$

C1	3
C2	3

$$P(C_1) = \frac{3}{6} = 0.5, \qquad P(C_2) = \frac{3}{6} = 0.5,$$

$$\therefore\ GINI = 1 - [P(C_1)^2 + P(C_2)^2] = 0.5$$

Splitting Based On GINI

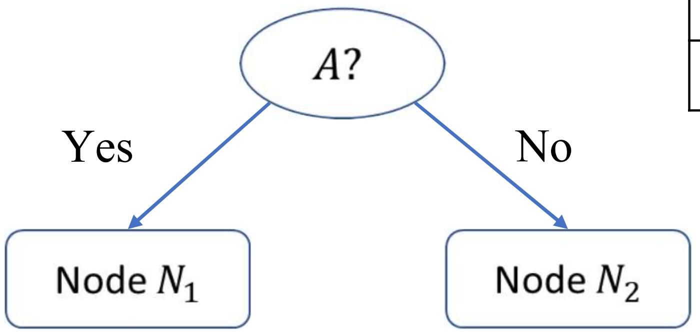
- When a parent node p is split into k partitions (children), the GINI of split is following:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Where n_i = number of samples at Child-i $\quad$ n = number of samples at Node p

C1	6
C2	6

Distribution in Parent Node
GINI = 0.5



Node N_1:

C1	5
C2	2

Node N_2:

C1	1
C2	4

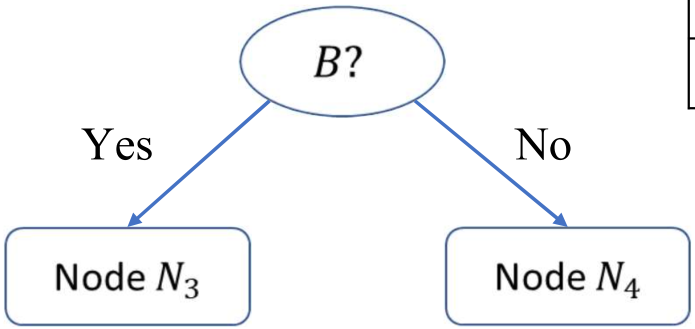
GINI in Node N_1: $1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = 0.408$

GINI in Node N_2: $1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$

$GINI_{split}^{(A)}$: $\left(\frac{7}{12}\right) * 0.408 + \left(\frac{5}{12}\right) * 0.32 = 0.3713$

Splitting Based On GINI

- Now let’s consider another attribute for splitting:



C1	3
C2	2

C1	3
C2	4

C1	6
C2	6

Distribution in Parent Node
GINI = 0.5

GINI in Node N_1: $1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$

GINI in Node N_2: $1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.49$

$GINI_{split}^{(B)}: \left(\frac{5}{12}\right) * 0.48 + \left(\frac{7}{12}\right) * 0.49 = 0.4858$

- Gain while splitting with Attribute A: $GINI_{parent} - GINI_{split}^{(A)} = 0.5 - 0.3713 = 0.1287$
- Gain while splitting with Attribute B: $GINI_{parent} - GINI_{split}^{(B)} = 0.5 - 0.4858 = 0.0142$

Hence, we shall split the parent node based on attribute A as it provides more gain.

Another Measure Of Impurity: Entropy

The formula for entropy at a particular node t is:

$$Entropy(t) = -\sum_{j} p(j|t)\log_2\big(p(j|t)\big)$$

Where $p(j|t)$ is the relative frequency of $class - j$ at node t.

- Maximum Value = $\log_2(n_c)$, where n_c is the number of classes. It occurs when all the samples at the node are equally distributed among the classes, implies high impurity.

- Minimum Value = 0, occurs when all the samples at the node belong to one class. Implies homogeneity.

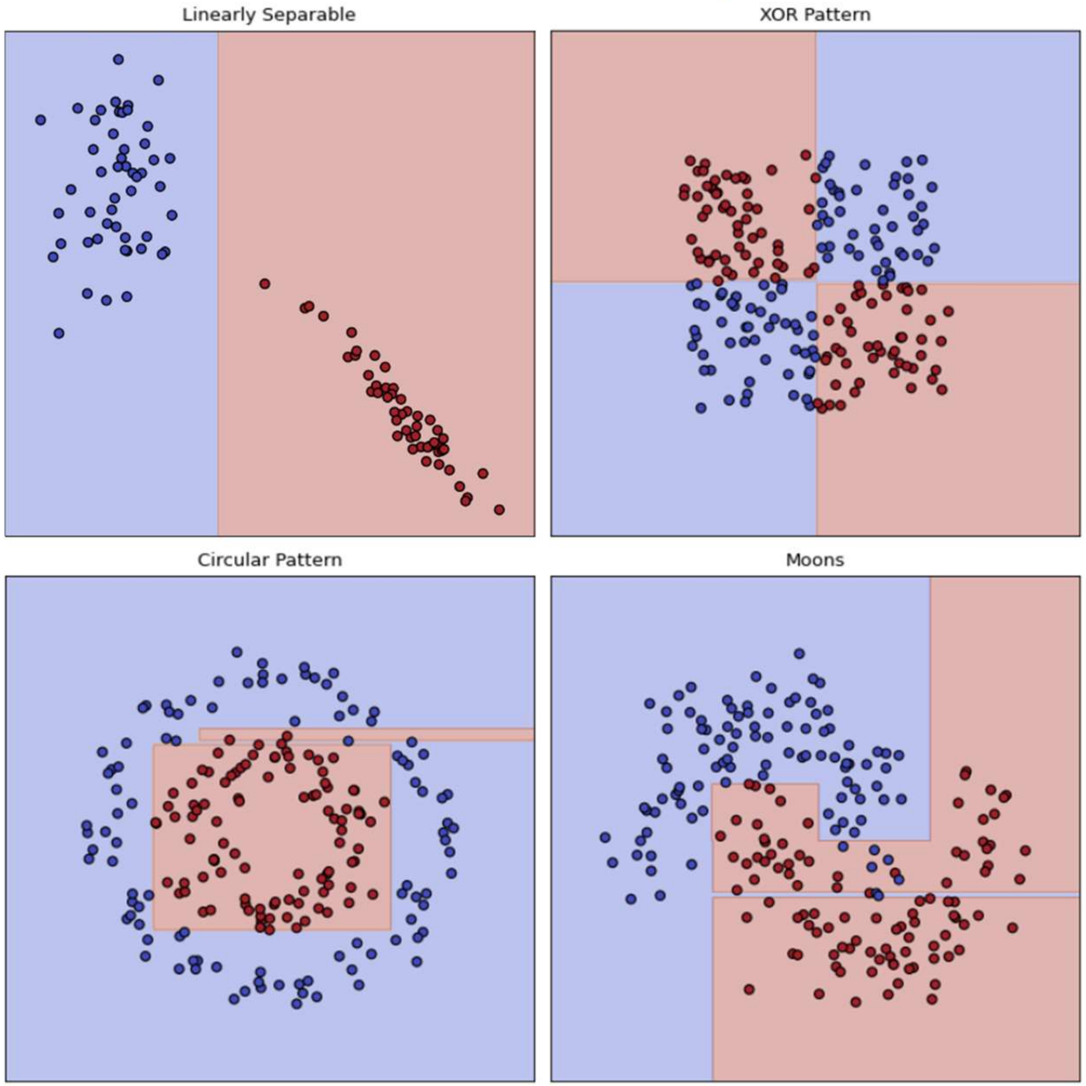
Information Gain (Entropy)

Information gain from the split is calculated as:

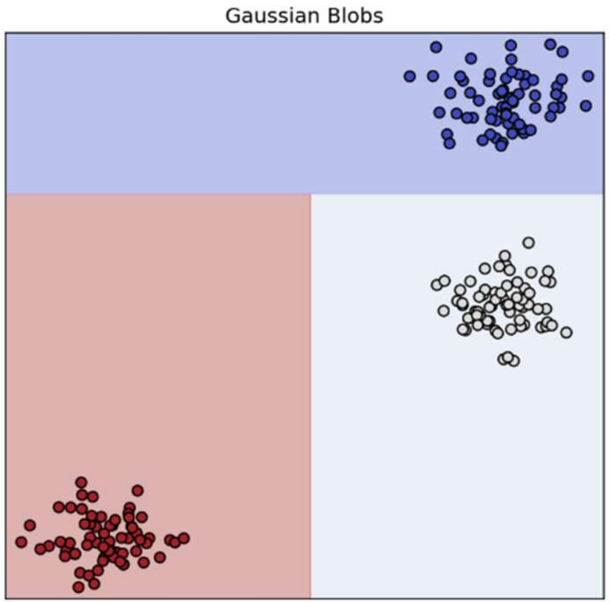
$$Information\ Gain = Entropy\ (parent)\ - \sum_{i=1}^{k} \frac{n_i}{n}\ Entropy(child_i)$$

Where k is the number of child nodes (partitions), n is the total number of data points, and n_i is the number of data points in child node i. The goal is to maximize information gain.

Decision Boundary of a Decision Tree

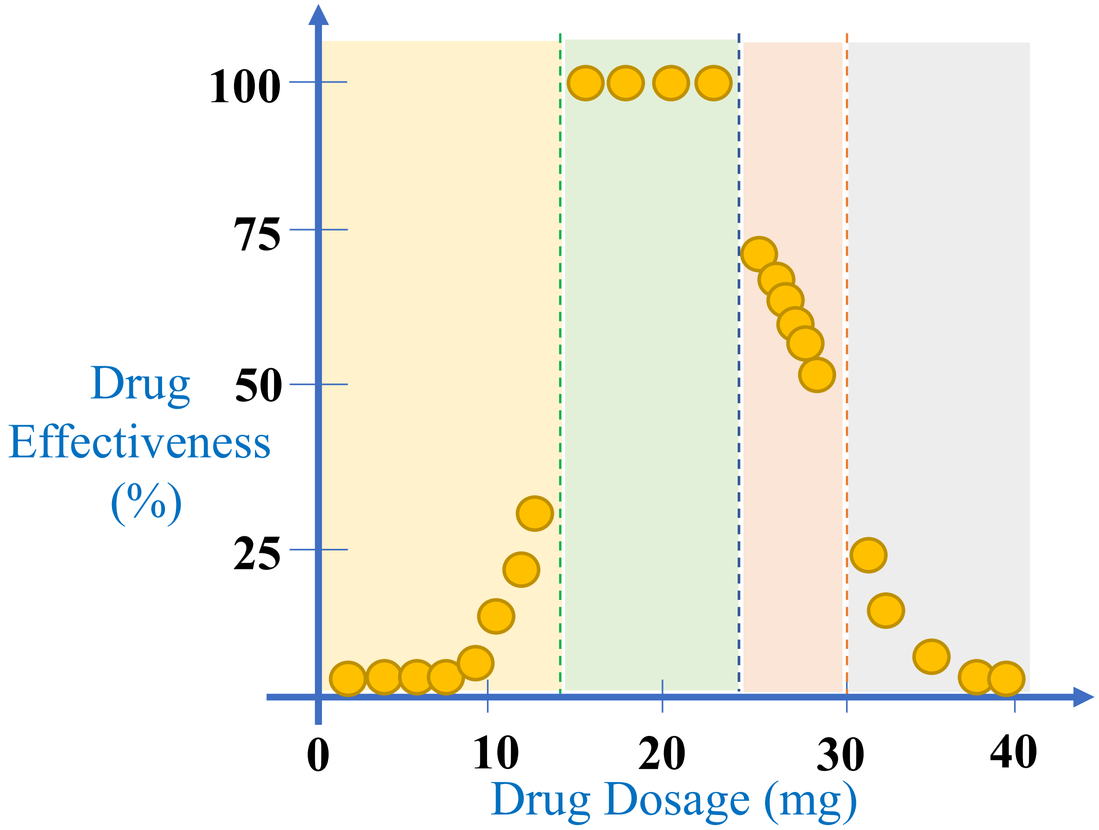


- A **decision tree** creates **axis-aligned** decision boundaries based on feature values. Each split in the tree corresponds to a threshold on a single feature, resulting in rectangular regions in the feature space.

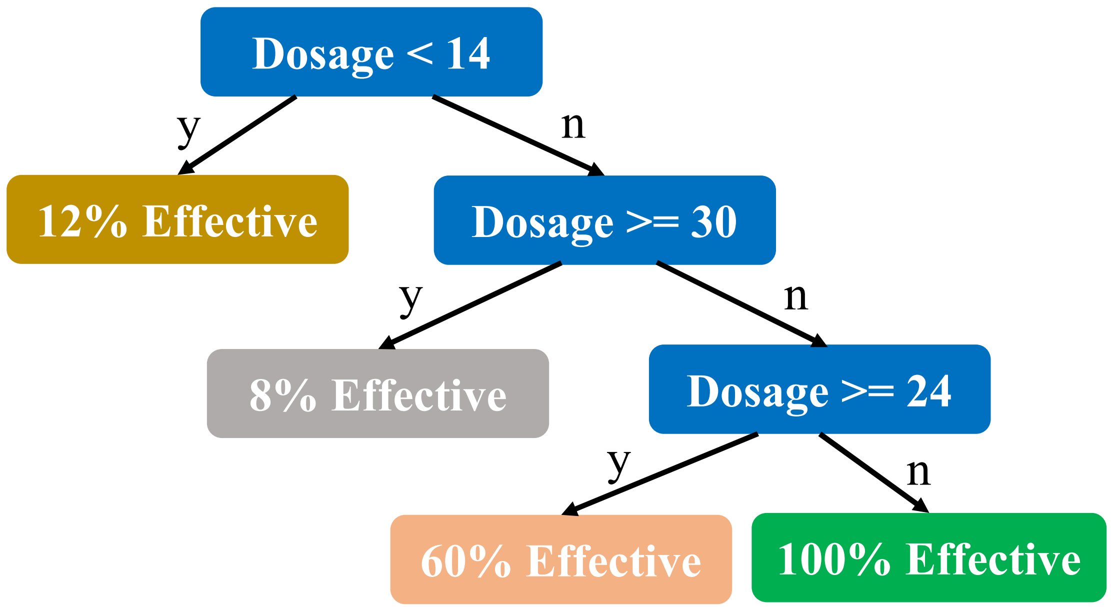


Decision Tree Regression

- Suppose, the clinical trial data, where we want to determine the effectiveness of a drug vs the dosage (mg) looks like following:



- In this case, fitting a straight line to the data will not be very useful.
- One option is to use a **Regression Tree**.



- A Decision Tree Regressor doesn't try to fit a single, global model. Instead, it partitions the data into smaller, more manageable subsets and fits a simple model (the average value) to each.

Decision Tree Regression: Impurity Measure

- The core of a decision tree algorithm is to find the best way to split the data at each node. For regression, "impurity" refers to the variance of the target values within a node. The goal is to find splits that minimize this **variance**, creating child nodes that are as homogeneous as possible.

- **The Mean Squared Error (MSE) Metric:** The most common metric for measuring impurity in a regression tree is the Mean Squared Error (MSE). A lower MSE indicates that the target values in the node are closer to their average, meaning the node is "purer."

For a given node t containing a set of target values $\{y_1, y_2, y_3, \dots, y_{n_t}\}$, the MSE is calculated as:

$$MSE = \frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \bar{y})^2 \quad ; \quad where, \bar{y} = \frac{1}{n_t}\sum_{i=1}^{n_t} y_i$$

Where, n_t is the number of samples in the node.

When To Stop Splitting

- Growing of decision tree (Tree Induction Algorithm) can be stopped by following criteria:
  - When there are no records / samples to split further, that is when each of the training sample belong to one of the leaf nodes.
  - When the leaf nodes are homogeneous or nearly homogeneous for classification.
  - When the tree height is equal to some predefined height.
  - When the gain of split at a node is not more than a predefined value for classification or when the MSE of a node falls below a specific threshold for regression.

Decision Tree : Merits And Demerits

Merits:

- Inexpensive to construct, i.e. it takes less time and effort to build a decision tree.
- Extremely fast at classifying unknown records.
- Highly interpretable model for small-sized tree.
- Can work on both categorical and quantitative attributes.
- Feature scaling and normalization is not required.
- Somewhat robust to missing values.

Demerits:

- Highly prone to overfit (i.e. low bias but high variance).
- A small change in the training data can cause a large change in the structure of the decision tree causing instability.

Thank You