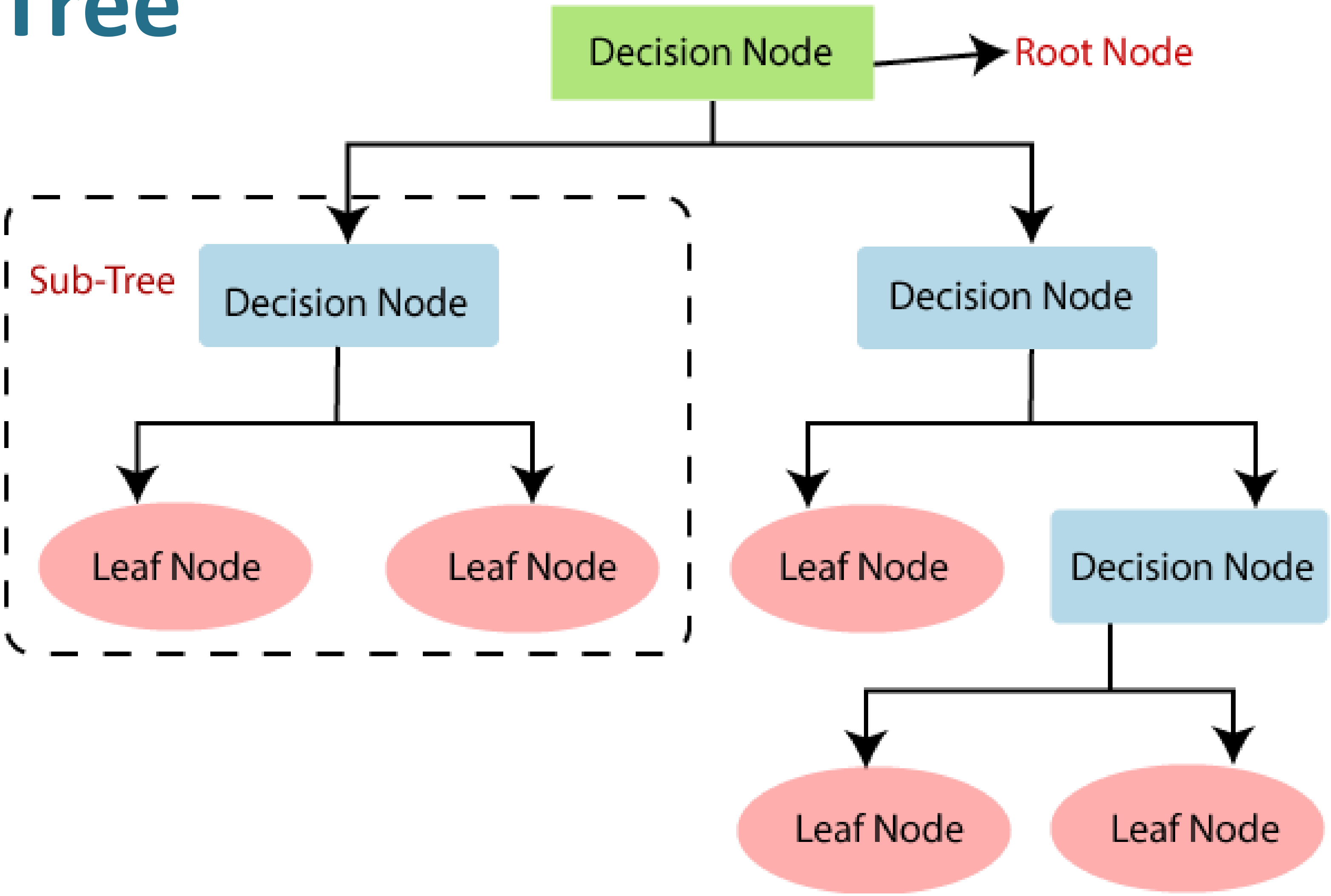
Decision Tree



By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Decision Tree

How Decision Trees Work?

The process of creating a decision tree involves:

1. **Selecting the Best Attribute**: Using a metric like Gini impurity, entropy, or information gain, the best attribute to split the data is selected.
2. **Splitting the Dataset**: The dataset is split into subsets based on the selected attribute.
3. **Repeating the Process**: The process is repeated recursively for each subset, creating a new internal node or leaf node until a stopping criterion is met (e.g., all instances in a node belong to the same class or a predefined depth is reached)

Decision Tree - Metrics for Splitting

Gini Impurity: Measures the amount of uncertainty or impurity in the dataset.

$$Gini = 1 - \sum (p_i)^2$$

where *pi* is the probability of an instance being classified into a particular class.

Entropy: Measures the amount of uncertainty or impurity in the dataset.

$$Entropy = -\sum p_i \log_2(p_i)$$

where *pi* is the probability of an instance being classified into a particular class.

Information Gain: Measures the reduction in entropy or Gini impurity after a dataset is split on an attribute.

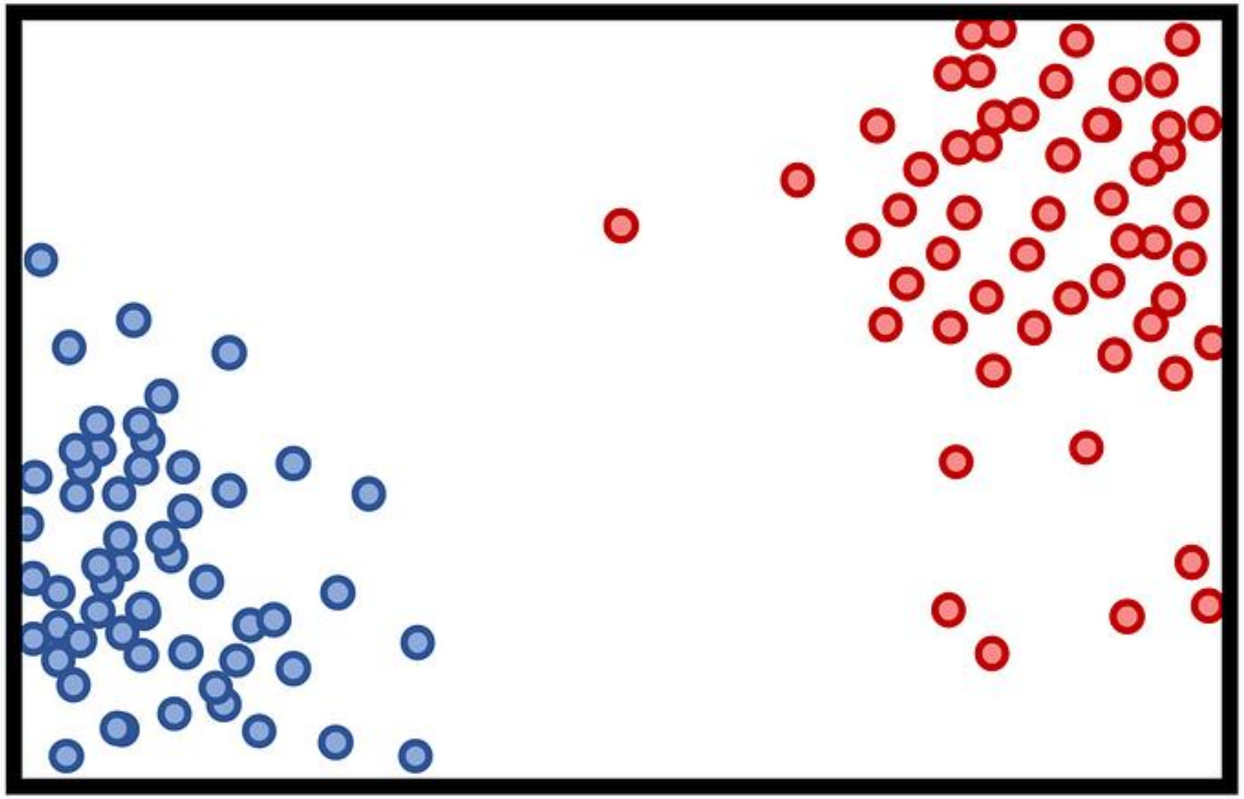
Decision Tree

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

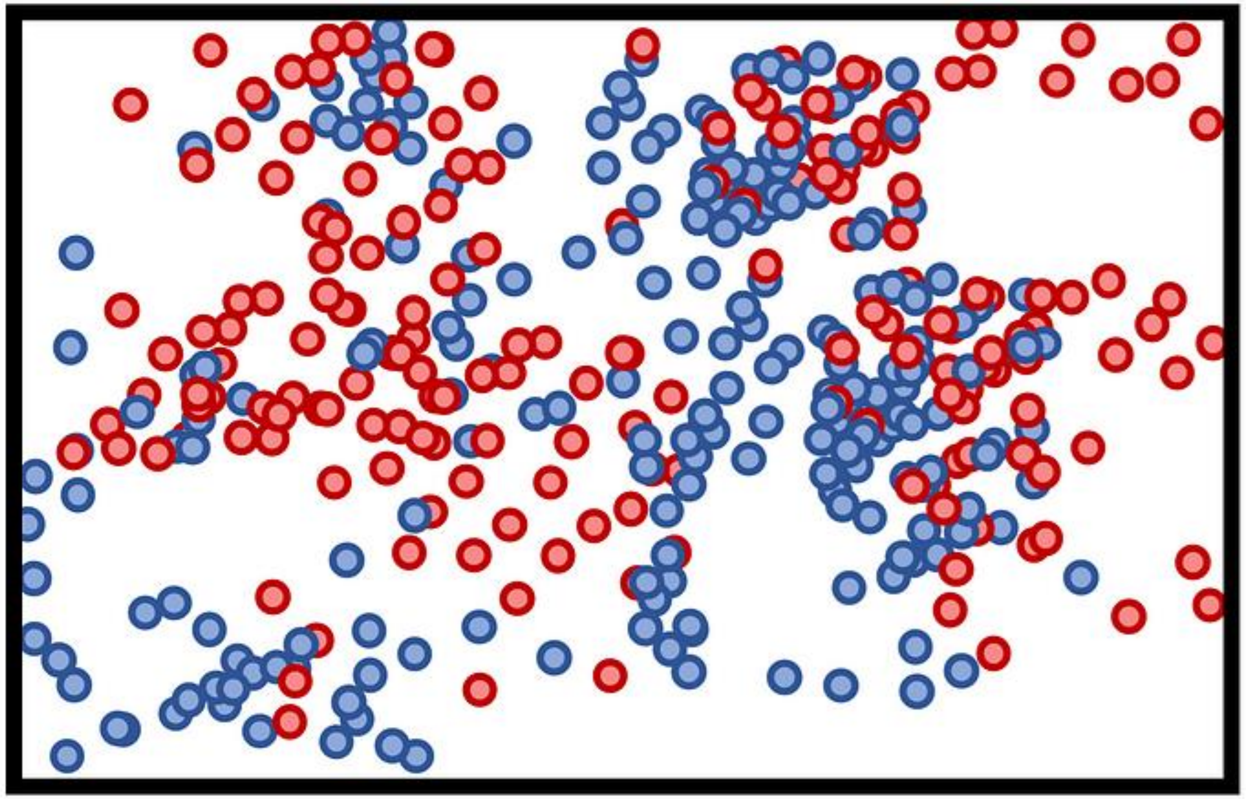
Entropy

Entropy is a concept that stems from information theory, which measures the impurity of the sample values. It is defined with by the following formula, where:

By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar



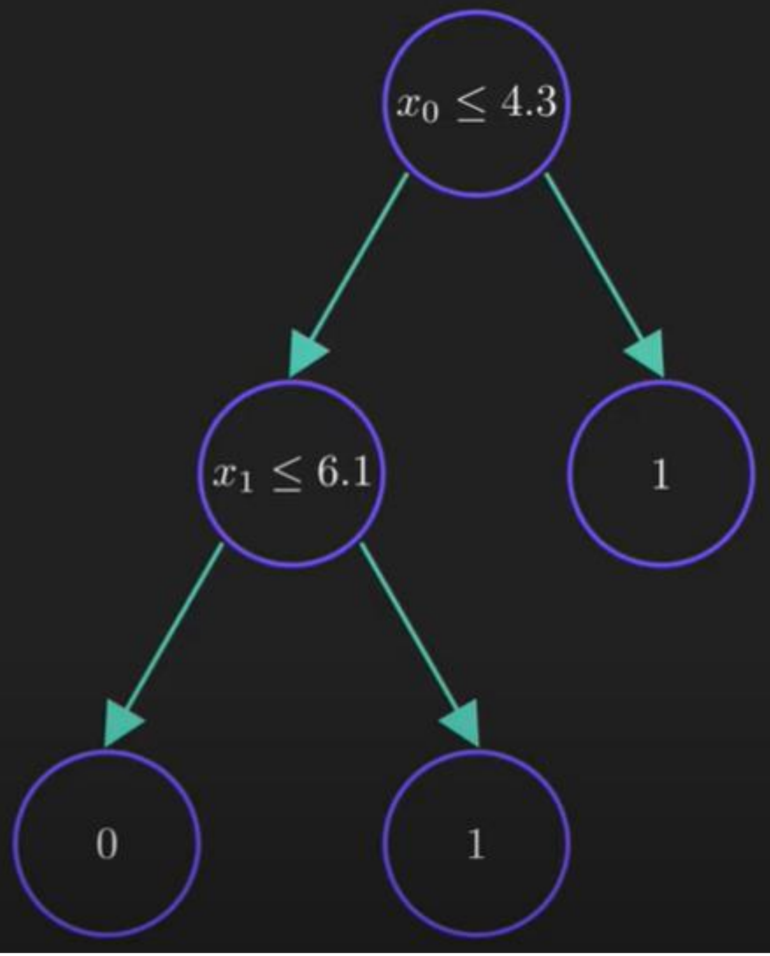
Low Entropy



High Entropy

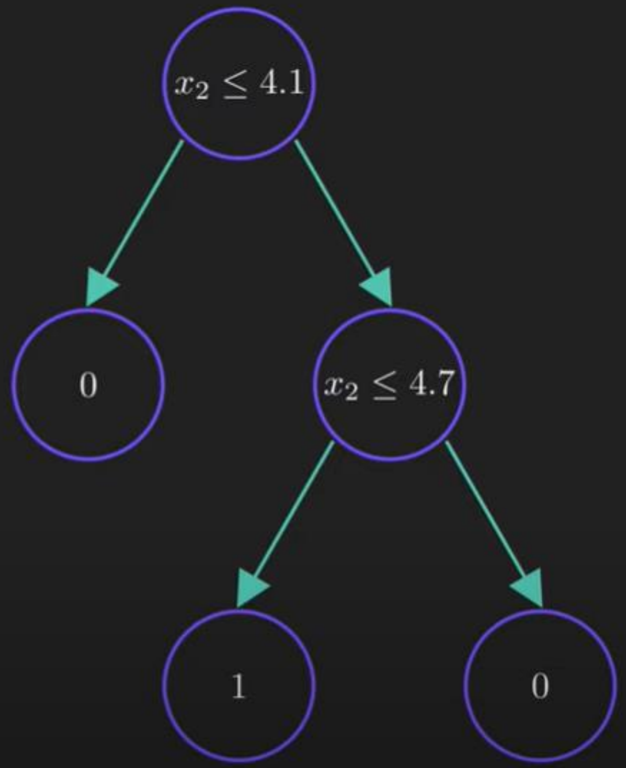
Decision Tree

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
[redacted]						
5	2.7	6.7	4.2	5.3	4.8	1



Decision Tree

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Decision Tree - *Entropy*

$$Entropy([9+,5-]) = -(\frac{9}{14}log_2\frac{9}{14} + \frac{5}{14}log_2\frac{5}{14}) = 0.940 \qquad (1.2)$$

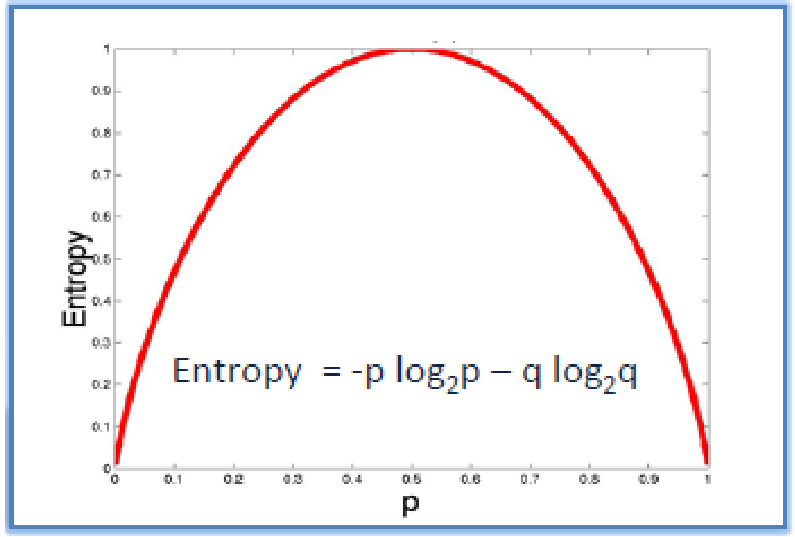
Which concludes, the dataset is 94% impure or 94% non-homogeneous.

Let's do some more calculations and try to understand the nature of *Entropy*.
What could be the Entropy of [7+,7-] & [14+,0-]?

$$Entropy[7+,7-] = -(\frac{7}{14}log_2\frac{7}{14} + \frac{7}{14}log_2\frac{7}{14}) = 1$$

And,

$$Entropy[14+,0-] = -(\frac{14}{14}log_2\frac{14}{14} + \frac{0}{14}log_2\frac{0}{14}) = 0$$



Gini Impurity

Gini impurity is the probability of incorrectly classifying a random data point in a dataset. It is an impurity metric since it shows how the model differs from a pure division.

$$Gini = 1 - \sum_j p_j^2$$

By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Gini Impurity

Student Background	Online Courses	Working	Decision
Math	Yes	W	Hire
Math	No	W	Reject
CS	Yes	W	Hire
IT	Yes	NW	Hire
IT	No	W	Train
IT	Yes	NW	Hire
CS	No	NW	Train
CS	No	W	Hire
CS	Yes	W	Hire
Math	No	W	Reject

- Hire - 6 instances.
- Reject - 2 instances.
- Train - 2 instances.

So the Gini Impurity on Decision will be:

$$Gini(S) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right] = 0.56$$

Decision Tree - How to choose the best attribute at each node

- Entropy values can fall between 0 and 1. If all samples in data set, S, belong to one class, then entropy will equal zero. If half of the samples are classified as one class and the other half are in another class, entropy will be at its highest at 1.
- Information gain represents the difference in entropy before and after a split on a given attribute. **The attribute with the highest information gain will produce the best split as it's doing the best job at classifying the training data according to its target classification**.

Information Gain

information gain, is simply the expected reduction in entropy caused by partitioning the data set. The information gain of an attribute A relative to a collection of data set S, is defined as-

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where, $Values(A)$ is the all possible values for attribute A, and S_v is the subset of S for which attribute A has value v.

Information Gain

$$Values(Outlook) = Sunny, Overcast, Rain$$
$$S = [9+, 5-]$$
$$S_{\text{sunny}} = [2+, 3-]$$
$$S_{\text{overcast}} = [4+, 0-]$$
$$S_{\text{rain}} = [3+, 2-]$$

$$G(S, Outlook) = Entropy(S) - \left(\frac{5}{14}Entropy(S_{\text{sunny}}) + \frac{4}{14}Entropy(S_{\text{overcast}}) + \frac{5}{14}Entropy(S_{\text{rain}})\right)$$

$$Entropy(S_{\text{sunny}}) = -\left(\frac{2}{5}log_2\frac{2}{5} + \frac{3}{5}log_2\frac{3}{5}\right) = 0.971$$

$$Entropy(S_{\text{overcast}}) = -\left(\frac{4}{4}log_2\frac{4}{4} + \frac{0}{4}log_2\frac{0}{4}\right) = 0$$

$$Entropy(S_{\text{rain}}) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

Information Gain

Wind	Play Tennis
Weak	No
Strong	No
Weak	Yes
Weak	Yes
Weak	Yes
Strong	No
Strong	Yes
Weak	No
Weak	Yes
Weak	Yes
Strong	Yes
Strong	Yes
Weak	Yes
Strong	No

$S = [9+, 5-]$
$S_{\text{weak}} = [6+, 2-]$
$S_{\text{strong}} = [3+, 3-]$
$Entropy(S) = 0.940$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - (\frac{8}{14} Entropy(S_{\text{weak}}) + \frac{6}{14} Entropy(S_{\text{strong}}))$$

$$Entropy(S_{\text{weak}}) = -(\frac{6}{8} log_2 \frac{6}{8} + \frac{2}{8} log_2 \frac{2}{8}) = 0.811$$

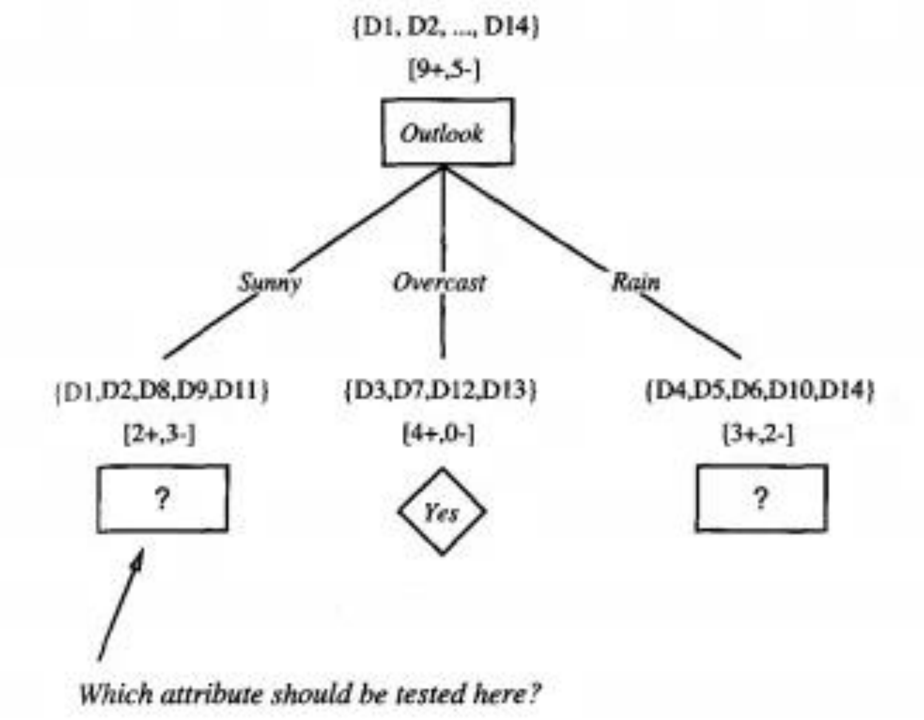
$$Entropy(S_{\text{strong}}) = -(\frac{3}{3} log_2 \frac{3}{3} + \frac{3}{3} log_2 \frac{3}{3}) = 1.00$$

Put the values of $Entropy(S_{\text{weak}})$ and $Entropy(S_{\text{strong}})$ in eqtn 1.6

$$\begin{aligned} Gain(S, Wind) &= Entropy(S) - (\frac{8}{14} 0.811 + \frac{6}{14} 1.00) \\ &= 0.940 - (0.463 + 0.429) \\ &= 0.048 \end{aligned}$$

Information Gain

The most useful attribute is "Outlook" as it is giving us more information than others. So, "Outlook" will be the root of our tree.



$Gain(S, Outlook) = 0.246$
$Gain(S, Humidity) = 0.151$
$Gain(S, Wind) = 0.048$
$Gain(S, Temperature) = 0.029$

By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Information Gain for second level

$S_{suuny} = 5 = S$
$Humidity = High, Normal$
$Humidity_{high} = [0+, 3-]$
$Humidity_{normal} = [2+, 0-]$
$Gain(S, Humidity) = ?$

$$Gain(S_{sunny}, Humidity) = Entropy(S) - (\frac{3}{5} Entropy(Humidity_{high}) + \frac{2}{5} Humidity_{normal}) \tag{1.15}$$

$$Entropy(Humidity_{high}) = -(\frac{0}{3} log_2 \frac{0}{3} + \frac{3}{3} log_2 \frac{3}{3}) = 0 \tag{1.16}$$

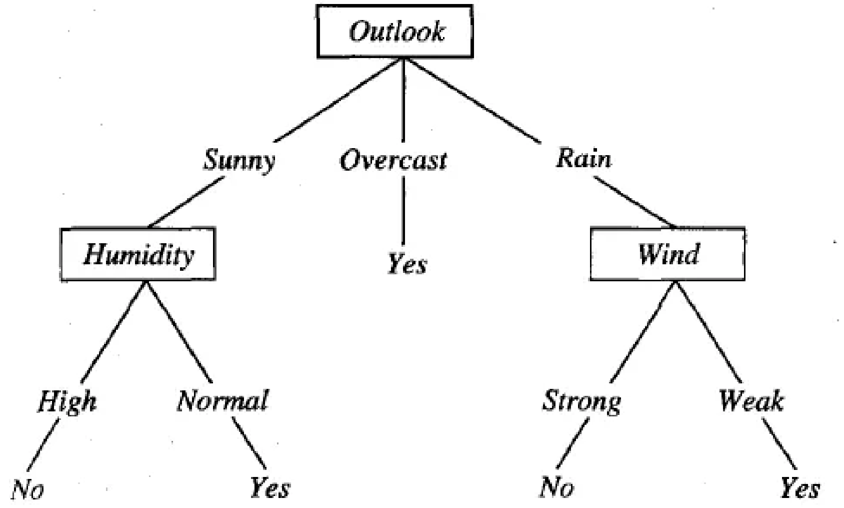
$$Entropy(Humidity_{normal}) = -(\frac{2}{2} log_2 \frac{2}{2} + \frac{0}{2} log_2 \frac{0}{2}) = 0 \tag{1.17}$$

Put the values in eqtn 1.15

$$\begin{aligned} Gain(S_{sunny}, Humidity) &= Entropy(S) - (\frac{3}{5} 0 + \frac{2}{5} 0) \\ &= 0.970 - 0 \\ &= 0.970 \end{aligned} \tag{1.18}$$

Information Gain for second level

$$Gain(S, Humidity) = 0.970$$
$$Gain(S, Temperature) = 0.570$$
$$Gain(S, Wind) = 0.019$$



So Humidity gives us the most information at this stage. The node after "Outlook" at Sunny descendant will be **Humidity**. The **High** descendant has only negative examples and the **Normal** descendant has only positive examples. So both of them become the leaf node and can not be furthered expanded. If we expand the **Rain** descendant by the same procedure we will see that the **Wind** attribute is providing most information.

Decision Tree

Advantages

-Easy to interpret: The Boolean logic and visual representations of decision trees make them easier to understand and consume. The hierarchical nature of a decision tree also makes it easy to see which attributes are most important, which isn't always clear with other algorithms, like neural networks.

-Little to no data preparation required: Decision trees have a number of characteristics, which make it more flexible than other classifiers. It can handle various data types—i.e. discrete or continuous values, and continuous values can be converted into categorical values through the use of thresholds. Additionally, it can also handle values with missing values, which can be problematic for other classifiers, like Naïve Bayes.

-More flexible: Decision trees can be leveraged for both classification and regression tasks, making it more flexible than some other algorithms. It's also insensitive to underlying relationships between attributes; this means that if two variables are highly correlated, the algorithm will only choose one of the features to split on.

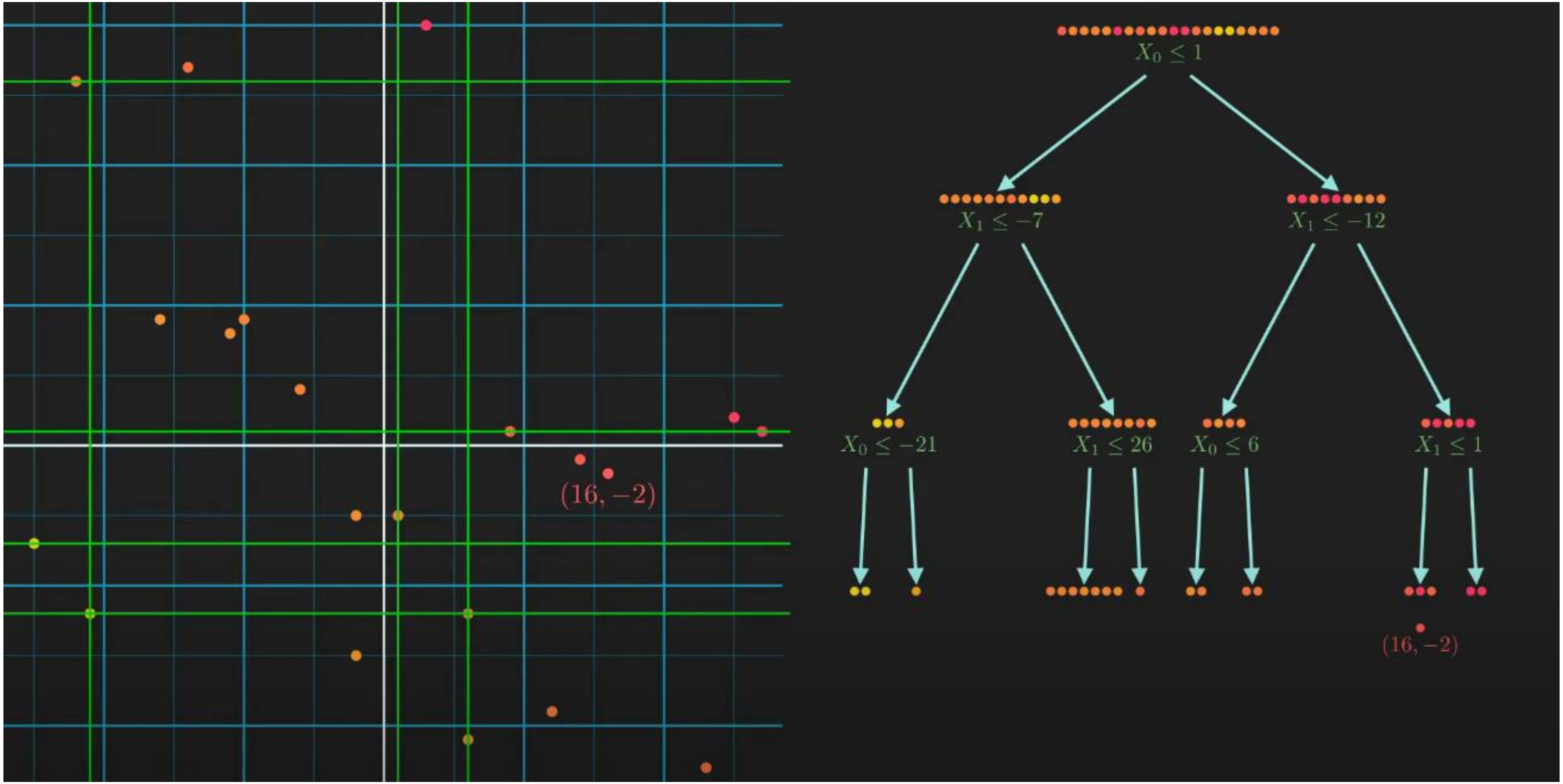
Decision Tree

Disadvantages

- **Prone to overfitting:** Complex decision trees tend to overfit and do not generalize well to new data. This scenario can be avoided through the processes of pre-pruning or post-pruning. Pre-pruning halts tree growth when there is insufficient data while post-pruning removes subtrees with inadequate data after tree construction.

- **High variance estimators:** Small variations within data can produce a very different decision tree. Bagging, or the averaging of estimates, can be a method of reducing variance of decision trees. However, this approach is limited as it can lead to highly correlated predictors.

- **More costly:** Given that decision trees take a greedy search approach during construction, they can be more expensive to train compared to other algorithms.



By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Random Forest

Random Forest is a popular ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. It's a versatile algorithm that can be applied to both classification and regression problems.

Random Forest

How it works ?

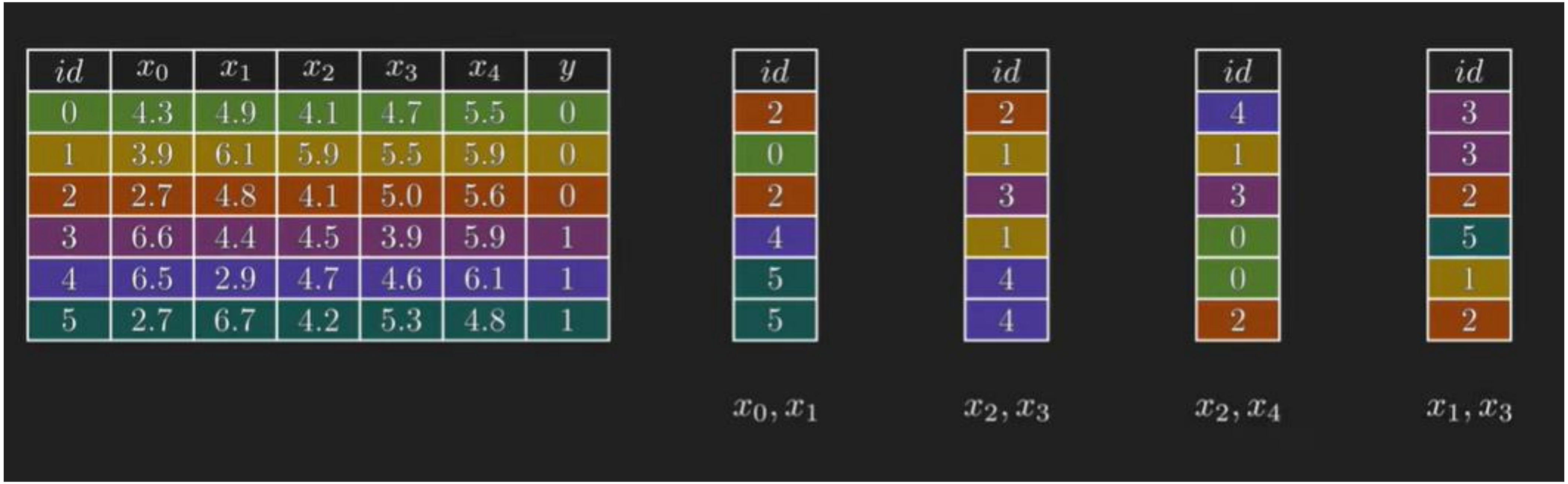
- **Bootstrap Sampling:** The algorithm randomly selects multiple samples (with replacement) from the original dataset. Each sample is used to train a separate decision tree.

- **Feature Randomization:** At each node of each decision tree, only a random subset of features is considered for splitting. This helps to decorrelate the trees and reduce overfitting.

- **Tree Growth:** Each decision tree is grown to its maximum depth without pruning.

- **Prediction:** To make a prediction for a new instance, the predictions of all trees are combined. For classification, the most frequent class among the predictions is chosen. For regression, the average of the predictions is taken.

Random Forest



id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id	id	id	id
2	2	4	3
0	1	1	3
2	3	3	2
4	1	0	5
5	4	0	1
5	4	2	2
x_0, x_1	x_2, x_3	x_2, x_4	x_1, x_3

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

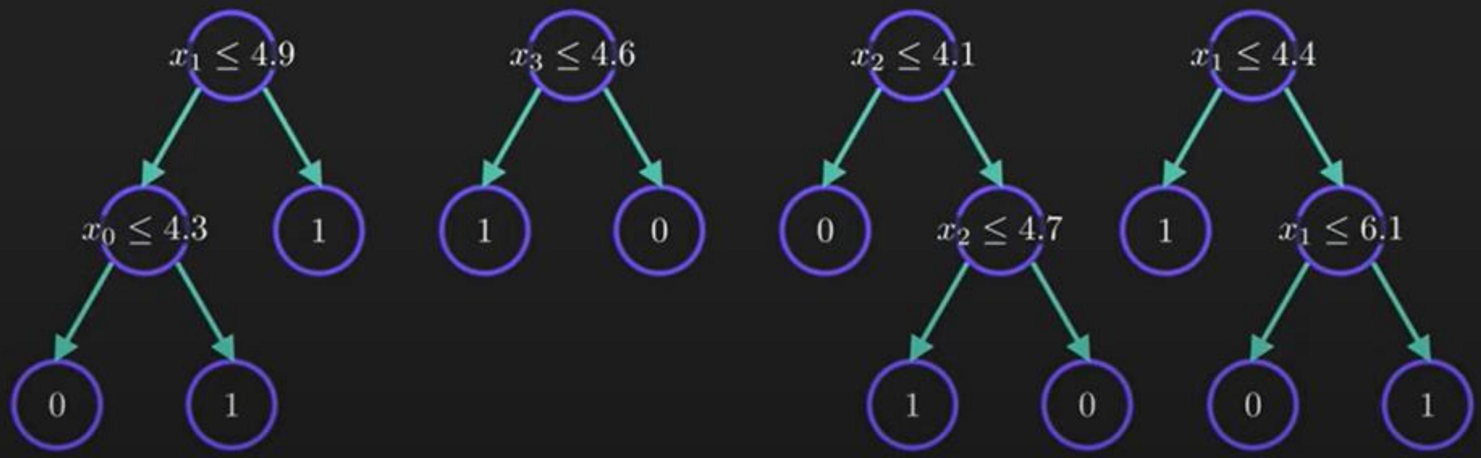
id	id	id	id
2	2	4	3
0	1	1	3
2	3	3	2
4	1	0	5
5	4	0	1
5	4	2	2

x_0, x_1	x_2, x_3	x_2, x_4	x_1, x_3

2.8	6.2	4.3	5.3	5.5

Bootstrap + Aggregating

(Bagging)



1	0	1	1

By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Random Forest

Advantages

- **Reduced Overfitting:** By averaging the predictions of multiple trees, random forests can help to reduce overfitting, which is a common problem with decision trees.

- **Improved Accuracy:** Random forests often achieve higher accuracy than individual decision trees, especially on large and complex datasets.

- **Robustness:** Random forests are relatively robust to noise and outliers in the data.

- **Feature Importance:** The algorithm can be used to assess the importance of different features in the prediction task.

Ensemble methods Random Forest

Disadvantage

- **Computational Cost:** Training a random forest can be computationally expensive, especially for large datasets with many features.

- **Interpretability:** While random forests can be more accurate than individual decision trees, they can be less interpretable due to the complexity of the ensemble.

FEATURE GENERATION

PATIENT ID	PATIENT AGE	NUMBER OF DIAGNOSES
55629189	15	9
86057875	25	6
82442376	35	7
42519267	45	5
82637451	55	9
114882984	65	7
48330782	75	8

AGE X NUMBER DIAGNOSES
135
150
245
225
495
455
600

Feature Generation

Feature Generation (also known as **feature construction**) is the process of transforming features into new features that better relate to the target.

Examples of Feature Generation techniques

A transformation is a mapping that is used to transform a feature into a new feature. The right transformation depends on the type and structure of the data, data size and the goal. This can involve transforming single feature into a new feature using **standard operators like log, square, power, exponential, reciprocal, addition, division, multiplication** etc.

Why Feature Generation?

Enhanced Model Performance: Well-crafted features can significantly improve a model's ability to learn and make accurate predictions.

Reduced Feature Engineering: By generating informative features, you might be able to reduce the need for extensive feature engineering.

Better Interpretability: Generated features can sometimes provide insights into the underlying relationships between variables.

Example: Feature Generation

For a Customer Churn Prediction Model

Given a dataset with customer information (age, tenure, monthly bill, etc.)

You could generate new features like:

Customer tenure in month: Divide tenure by 12.

Around Monthly bill per year: Multiply monthly bill by 12.

Age group: Categorize age into bins (e.g., young, middle-aged, elderly).

Interaction between tenure and monthly bill: Multiply tenure by monthly bill.

Tips for Effective Feature Generation

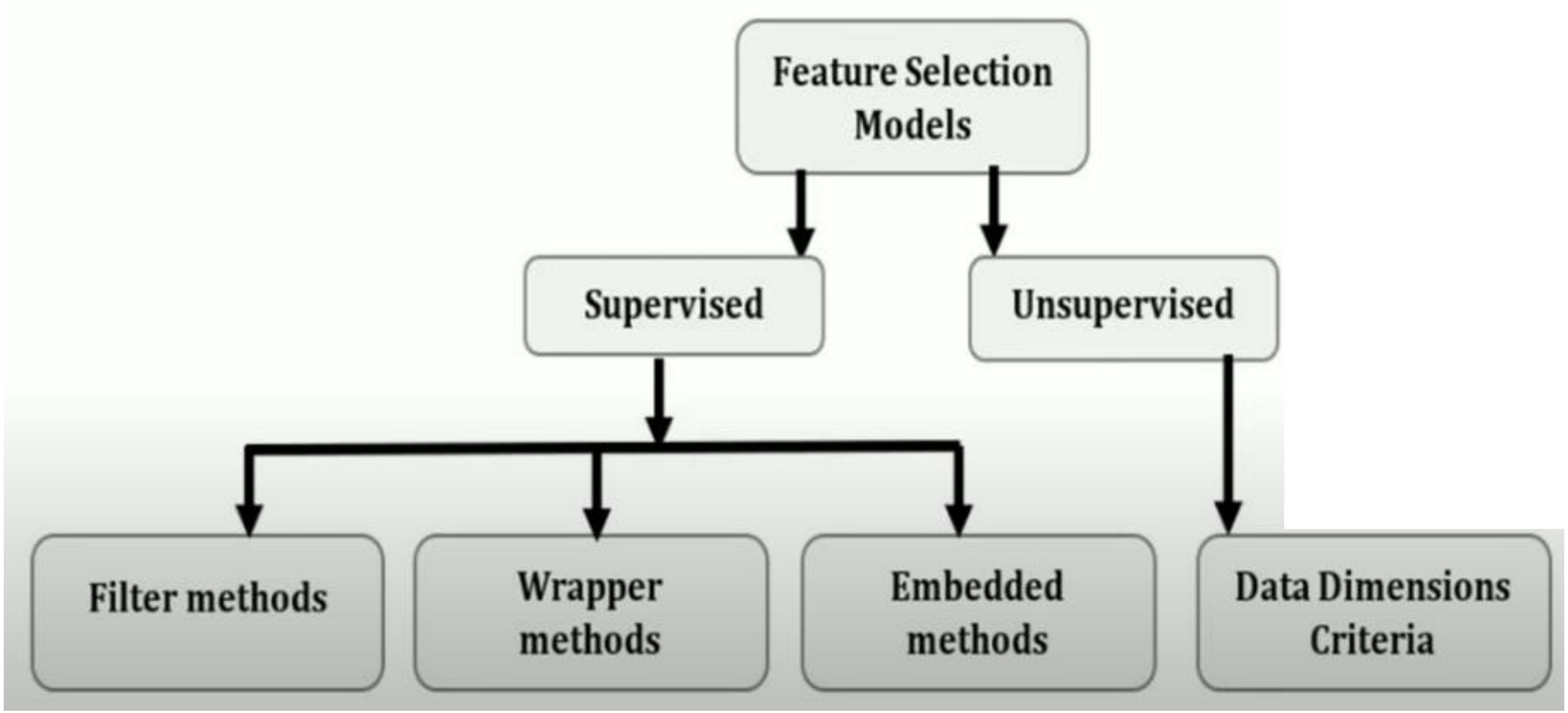
Domain Knowledge: Leverage your understanding of the problem domain to create meaningful features.

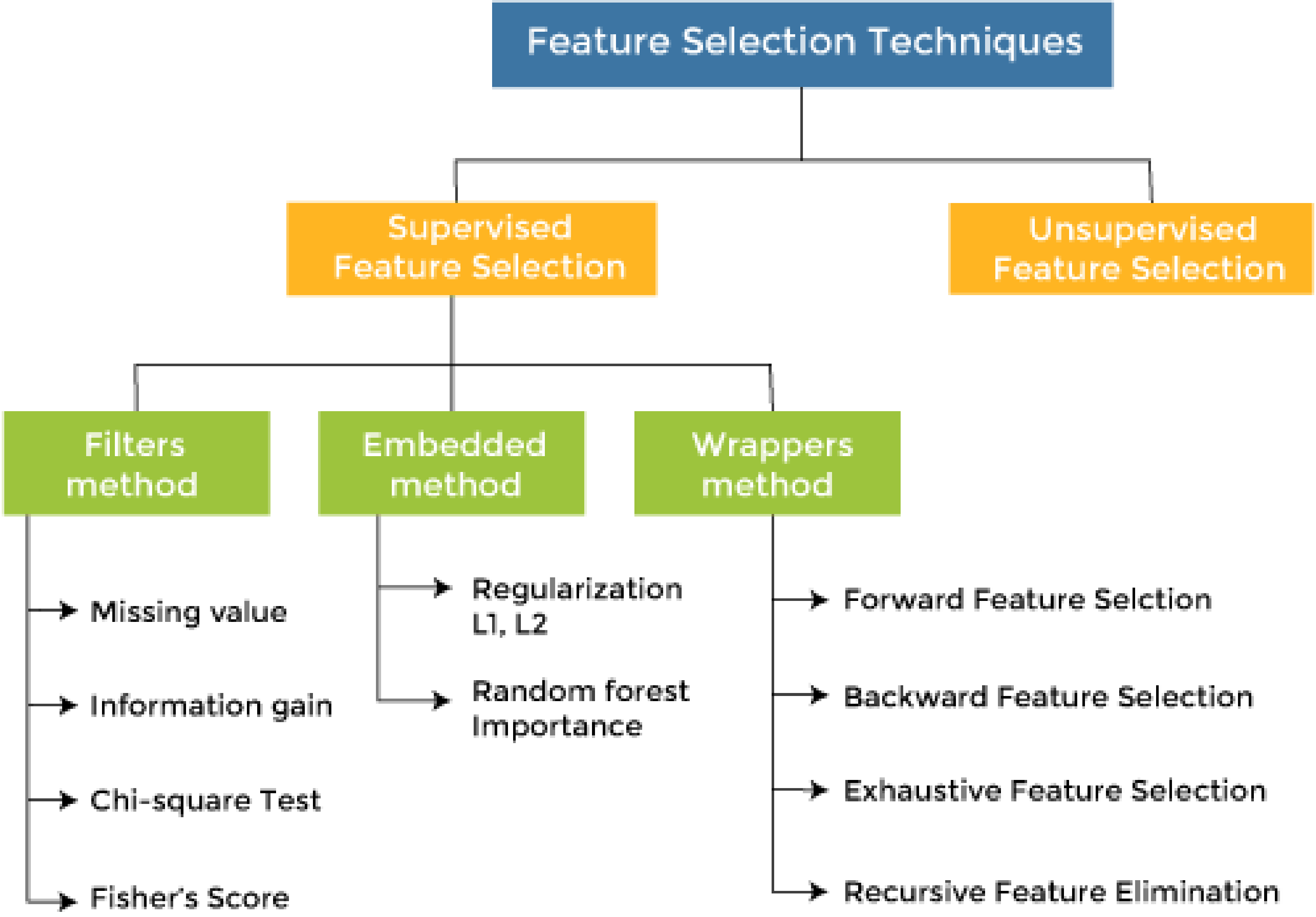
Experimentation: Try different feature generation techniques and evaluate their impact on model performance.

Feature Selection: After generating new features, consider using feature selection techniques to identify the most relevant ones.

Avoid Overfitting: Be cautious of creating too many features, as it can lead to overfitting.

Feature Selection algorithms




Feature Selection Techniques
Supervised Feature Selection
Unsupervised Feature Selection
Filters method
Embedded method
Wrappers method
Missing value
Information gain
Chi-square Test
Fisher's Score
Regularization L1, L2
Random forest Importance
Forward Feature Selction
Backward Feature Selection
Exhaustive Feature Selection
Recursive Feature Elimination

Feature Selection Methods

Method	Description	Advantages
Filter	Uses statistical tests to rank features	Simple and fast
Wrapper	Trains a model on different subsets of features	Accurate
Embedded	Selects features as part of the model training process	Accurate and efficient

Feature Selection

Advantages:

- Improves accuracy of machine learning models
- Reduces overfitting
- Reduces training time
- Improves interpretability of machine learning models

Disadvantages:

- Can be computationally expensive
- May not always find the optimal subset of features

Filters

At buffet, you wouldn't just take everything on a plate right? Filter methods are just like this, only focusing on specific traits of each feature.

statistical tests and measures to identify features that seem relevant based on their correlation with target variable.

chi-square tests, information gain, and mutual dependence methods.

Objective : To have quick elimination of irrelevant features and in terms of process, features are evaluated based on statistical measures or mathematical functions.

when to use filter methods: A) when you have higher dimensional data with a larger number of features, B) the pre-processing step as it acts as a preliminary filter before diving into more intricate methods.

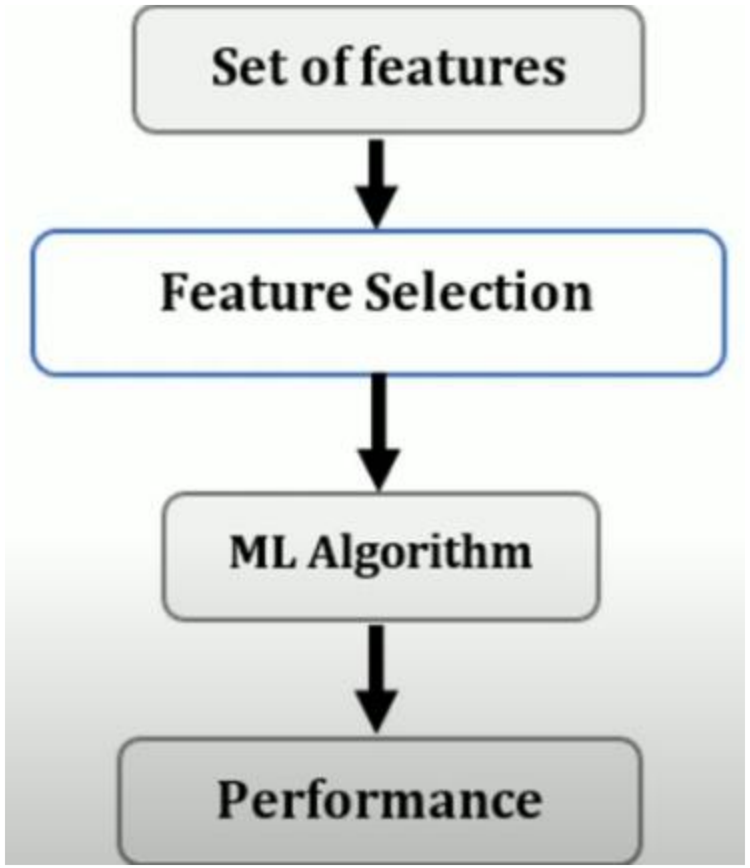
Filters

Advantages

- they are computationally cheaper as compared to others.
- they have the fastest running time with the ability of good generalization.
- It is also easily scaled to high dimensional data sets.

Disadvantages

- no interaction with classification models can happen for feature selection.
- it mostly ignores feature dependencies and considers each feature separately in case of univariate techniques.



Wrappers

It like trial and error.

Imagine building your plate one bit at a time, testing each feature and seeing how it affects your overall dining experience.

These methods train predictive models with different combinations of features and then choose the set that makes the tastiest model ideally choosing the one with the highest accuracy or lowest error.

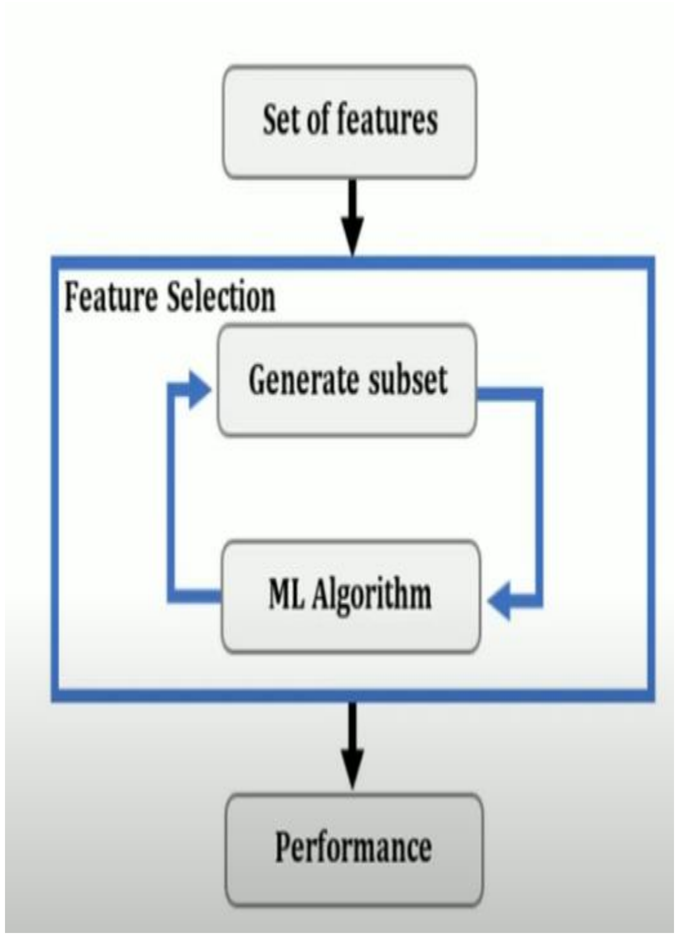
Methods : forward selection and recursive elimination.

Objective: To evaluate subsets of features as a group and in terms of process, it employs predictive models to assess feature subsets.

when to use wrapper methods:

A) when dealing with model-specific optimization

B) small to medium data sets.



By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Wrappers

Advantages:

- it interacts with the **classifier** for feature selection.
- more comprehensive search of feature set space can happen with it.
- it considers feature dependencies and is offering better generalization than filter approach.

Disadvantages:

- It surely has high **computational cost** alongside long running time.
- It also poses higher risk of **overfitting** as compared to filter and embedded methods.
- it is computationally more unfeasible with increased number of features.

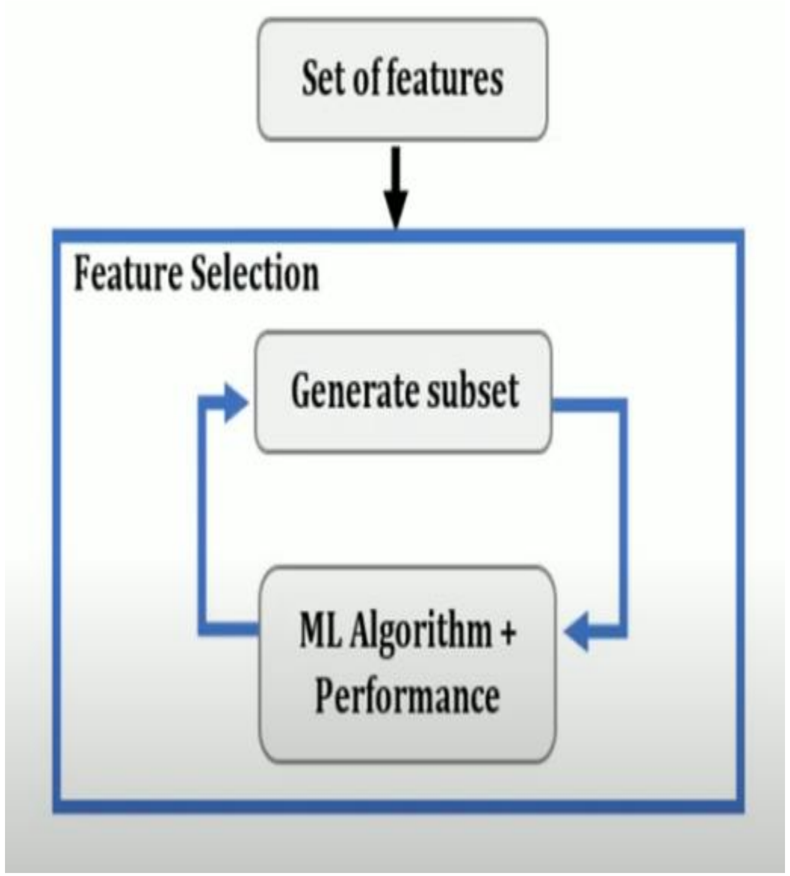
Embedded methods

Embedded methods blend the best of both worlds.

- They're like having a star chef guide you through the menu, highlighting hidden gems.
- These methods build the model and perform feature selection.
- They shrink the weight of irrelevant features or prune them all together, resulting in a leaner, meaner model.

Methods : tree-based methods

Objective: To incorporate feature selection within the model training process and in terms of process, features are selected as the model learns during the training.



Embedded methods

when to use wrapper methods:

A) when dealing with **integrated learning**,

B) when dealing with **large data sets** as it efficiently handles substantial amounts of data during the model building process.

By Ms. Drashti Garadharia
National Forensic Science University, Gandhinagar

Embedded methods

Advantages :

they're computationally less expensive as compared to wrapper methods.

They offer faster running time as compared to wrapper and interacts with the classification model for feature selection.

It offers a lower risk of overfitting as compared to wrapper.

Disadvantages:

The identification of a small set of features may be problematic in this method.