

Decision Tree



Course Overview

You are here...

Term	CDF	GCD	GCDAI	PGPDSAI
Term 1	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python
Term 2	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques
Term 3	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling
		Minor Project	Minor Project	Minor Project
Term 4		Machine Learning Foundation	Machine Learning Foundation	Machine Learning Foundation
Term 5		Machine Learning Intermediate	Machine Learning Intermediate	Machine Learning Intermediate
Term 6		Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)
		Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)
		Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)
		Capstone Project	Capstone Project	Capstone Project
Term 7		Bonus: Industrial ML (ML – 4 & 5)	Basics of AI, TensorFlow, and Keras	Basics of AI, TensorFlow, and Keras
Term 8			Deep Learning Foundation	Deep Learning Foundation
Term 9			NPL – I/CV – I	CV – I
Term 10			NLP – II/CV – II	NLP – I
			Capstone Project	Capstone Project
Term 11				CV – II
Term 12				NLP – II
				NLP – III + CV – III
				AutoVision & AutoNLP
				Building AI product

Term Context

- **Decision Tree** ← You are here...
- Random Forest
- Principal Component Analysis
- Naïve Bayes Classifier

Agenda

- | | |
|--|---|
| <input type="radio"/> Terminology Related to Trees | <input type="radio"/> CART Algorithm |
| <input type="radio"/> Decision Tree | <input type="radio"/> Gini Index |
| <input type="radio"/> Decision Tree Algorithms | <input type="radio"/> Steps to estimate Gini Index |
| <input type="radio"/> Attribute Selection Measures | <input checked="" type="radio"/> CART – Regression Example |
| <input type="radio"/> ID3 Algorithm | <input type="radio"/> Issues with Decision Trees |
| <input type="radio"/> Entropy & Information Gain | <input type="radio"/> Tree Pruning |
| <input type="radio"/> Steps to Estimate Entropy & Information Gain | <input type="radio"/> Decision Tree Applications |

CART – Regression Example

	X_1	X_2	X_3	X_4	y
#	Outlook	Temperature	Humidity	Windy	Hours Played
01	Rainy	Hot	High	False	26
02	Rainy	Hot	High	True	30
03	Overcast	Hot	High	False	46
04	Sunny	Mild	High	False	45
05	Sunny	Cool	Normal	False	52
06	Sunny	Cool	Normal	True	23
07	Overcast	Cool	Normal	True	43
08	Rainy	Mild	High	False	35
09	Rainy	Cool	Normal	False	38
10	Sunny	Mild	Normal	False	46
11	Rainy	Mild	Normal	True	48
12	Overcast	Mild	High	True	52
13	Overcast	Hot	Normal	False	44
14	Sunny	Mild	High	True	30

- In case of regression, standard deviation is used to calculate the homogeneity of a numerical sample.
- If the numerical sample is completely homogeneous, then its standard deviation is zero. *

Predictors: Outlook, Temperature, Humidity, Windy

Target: Hours Played

Decision Tree Regression Working

$n=14$

#	Outlook	Temperature	Humidity	Windy	Hours Played
01	Rainy	Hot	High	False	26
02	Rainy	Hot	High	True	30
03	Overcast	Hot	High	False	46
04	Sunny	Mild	High	False	45
05	Sunny	Cool	Normal	False	52
06	Sunny	Cool	Normal	True	23
07	Overcast	Cool	Normal	True	43
08	Rainy	Mild	High	False	35
09	Rainy	Cool	Normal	False	38
10	Sunny	Mild	Normal	False	46
11	Rainy	Mild	Normal	True	48
12	Overcast	Mild	High	True	52
13	Overcast	Hot	Normal	False	44
14	Sunny	Mild	High	True	30

- Firstly, we will calculate the standard deviation of Target variable and its Coef. Of Variation.

Total (N)	Mean (\bar{x})	Standard Deviation (S)	Coef. Of Variation
14	39.8	9.32	23%

$$\frac{9.32}{39.8} \times 100 = 23\%$$

$$\text{Coefficient of Variation or CV} = \frac{S}{\bar{x}} * 100$$

Requires for stopping criteria

Decision Tree Regression Working

- Now we will estimate standard deviation of predictor with respect to the target variable.
- The formula that we will use is as follows:

$$S(X, T) = \sum_{c \in X} P(c) * S(c)$$

$S(X, T)$ = Standard deviation of Predictor w.r.t. Target variable

$P(c)$ = Probability of class c

$S(c)$ = Standard deviation of class c in Predictor w.r.t. Target variable

Decision Tree Regression Working

- We will estimate standard deviation of each predictor with respect to the target variable for feature split.
- To ease the calculations, you can refer to the following table:

	Outlook	S(Hours Played)	Mean(Hours Played)	CV	Count
1	Overcast	3.49 ✓	46.25	7.54%	4
2	Rainy	7.6 ✓	35.40	21.46%	5
3	Sunny	10.87 ✓	39.20	27.72%	5



14

$$\begin{aligned} S(\text{Outlook, Hours Played}) &= P(\text{Overcast}) * S(\text{Overcast}) + P(\text{Rainy}) * S(\text{Rainy}) + P(\text{Sunny}) * S(\text{Sunny}) \\ S(\text{Outlook, Hours Played}) &= \frac{4}{14} * 3.49 + \frac{5}{14} * 7.6 + \frac{5}{14} * 10.87 \\ S(\text{Outlook, Hours Played}) &= 7.59 \end{aligned}$$

$X_1 \sim y = 7.59$

Decision Tree Regression Working

$$X_2 \sim y = 8.75$$

$$S(\text{Temperature, Hours Played}) = P(\text{Cool}) * S(\text{Cool}) + P(\text{Hot}) * S(\text{Hot}) + P(\text{Mild}) * S(\text{Mild})$$

$$S(\text{Temperature, Hours Played}) = (4 \div 14) * 10.51 + (4 \div 14) * 8.64 + (6 \div 14) * 7.65$$

$$S(\text{Temperature, Hours Played}) = 8.75$$

$$X_3 \sim y = 8.95$$

$$S(\text{Humidity, Hours Played}) = P(\text{High}) * S(\text{High}) + P(\text{Normal}) * S(\text{Normal})$$

$$S(\text{Humidity, Hours Played}) = (7 \div 14) * 9.17 + (7 \div 14) * 8.73$$


$$S(\text{Humidity, Hours Played}) = 8.95$$

$$X_4 \sim y = 8.88$$


$$S(\text{Windy, Hours Played}) = P(\text{False}) * S(\text{False}) + P(\text{True}) * S(\text{True})$$

$$S(\text{Windy, Hours Played}) = (8 \div 14) * 7.61 + (6 \div 14) * 10.59$$


$$S(\text{Windy, Hours Played}) = 8.88$$



Temperature	S(Hours Played)	Mean(Hours Played)	CV	Count
Cool	10.51	39	26.94%	4
Hot	8.64	36.5	23.67%	4
Mild	7.65	42.67	17.92%	6



Humidity	S(Hours Played)	Mean(Hours Played)	CV	Count
High	9.17	37.71	24.31%	7
Normal	8.73	42	20.78%	7



Windy	S(Hours Played)	Mean(Hours Played)	CV	Count
False	7.61	41.50	18.33%	8
True	10.59	37.67	28.11%	6

Decision Tree Regression Working

- Next, we will calculate the reduction in impurity.
- The formula that we will use is as follows:

$$\text{SDR}(X, T) = S(T) - S(X, T)$$

S(Hours Played)		
9.32		

Predictor	$S(X, T)$	$\text{SDR}(X, T)$
Outlook	7.59	$9.32 - 7.59 = 1.73$
Temperature	8.75	$9.32 - 8.75 = 0.57$
Humidity	8.95	$9.32 - 8.95 = 0.37$
Windy	8.88	$9.32 - 8.88 = 0.44$

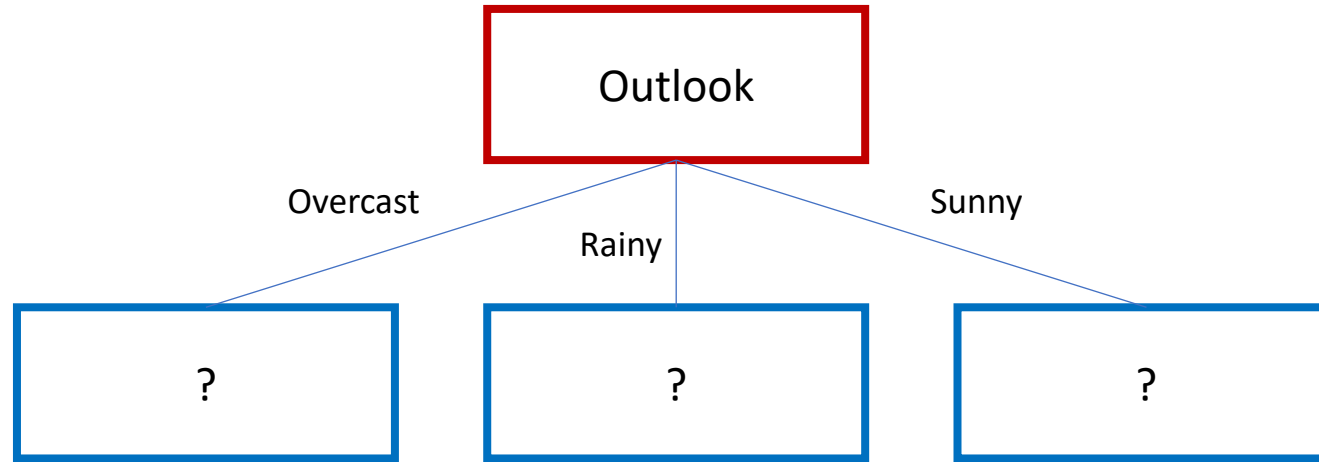
S(X, T)

Highest reduction in impurity

First Node

Decision Tree Regression Working

- Now our tree representation will look like as shown below:



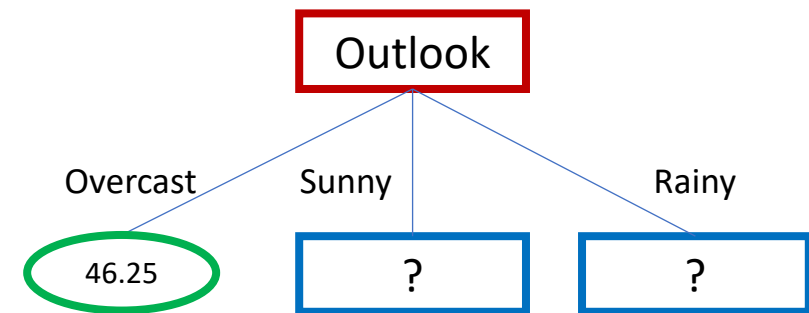
Decision Tree Regression Working

- Next, we need some stopping criteria. This is where Coef. Of Variation comes into picture.
- When CV for a branch becomes smaller than a certain threshold (e.g. 10%), the leaf node gets averaged.
- If you observe that the CV for Overcast is less than 10%, so the leaf node will be averaged.

CV Threshold = 10%

X₁

Outlook	S(Hours Played)	Mean(Hours Played)	CV	Count
Overcast	3.49	46.25	7.54%	4
Rainy	7.6	35.40	21.46%	5
Sunny	10.87	39.20	27.72%	5



- Note: We can also stop the further split if number of points remains less than a certain threshold (e.g. 3).


Decision Tree Regression Working

#	Outlook	Temperature	Humidity	Windy	Hours Played
01	Sunny	Mild	High	True	30
02	Sunny	Mild	Normal	False	46
03	Sunny	Mild	High	False	45
04	Sunny	Cool	Normal	False	52
05	Sunny	Cool	Normal	True	23
06	Rainy	Hot	High	False	26
07	Rainy	Hot	High	True	30
08	Rainy	Mild	High	False	35
09	Rainy	Cool	Normal	False	38
10	Rainy	Mild	Normal	True	48
11	Overcast	Hot	High	False	46
12	Overcast	Cool	Normal	True	43
13	Overcast	Mild	High	True	52
14	Overcast	Hot	Normal	False	44

- Now, we need to focus on the remaining data.
- We need to **identify** the **next node** for Sunny and Rainy branch.
- To do that we will **repeat** all the **steps** that we have done so far.

Decision Tree Regression Working


- We will calculate the standard deviation of Target variable and its Coef. Of Variation.
- We will be calculating the value for both sub-branches simultaneously.



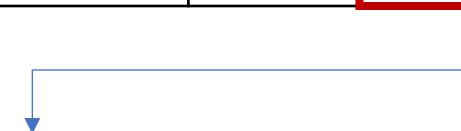
#	Outlook	Temperature	Humidity	Windy	Hours Played
01	Sunny	Mild	High	True	30
02	Sunny	Mild	Normal	False	46
03	Sunny	Mild	High	False	45
04	Sunny	Cool	Normal	False	52
05	Sunny	Cool	Normal	True	23



Total (N)	Mean (\bar{x})	Standard Deviation (S)	Coef. Of Variation
5	39.2	12.15	30.99%



#	Outlook	Temperature	Humidity	Windy	Hours Played
01	Rainy	Hot	High	False	26
02	Rainy	Hot	High	True	30
03	Rainy	Mild	High	False	35
04	Rainy	Cool	Normal	False	38
05	Rainy	Mild	Normal	True	48



Total (N)	Mean (\bar{x})	Standard Deviation (S)	Coef. Of Variation
5	35.4	8.41	23.75%

Decision Tree Regression Working

- Next, we will calculate the standard deviation with respect to the Target variable.

For Outlook = Sunny					
Predictor	Values	S(Hours Played)	Mean(Hours Played)	CV	Count
Temperature	Cool	14.5	37.5	38.67%	2
	Hot	0	0	0%	0
	Mild	7.31	40.33	18.12%	3
Humidity	High	7.5	37.5	20%	2
	Normal	12.49	40.33	30.96%	3
Windy	False	3.09	47.67	6.48%	3
	True	3.50	26.50	13.20%	2

For Outlook = Rainy					
Predictor	Values	S(Hours Played)	Mean(Hours Played)	CV	Count
Temperature	Cool	0	38	0%	1
	Hot	2	28	7.14%	2
	Mild	6.5	41.5	15.66%	2
Humidity	High	3.68	30.33	12.13%	3
	Normal	5	43	11.62%	2
Windy	False	5.09	33	15.42%	3
	True	9	39	23.07%	2

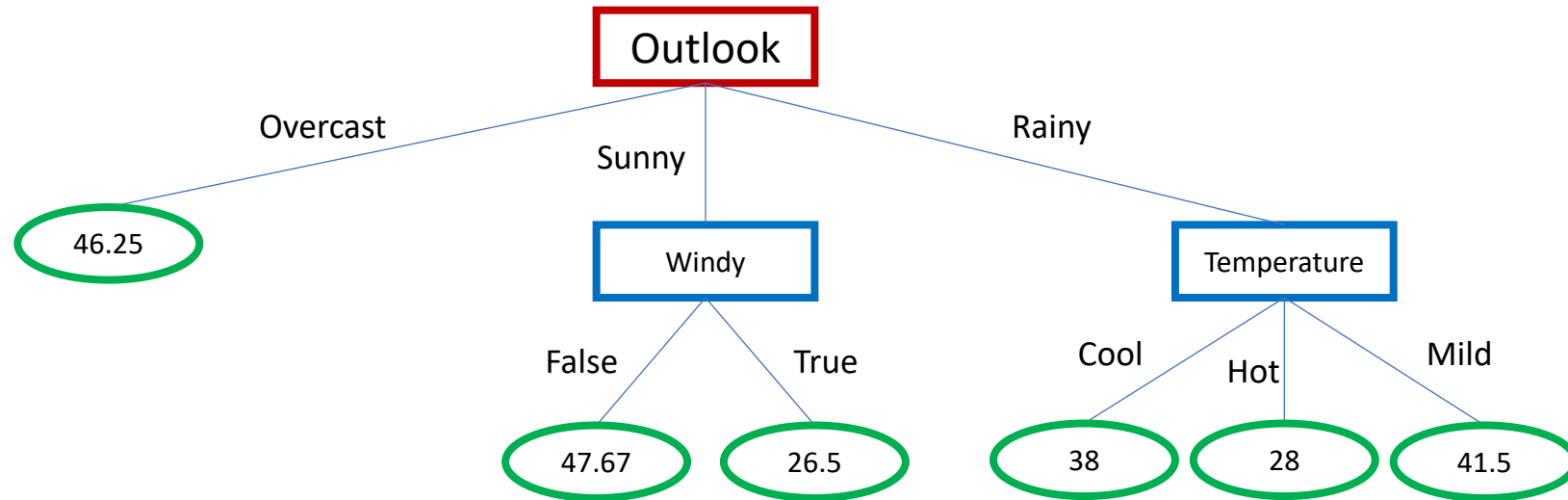
Predictor	S(X, T)	SDR(X, T)
Temperature	10.18	$12.15 - 10.18 = 1.97$
Humidity	10.49	$12.15 - 10.49 = 1.66$
Windy	3.25	$12.15 - 3.25 = 8.90$

Predictor	S(X, T)	SDR(X, T)
Temperature	3.40	$8.41 - 3.40 = 5.01$
Humidity	4.20	$8.41 - 4.20 = 4.21$
Windy	6.65	$8.41 - 6.65 = 1.76$

Next Node

Decision Tree Regression Working

- Windy variable's sub-branches are left with only two and three data points, so, we can just average out the values.
- Similar is the case with Temperature variable. We can just average out the values.

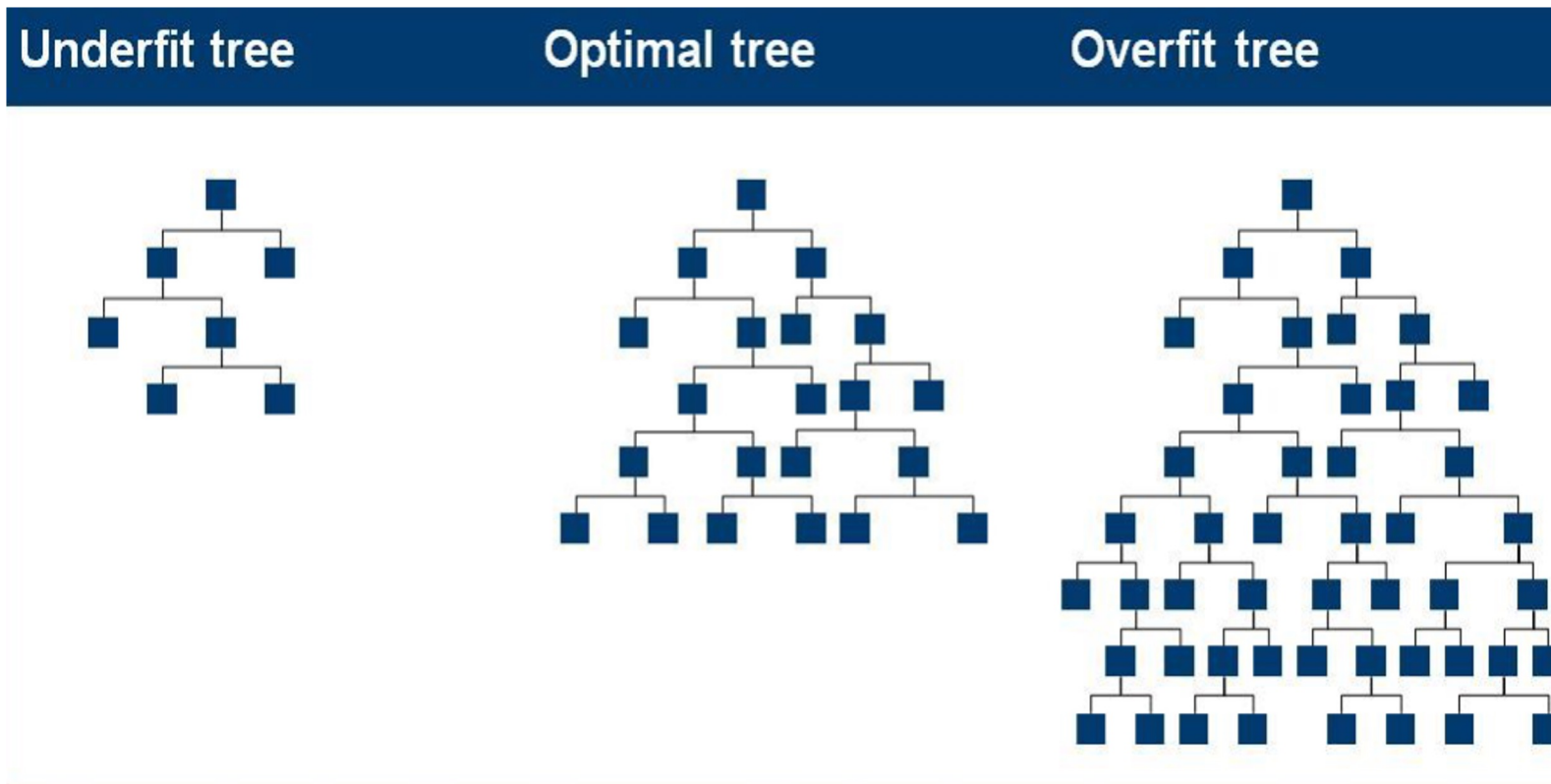


Agenda

- | | |
|--|--|
| <input type="radio"/> Terminology Related to Trees | <input type="radio"/> CART Algorithm |
| <input type="radio"/> Decision Tree | <input type="radio"/> Gini Index |
| <input type="radio"/> Decision Tree Algorithms | <input type="radio"/> Steps to estimate Gini Index |
| <input type="radio"/> Attribute Selection Measures | <input type="radio"/> CART – Regression Example |
| <input type="radio"/> ID3 Algorithm | <input checked="" type="radio"/> Issues with Decision Trees |
| <input type="radio"/> Entropy & Information Gain | <input type="radio"/> Tree Pruning |
| <input type="radio"/> Steps to Estimate Entropy & Information Gain | <input type="radio"/> Decision Tree Applications |

Issues with Decision Trees

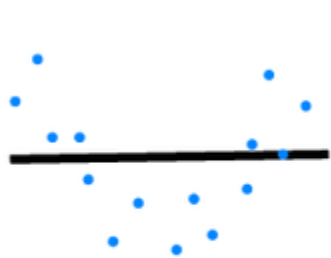
- **Underfitting**: Classifier generating too **simple trees**.
- **Overfitting**: Classifier generating too **complex trees** resulting in **poor generalization**.



Bias - Variance Trade Off

Bias

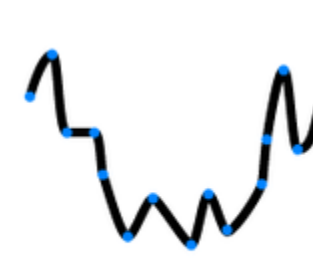
- It is also referred as **error** and it is computed as difference of predicted and actual value.
- When the bias is high and variance is low, it's called **Underfitting**.



Underfitting



Desired



Overfitting

Variance

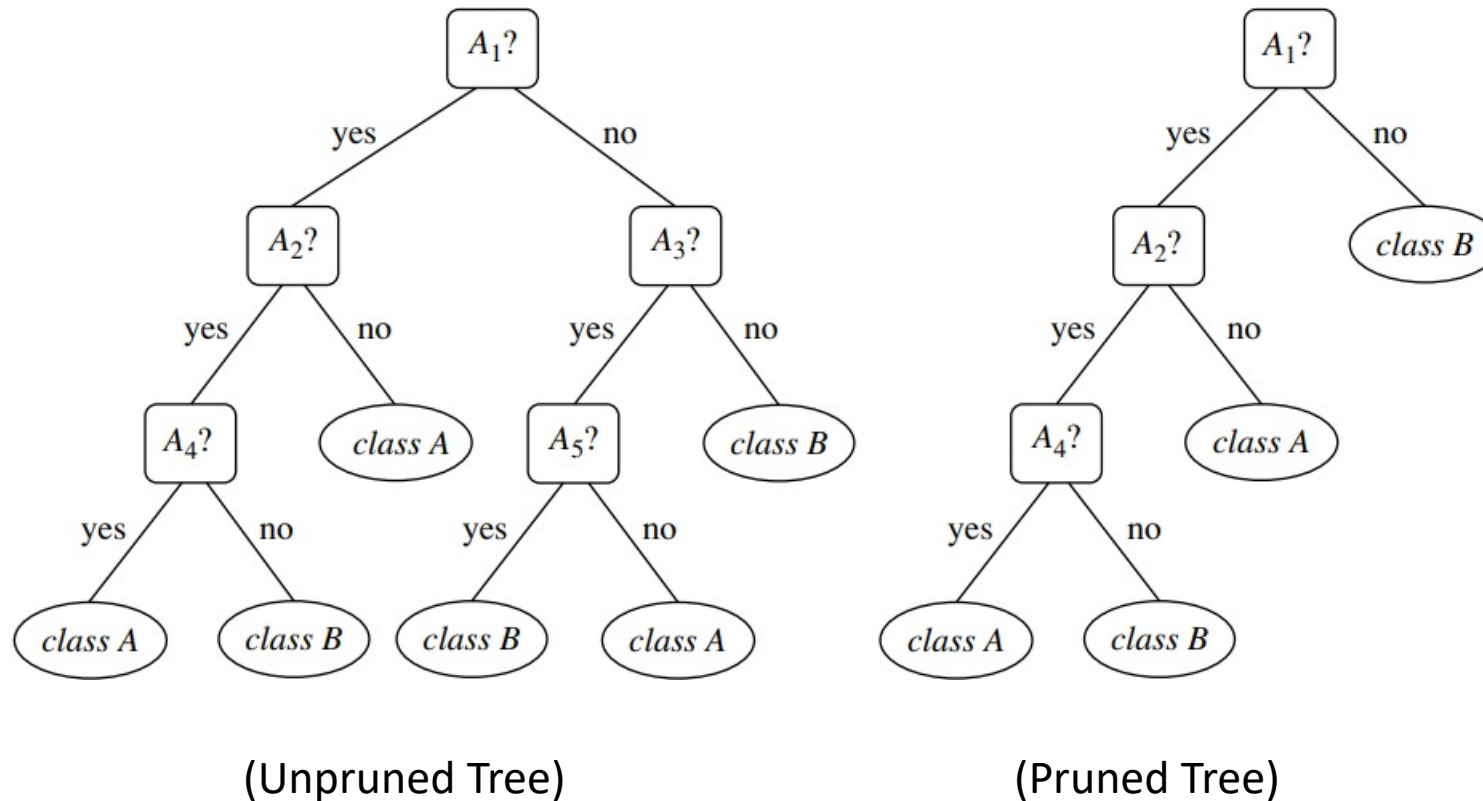
- It is the **dispersion** of the predicted values after providing input data.
- When the bias is low and variance is high, it's called **Overfitting**.

Agenda

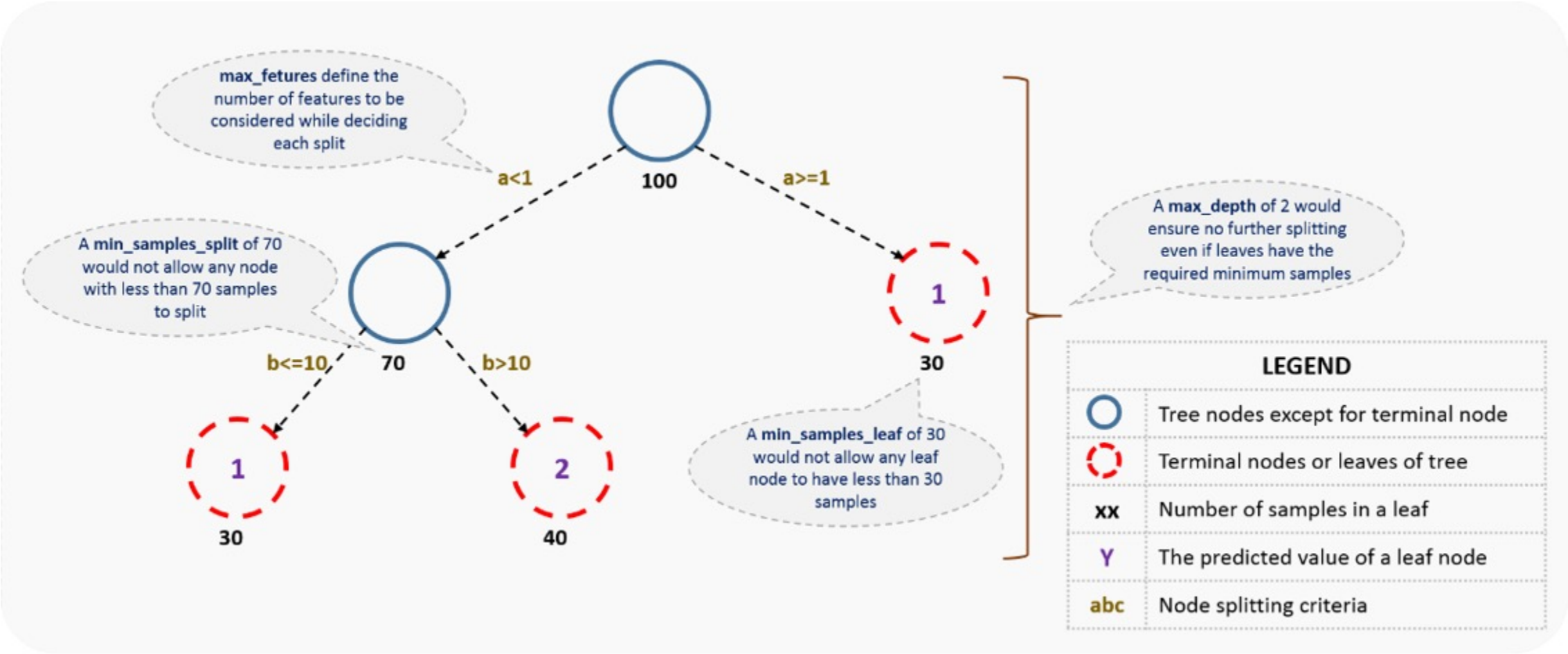
- | | |
|--|--|
| <input type="radio"/> Terminology Related to Trees | <input type="radio"/> CART Algorithm |
| <input type="radio"/> Decision Tree | <input type="radio"/> Gini Index |
| <input type="radio"/> Decision Tree Algorithms | <input type="radio"/> Steps to estimate Gini Index |
| <input type="radio"/> Attribute Selection Measures | <input type="radio"/> CART – Regression Example |
| <input type="radio"/> ID3 Algorithm | <input type="radio"/> Issues with Decision Trees |
| <input type="radio"/> Entropy & Information Gain | <input checked="" type="radio"/> Tree Pruning |
| <input type="radio"/> Steps to Estimate Entropy & Information Gain | <input type="radio"/> Decision Tree Applications |

Tree Pruning

- When a decision tree is built, many of the **branches** will **reflect anomalies in** the training **data due to noise or outliers**.
- Tree **pruning** methods **address** this **problem of overfitting** the **data**.



Solution to Over/Under-fitting: Parameters Adjustment

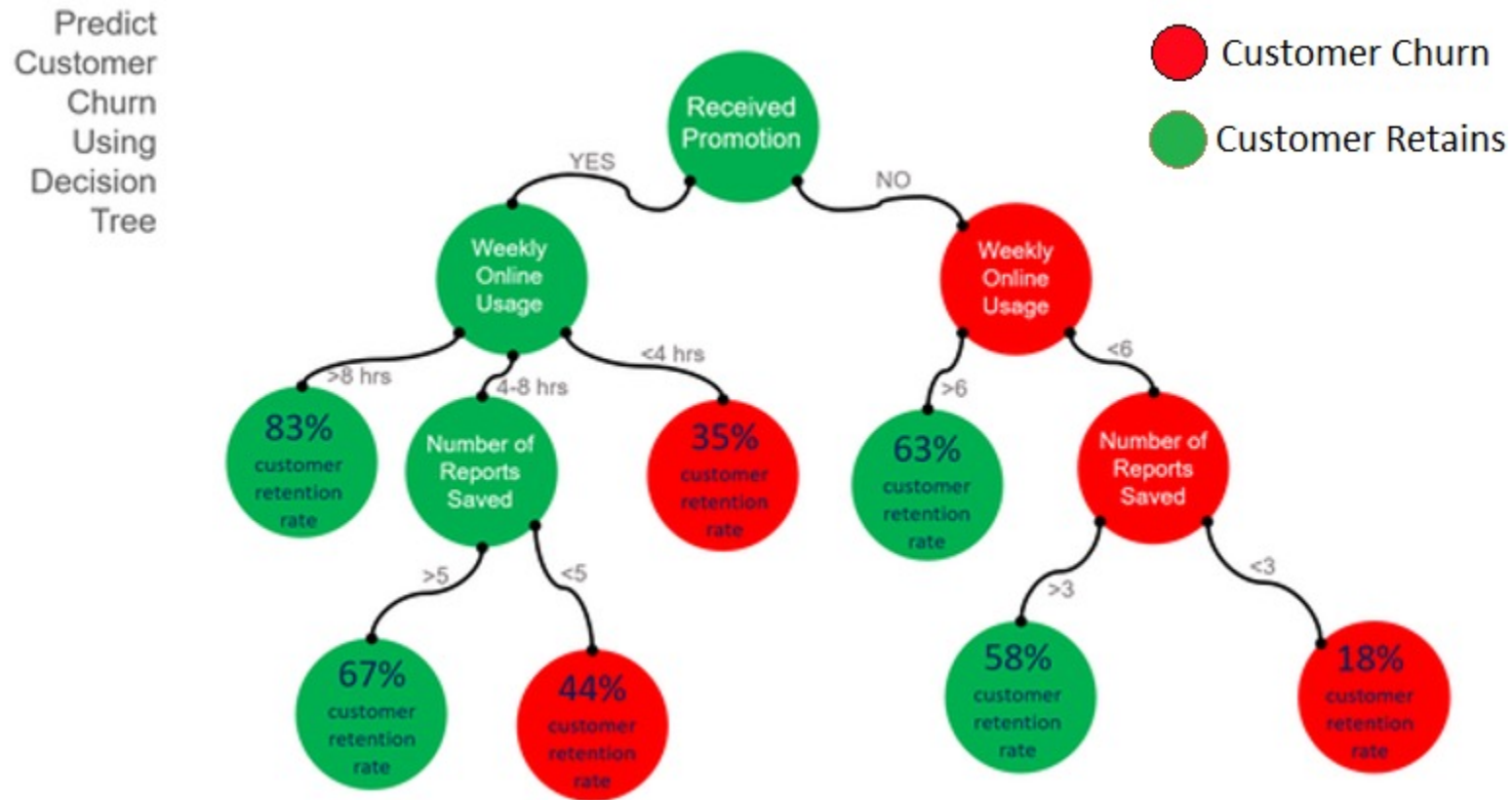


Agenda

- | | |
|--|--|
| <input type="radio"/> Terminology Related to Trees | <input type="radio"/> CART Algorithm |
| <input type="radio"/> Decision Tree | <input type="radio"/> Gini Index |
| <input type="radio"/> Decision Tree Algorithms | <input type="radio"/> Steps to estimate Gini Index |
| <input type="radio"/> Attribute Selection Measures | <input type="radio"/> CART – Regression Example |
| <input type="radio"/> ID3 Algorithm | <input type="radio"/> Issues with Decision Trees |
| <input type="radio"/> Entropy & Information Gain | <input type="radio"/> Tree Pruning |
| <input type="radio"/> Steps to Estimate Entropy & Information Gain | <input checked="" type="radio"/> Decision Tree Applications |

Decision Tree Applications

- **Customer Churn:** Identification of Customer Churn and Customer Retains



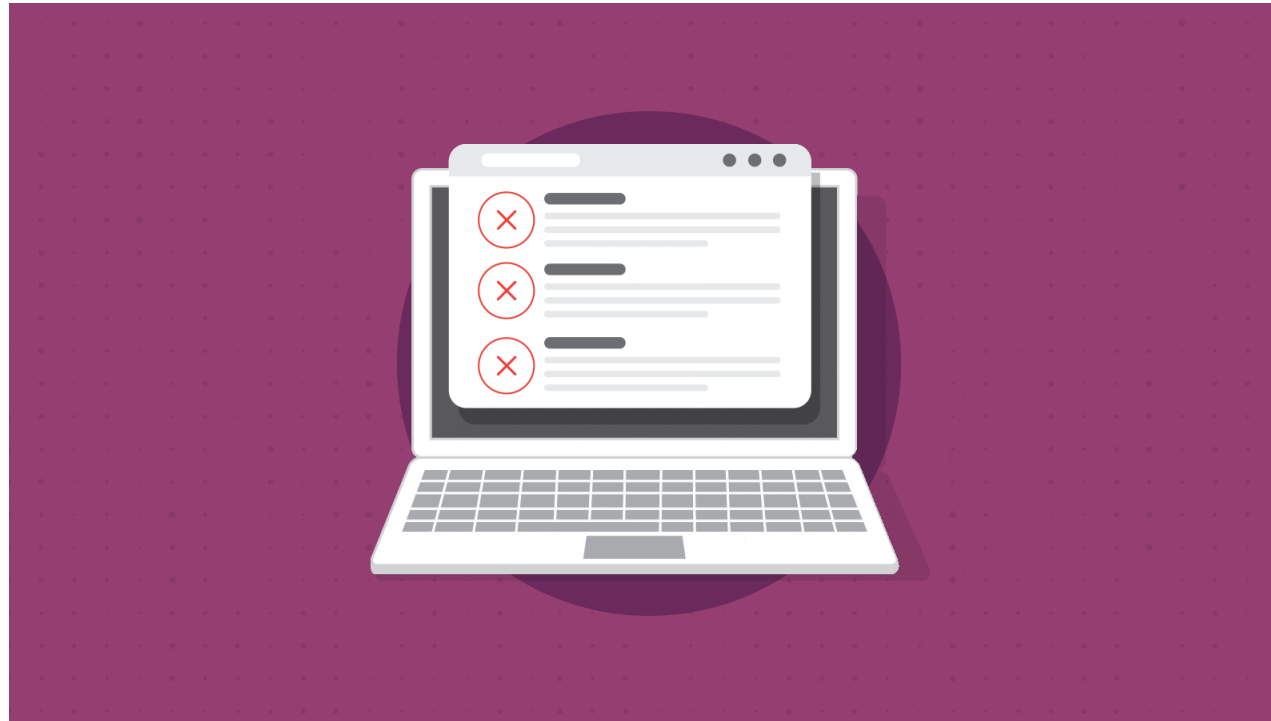
Decision Tree Applications

- **Credit-card Fraud Detection:** Classification of whether a person is fraud or not.



Decision Tree Applications

- **Manufacturing:** Chemical material evaluation for manufacturing/production.





Thanks for watching