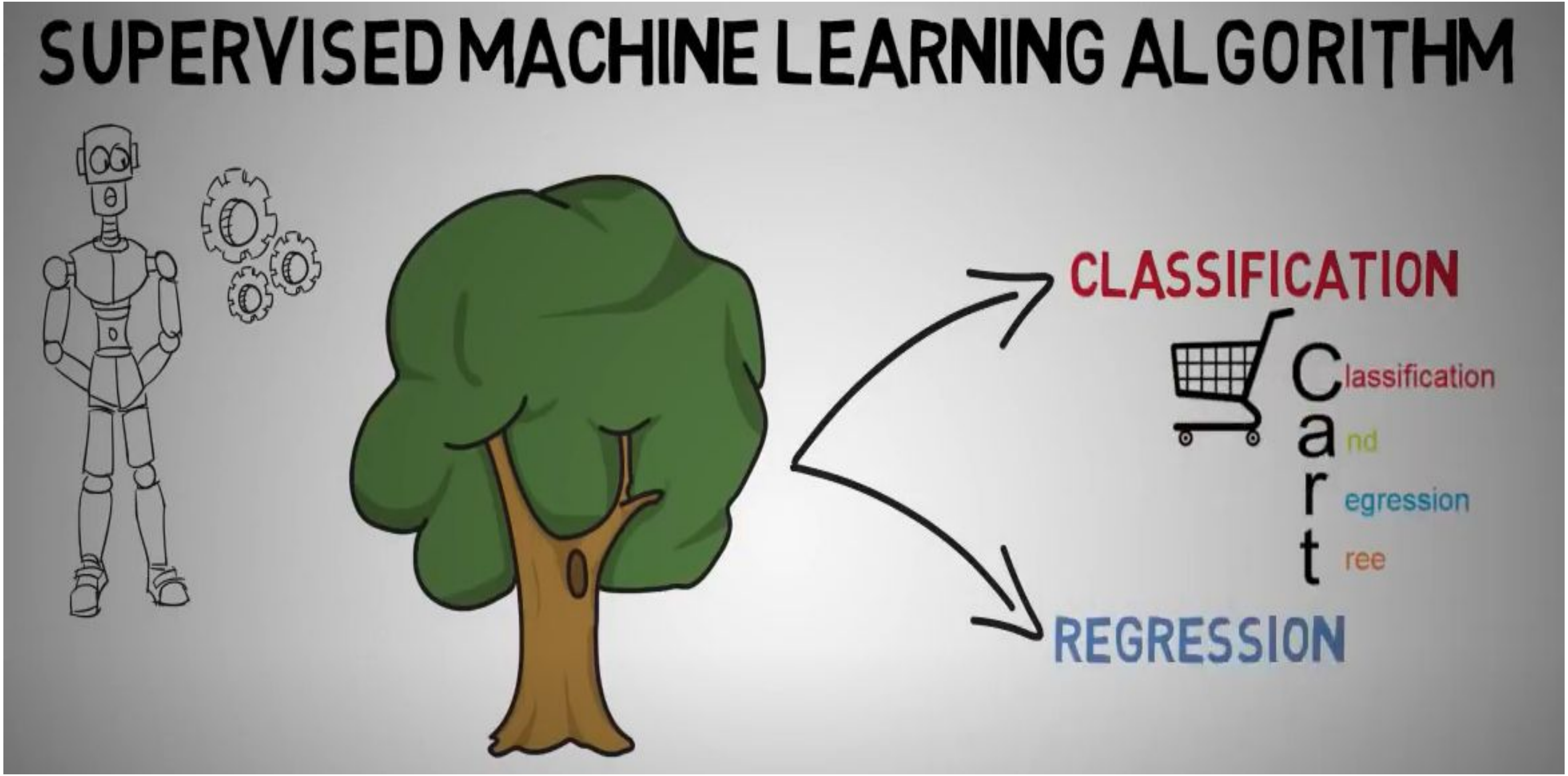

SUPERVISED MACHINE LEARNING ALGORITHM
CLASSIFICATION
Classification
and
regression
tree
REGRESSION

CART..

- **Classification Trees:** The tree is used to determine which **"class"** the target variable is most likely to fall into when it is **continuous.**
- **Regression trees:** These are used to **predict a continuous variable's value.**

CART Algorithm

- Classification and Regression Trees (CART) is a decision tree algorithm that is used for both classification and regression tasks. It is a supervised learning algorithm that learns from labelled data to predict unseen data.

- **Tree structure:** CART builds a tree-like structure consisting of nodes and branches. The nodes represent different decision points, and the branches represent the possible outcomes of those decisions. The leaf nodes in the tree contain a predicted class label or value for the target variable.

- **Splitting criteria:** CART uses a greedy approach to split the data at each node. It evaluates all possible splits and selects the one that best reduces the impurity of the resulting subsets. For **classification tasks,** CART uses **Gini impurity** as the splitting criterion. The lower the Gini impurity, the more pure the subset is. For **regression tasks**, CART uses **residual reduction** as the splitting criterion. The lower the residual reduction, the better the fit of the

- **Pruning:** To prevent overfitting of the data, pruning is a technique used to **remove the nodes that contribute little to the model accuracy**. Cost complexity pruning and information gain pruning are two popular pruning techniques.
- **Cost complexity pruning** involves calculating the cost of each node and removing nodes that have a negative cost.
- **Information gain pruning** involves calculating the information gain of each node and removing nodes that have a low information gain.

- **How does CART algorithm works?**

- The CART algorithm works via the following process:
  - The best-split point of each input is obtained.
  - Based on the best-split points of each input in Step 1, the new "best" split point is identified.
  - Split the chosen input according to the "best" split point.
  - Continue splitting until a stopping rule is satisfied or no further desirable splitting is available.

- CART algorithm uses Gini Impurity to split the dataset into a decision tree .It does that by searching for the best Homogeneity for the sub nodes, with the help of the Gini index criterion.
- the Gini index is a metric for the classification tasks in CART. It stores the sum of squared probabilities of each class.

Mathematically, we can write Gini Impurity as follows:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

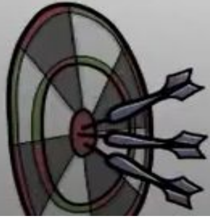
where p_i is the probability of an object being classified to a particular class.

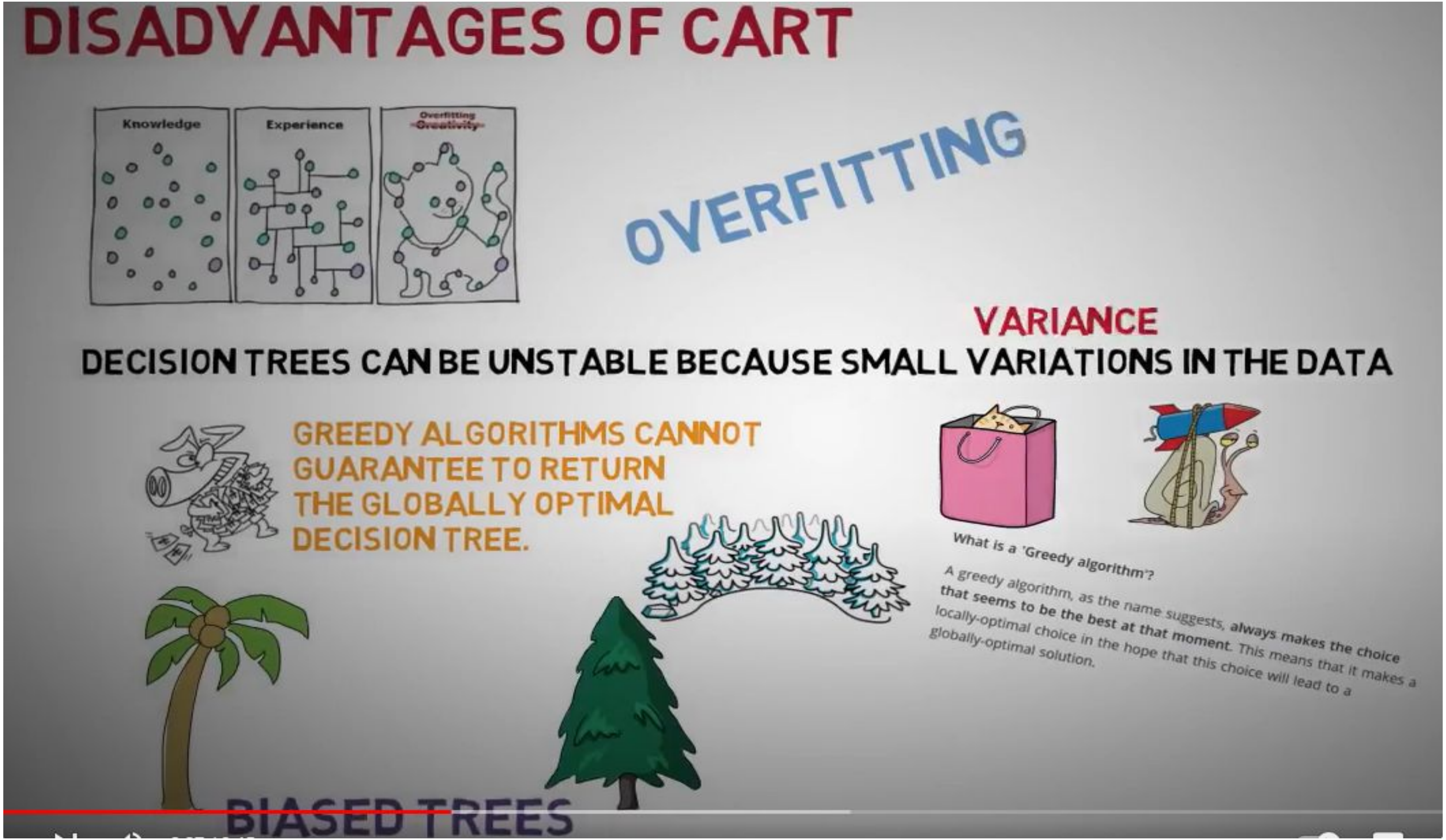
ADVANTAGES OF CART

SIMPLE TO UNDERSTAND, INTERPRET, VISUALIZE.

VARIABLE SCREENING OR FEATURE SELECTION

LITTLE EFFORT FOR DATA PREPERATION




DISADVANTAGES OF CART
Knowledge
Experience
Overfitting
~~Creativity~~
OVERFITTING
VARIANCE
DECISION TREES CAN BE UNSTABLE BECAUSE SMALL VARIATIONS IN THE DATA
GREEDY ALGORITHMS CANNOT GUARANTEE TO RETURN THE GLOBALLY OPTIMAL DECISION TREE.
What is a 'Greedy algorithm'?
A greedy algorithm, as the name suggests, **always makes the choice that seems to be the best at that moment.** This means that it makes a locally-optimal choice in the hope that this choice will lead to a globally-optimal solution.
BIASED TREES

CLASSIFICATION AND REGRESSION TREES

REGRESSION TREES ARE USED WHEN DEPENDENT VARIABLE IS CONTINUOUS.
CLASSIFICATION TREES ARE USED WHEN DEPENDENT VARIABLE IS CATEGORICAL.

REGRESSION - USE MEAN/AVERAGE

Mean Average

CLASSIFICATION - USE MODE / CLASS

Mode = happens most often

THE SPLITTING PROCESS RESULTS IN FULLY GROWN TREES UNTIL THE STOPPING CRITERIA IS REACHED.
BUT, THE FULLY GROWN TREE IS LIKELY TO OVERFIT DATA, LEADING TO POOR ACCURACY ON UNSEEN DATA.

PRUNING

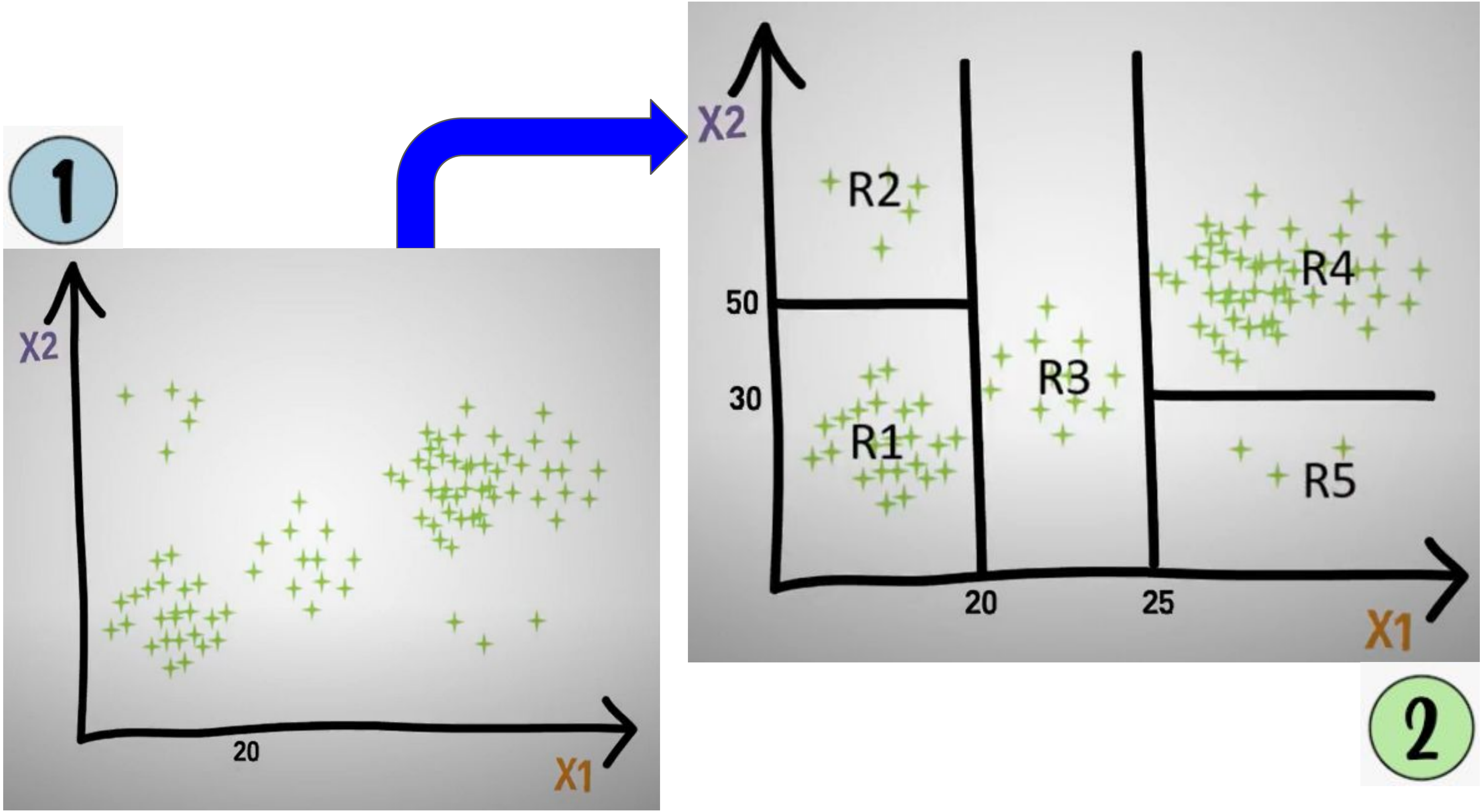
GROWING A TREE

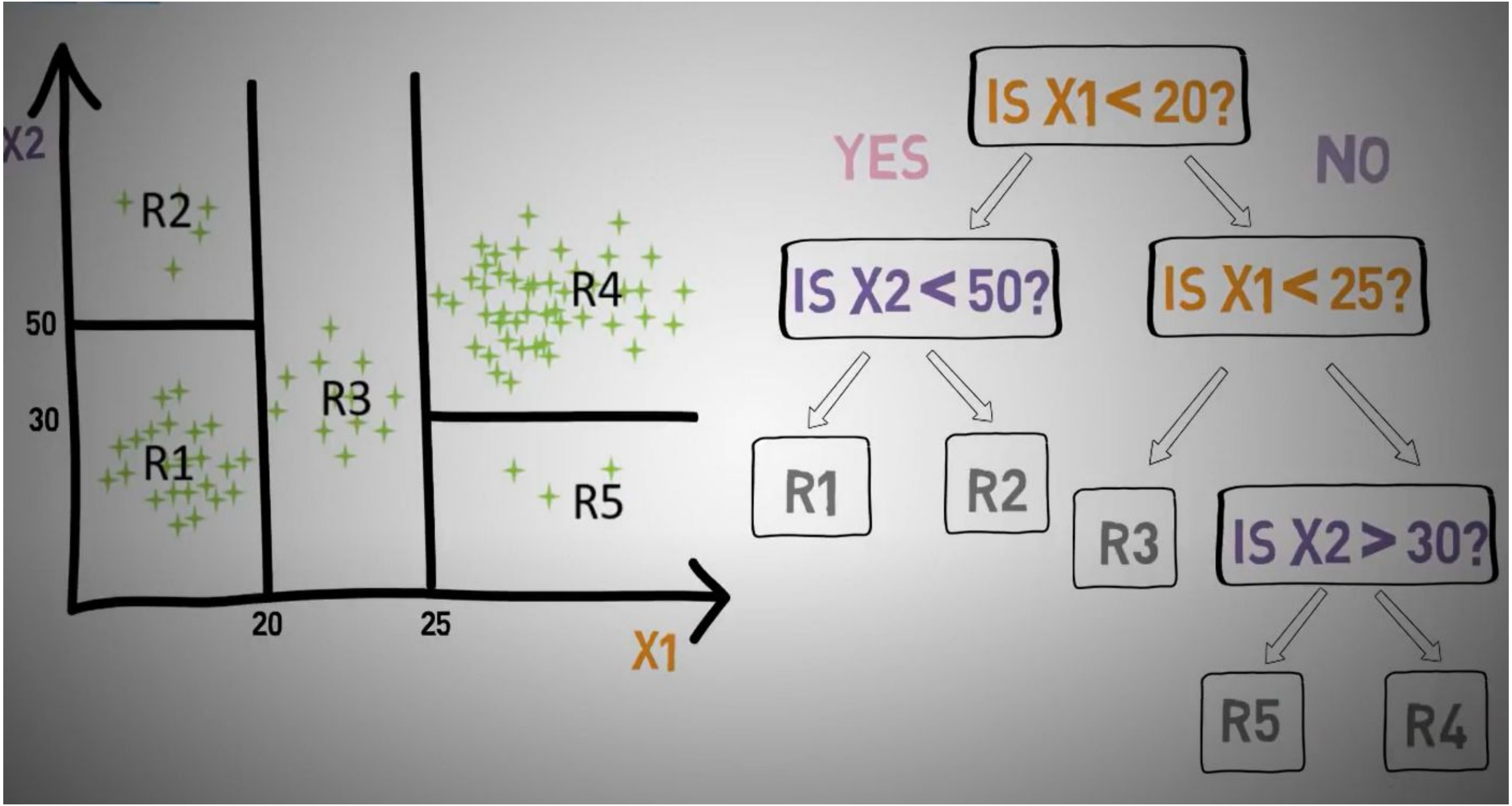
1) FEATURES TO CHOOSE

2) CONDITIONS FOR SPLITTING

3) KNOWING WHEN TO STOP

4) PRUNING


1
X2
20
X1
X2
R2
50
30
R1
R3
R4
R5
20
25
X1
2


X2
R2
50
R4
R3
30
R1
R5
20
25
X1
IS X1 < 20?
YES
NO
IS X2 < 50?
IS X1 < 25?
R1
R2
R3
IS X2 > 30?
R5
R4

Build Decision Tree using Gini Index Solved Numerical Example Machine Learni...

Decision Tree using Gini Index

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Solved

Numerical

Example

Machine Learning

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Compute the Gini Index for the overall collection of training examples.
- There are four possible output variables Cinema, Tennis, Stay In and Shopping.
- The data has 6 instances of Cinema, 2 instances of Tennis, 1 instance of Stay In and 1 of shopping.
- $Gini(S) = 1 - [\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2] = 0.58$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of Gini Index for Money Attribute
- It has two possible values of Rich (7 examples) and Poor (3 examples).
- For **Money = Poor,** there are 3 examples with "Cinema".
- $Gini(S) = 1 - [(\frac{3}{3})^2] = 0$ ✓

7

- For **Money = Rich,** there are 2 examples with "Tennis", 3 examples with "Cinema" and 1 example with "Stay in", "Shopping" each
- $Gini(S) = 1 - [(\frac{2}{7})^2 + (\frac{3}{7})^2 + (\frac{1}{7})^2 + (\frac{1}{7})^2] =$ 0.694
- $\boldsymbol{Weighted\ Average(Money)}$

$$= 0 * \left(\frac{3}{10}\right) + 0.694 * \left(\frac{7}{10}\right) = 0.486$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of **Gini Index for Parents** Attribute
- It has two possible values of **Yes (5 examples)** and **No (5 examples).**
- For **Parents = Yes,** there are **5 examples**, all with "Cinema".
- $Gini(S) = 1 - [(\frac{5}{5})^2] = 0$
- For **Parents = No,** there are **2 examples with "Tennis", 1 example with "Stay in", "Shopping" and "Cinema" each**
- $Gini(S) = 1 - [(\frac{2}{5})^2 + (\frac{1}{5})^2 + (\frac{1}{5})^2 + (\frac{1}{5})^2] =$ 0.72
- $\boldsymbol{Weighted\ Average(Parents)}$

$$= 0 * \left(\frac{5}{10}\right) + [0.72 * \left(\frac{5}{10}\right) = 0.36$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

- Computation of **Gini Index for Weather** Attribute
- It has three possible values of **Sunny (3 examples)**, **Rainy (3 examples)** and **Windy (4 examples).**
- For **Weather = Sunny**, there are **2 examples with "Cinema"** and **1 with "Tennis".**
- $Gini(Sunny) = 1 - [(\frac{2}{3})^2 + (\frac{1}{3})^2] = 0.444$
- For **Weather = Rainy**, there are **2 examples with "Cinema"** and **1 example with "Stay in"**
- $Gini(Rainy) = 1 - [(\frac{2}{3})^2 + (\frac{1}{3})^2] = 0.444$
- For **Weather = Windy**, there are **3 examples with "Cinema"** and **1 example with "Shopping"**
- $Gini(Windy) = 1 - [(\frac{3}{4})^2 + (\frac{1}{4})^2] = 0.375$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$Weighted\ Average(Weather)$$

$$= 0.444 * \left(\frac{3}{10}\right) + 0.444 * \left(\frac{3}{10}\right) + 0.375 * \left(\frac{4}{10}\right)$$

$$= 0.416$$

For Weather - Gini Index: 0.416

For Parents - Gini Index: 0.36

For Money - Gini Index: 0.486

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

$$Weighted\ Average(Weather)$$

$$= 0.444 * \left(\frac{3}{10}\right) + 0.444 * \left(\frac{3}{10}\right) + 0.375 * \left(\frac{4}{10}\right)$$

$$= 0.416$$

For Weather - Gini Index: 0.416

For Parents - Gini Index: 0.36

For Money - Gini Index: 0.486

Parents is selected as it has smallest Gini index.

Parents

Cinema Yes

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W6	Rainy	Yes	Poor	Cinema
W9	Windy	Yes	Rich	Cinema

No

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute

- **Sunny (2 examples)**
- For Parent= No | Weather = Sunny, there are 2 example with "Tennis.
- $Gini(S) = 1 - [\left(\frac{2}{2}\right)^2] = 0$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No | Weather Attribute

- **Windy (2 example)**
- For Parents = No | Weather = Windy, there is 1 example with "Cinema" and 1 example with "Shopping".
- $Gini(S) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0.5$

$$Weighted\ Average(Parents = No \mid Weather) = 0 * \left(\frac{2}{5}\right) + 0 * \left(\frac{1}{5}\right) + 0.5 * \left(\frac{2}{5}\right) = 0.2$$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Computation of Gini Index for Parents = No|Money Attribute

- Rich (4 examples)
- For Parents = No| Money = Rich, there is 1 example with "stay in" and "Shopping" each and 2 examples of "Tennis".
- $Gini(S) = 1 - [\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{2}{4}\right)^2] = 0.625$

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

For Parents = No | Weather - Gini Index: 0.2

For Parents = No | Money - Gini Index: 0.5

Weather is selected as it has smallest Gini index.

Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W5	Rainy	No	Rich	Stay In
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W10	Sunny	No	Rich	Tennis

Now, for Parent=No & Weather=Sunny, we have all instances as Tennis.

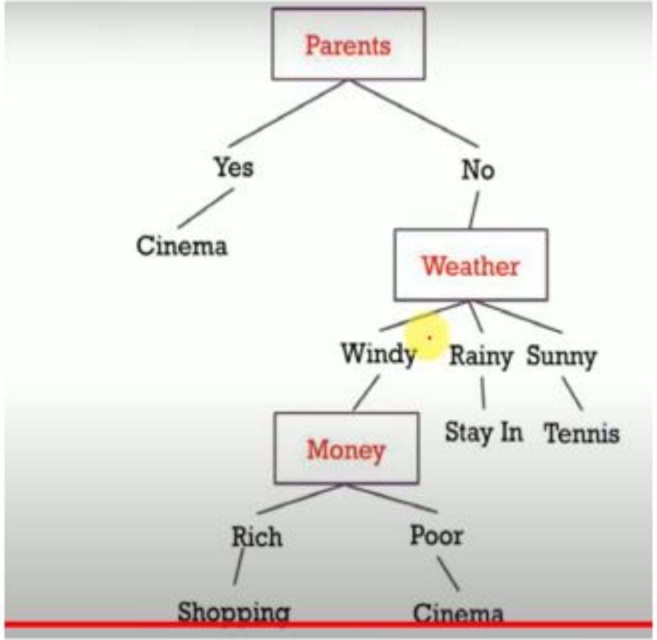
Weekend	Weather	Parents	Money	Decision
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Now, for Parents=No & Weather=Rainy, we have all instances as Stay In.

Weekend	Weather	Parents	Money	Decision
W5	Rainy	No	Rich	Stay In

Now, for Parent=No & Weather=Windy, we need to split.

Weekend	Weather	Parents	Money	Decision
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping


Parents
Yes
No
Cinema
Weather
Windy
Rainy
Sunny
Money
Stay In
Tennis
Rich
Poor
Shopping
Cinema

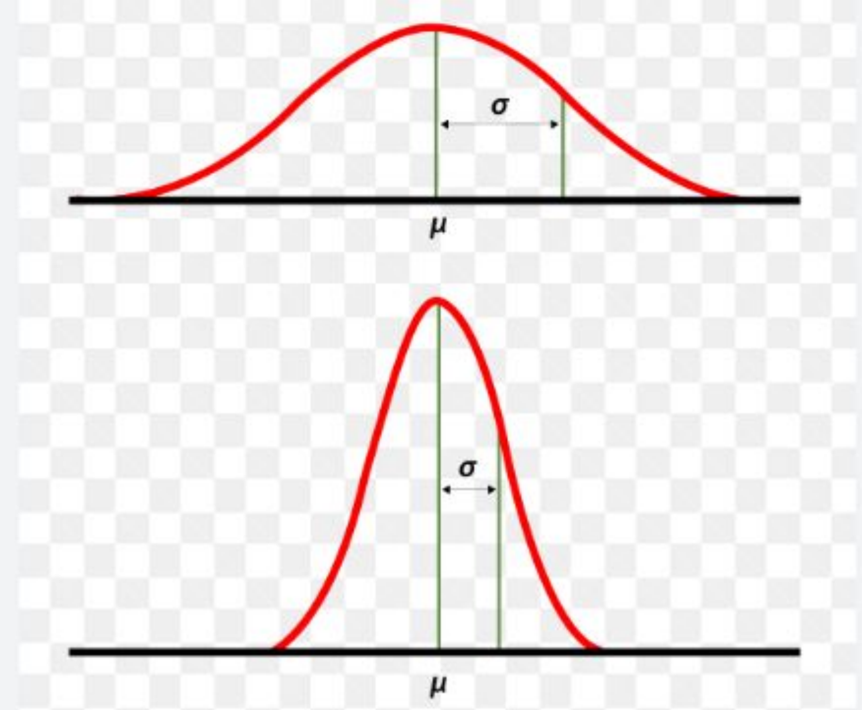
Regression Trees

Day	Outlook	Temp	Humidity	Wind	Golf Players
D1	Sunny	Hot	High	Weak	25
D2	Sunny	Hot	High	Strong	30
D3	Overcast	Hot	High	Weak	46
D4	Rain	Mild	High	Weak	45
D5	Rain	Cool	Normal	Weak	52
D6	Rain	Cool	Normal	Strong	23
D7	Overcast	Cool	Normal	Strong	43
D8	Sunny	Mild	High	Weak	35
D9	Sunny	Cool	Normal	Weak	38
D10	Rain	Mild	Normal	Weak	46
D11	Sunny	Mild	Normal	Strong	48
D12	Overcast	Mild	High	Strong	52
D13	Overcast	Hot	Normal	Weak	44
D14	Rain	Mild	High	Strong	30

Standard Deviation is used **to calculate the** ***homogeneity of a numerical sample***

How does standard deviation help data?

Standard deviation describes how dispersed a set of data is. It compares each data point to the mean of all data points, and standard deviation returns a calculated value **that describes whether the data points are in close proximity or whether they are spread out.**



- **Standard deviation**
- $Average\ of\ golf\ players\ =$

$$\frac{25+30+46+45+52+23+43+35+38+46+48+52+44+3(}{14}$$

- $Average\ of\ golf\ players\ =\ 39.78$
- $Standard\ deviation\ of\ golf\ players$

- $$sd=\sqrt{\frac{(25-39.78)^2+(30-39.78)^2+\cdots+(30-39.78)^2}{14}}$$

- $Standard\ deviation\ of\ golf\ players\ =\ \mathbf{9.32}$

Day	Outlook	Temp	Humidity	Wind	Golf Players
D1	Sunny	Hot	High	Weak	25 ✓
D2	Sunny	Hot	High	Strong	30
D8	Sunny	Mild	High	Weak	35
D9	Sunny	Cool	Normal	Weak	38
D11	Sunny	Mild	Normal	Strong	48

- **Outlook**
- Outlook →{sunny, overcast, rain}
- Calculate standard deviation of golf players for all of these outlook candidates.

- **Sunny outlook**
- $Average\ of\ golf\ players\ for\ sunny\ outlook = \frac{25+30+35+38+48}{5} = 35.2$
- $SD\ of\ golf\ players\ for\ sunny\ outlook = \sqrt{\frac{(25-35.2)^2+(30-35.2)^2+(35-35.2)^2+(38-35.2)^2+(48-35.2)^2}{5}}$
- $\boldsymbol{SD\ of\ golf\ players\ for\ sunny\ outlook = 7.78}$

Day	Outlook	Temp	Humidity	Wind	Golf Players
D3	Overcast	Hot	High	Weak	46
D7	Overcast	Cool	Normal	Strong	43
D12	Overcast	Mild	High	Strong	52
D13	Overcast	Hot	Normal	Weak	44

- **Outlook**
- Outlook →{sunny, overcast, rain}
- Calculate standard deviation of golf players for all of these outlook candidates.

- **Sunny Overcast**
- $Average\ of\ golf\ players\ for\ Overcast\ outlook = \frac{46+43+52+44}{4} = 46.25$
- $SD\ of\ golf\ players\ for\ Overcast\ outlook = \sqrt{\frac{(46-46.25)^2+(43-46.25)^2+(52-46.25)^2+(44-46.25)^2}{4}}$
- $\boldsymbol{SD\ of\ golf\ players\ for\ sunny\ Overcast = 3.49}$

Day	Outlook	Temp	Humidity	Wind	Golf Players
D4	Rain	Mild	High	Weak	45
D5	Rain	Cool	Normal	Weak	52
D6	Rain	Cool	Normal	Strong	23
D10	Rain	Mild	Normal	Weak	46
D14	Rain	Mild	High	Strong	30

- **Outlook**
- Outlook →{sunny, overcast, rain}
- Calculate standard deviation of golf players for all of these outlook candidates.

- **Sunny Rain**
- $Average\ of\ golf\ players\ for\ Rain\ outlook = \frac{45+52+23+46+30}{5} = 39.2$
- $SD\ of\ golf\ players\ for\ Rain\ outlook = \sqrt{\frac{(45-39.2)^2+(52-39.2)^2+(23-39.2)^2+(46-39.2)^2+(30-39.2)^2}{5}}$
- $\boldsymbol{SD\ of\ golf\ players\ for\ sunny\ Rain = 10.87}$

Outlook	Stdev of Golf Players	Instances
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5

- $Weighted\ standard\ deviation\ for\ outlook = \left(\frac{4}{14}\right) * 3.49 + \left(\frac{5}{14}\right) * 10.87 + \left(\frac{5}{14}\right) * 7.78 = 7.66$
- Global standard deviation of golf players 9.32
- $Standard\ deviation\ reduction\ for\ outlook\ = \ 9.32 - 7.66\ = \ 1.66$

Temperature

Temperature can be hot, cool or mild. We will calculate standard deviations for those candidates.

Hot temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
13	Overcast	Hot	Normal	Weak	44

Golf players for hot temperature = {25, 30, 46, 44}

Standard deviation of golf players for hot temperature = 8.95

Cool temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38

Golf players for cool temperature = {52, 23, 43, 38}

Standard deviation of golf players for cool temperature = 10.51

Mild temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Golf players for mild temperature = {45, 35, 46, 48, 52, 30}

Standard Deviation of the golf players for mild temperature=7.65

Temp	Stdev of Golf Players	Instances
Hot	8.95	4
Mild	10.51	4
Cool	7.65	6

- Weighted Standard deviation for temp=
  (4/14)*8.95+(4/14)*10.51+(6/14)*7.65 =8.84
- Global standard Deviation of golf players is 9.32
- Standard Deviation reduction for Temp=
  9.32-8.84=0.47

Humidity

Humidity is a binary class. It can either be normal or high.

High humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
8	Sunny	Mild	High	Weak	35
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Standard Deviation for golf players for high humidity=9.36

Normal humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
13	Overcast	Hot	Normal	Weak	44

Golf players for normal humidity = {52, 23, 43, 38, 46, 48, 44}

Standard Deviation for golf players for normal humidity=8.73

Humidity	Stdev of Golf Players	Instances
High	9.36	7
Normal	8.73	7

- Weighted Standard Deviation for Humidity=
  (7/14)*9.36+(7/14)*8.73=9.04
- Global standard Deviation of golf players is 9.32
- Standard Deviation reduction for Humidity=
  9.32-9.04=0.27

Strong Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
2	Sunny	Hot	High	Strong	30
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
14	Rain	Mild	High	Strong	30

Standard Deviation for golf players for “Strong

Weak Wind

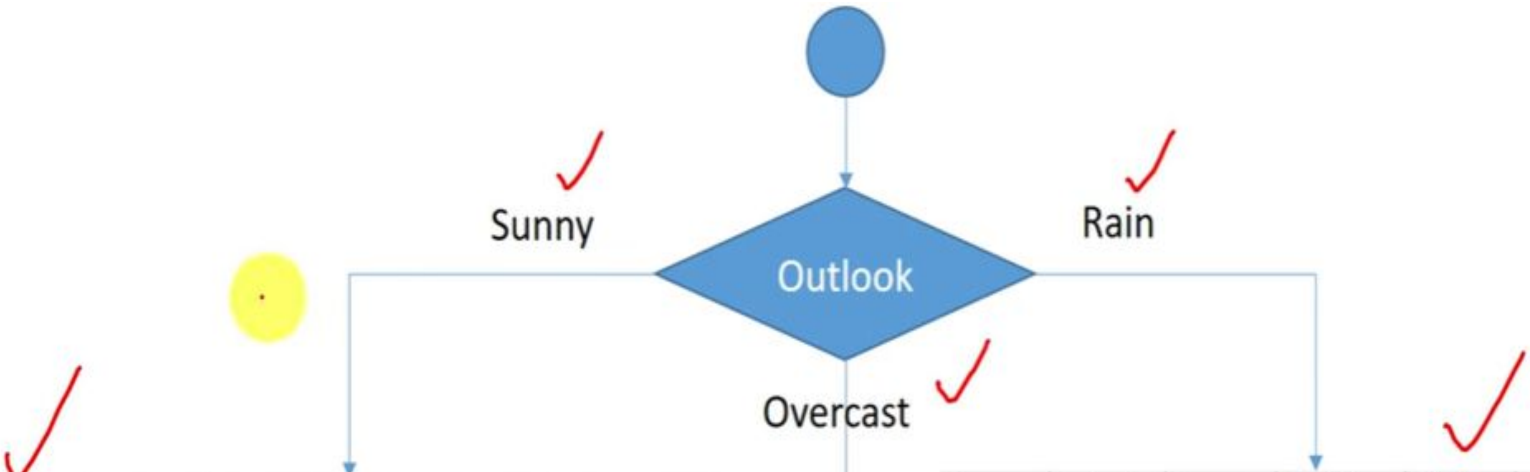
1	Sunny	Hot	High	Weak	25
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
13	Overcast	Hot	Normal	Weak	44

Standard Deviation for golf players for "Weak Wind"=7.87

Wind	Stdev of Golf Players	Instances
Strong	10.59 ✓	6
Weak	7.87	8

- Weighted Standard Deviation for Wind=
  (6/14)*10.59+(8/14)*7.87=9.03
- Global Standard Deviation of golf players=9.36
- Standard Deviation reduction wind=
  9.32-9.03=0.29

Feature	Standard Deviation Reduction
Outlook	1.66
Temperature	0.47
Humidity	0.27
Wind	0.29



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Sunny Outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Standard Deviation for “Sunny Outlook”=7.78

Sunny outlook and Hot Temperature

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

Standard deviation for sunny outlook and hot temperature = 2.5

Sunny outlook and Mild Temperature

Day	Outlook	Temp. ✓	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

Standard deviation for sunny outlook and mild temperature = 6.5

Temperature	Stdev for Golf Players	Instances
Hot	2.5	2
Cool	0	1
Mild	6.5	2

Weighted standard deviation for sunny outlook and temperature = (2/5)x2.5 + (1/5)x0 + (2/5)x6.5 = 3.6

Standard deviation reduction for sunny outlook and temperature = 7.78 – 3.6 = 4.18

Sunny outlook and high humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35

Standard deviation for sunny outlook and high humidity = 4.08

Sunny outlook and normal humidity

Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Standard deviation for sunny outlook and normal humidity = 5

Humidity	Stdev for Golf Players	Instances
High	4.08	3
Normal	5.00	2

Weighted standard deviations for sunny outlook and humidity = (3/5)x4.08 + (2/5)x5 = 4.45

Standard deviation reduction for sunny outlook and humidity = 7.78 – 4.45 = 3.33

Sunny Outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Golf players for sunny outlook = {25, 30, 35, 38, 48}

Standard Deviation for sunny outlook=7.79

Sunny outlook and Weak Wind

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38

Standard deviation for sunny outlook and weak wind = 5.56

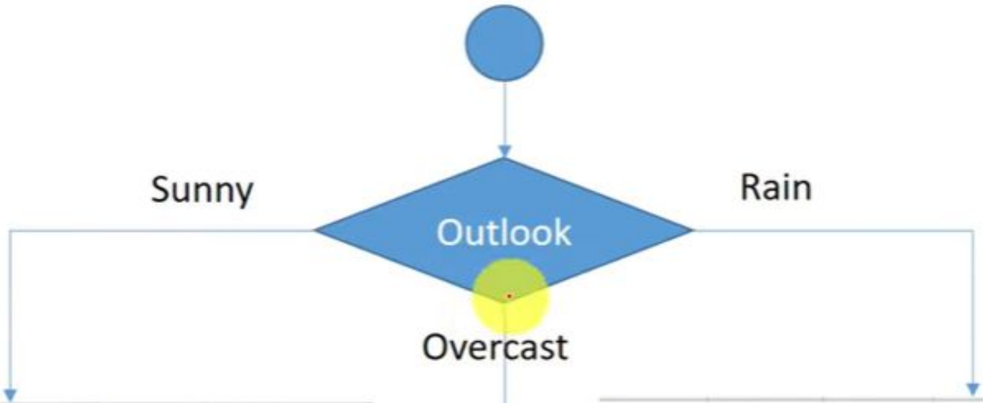
Standard deviation for sunny outlook and weak wind = 5.56

Wind	Stdev for Golf Players	Instances
Strong	9	2
Weak	5.56	3

Weighted standard deviations for sunny outlook and wind = (2/5)x9 + (3/5)x5.56 = 6.93

Standard deviation reduction for sunny outlook and wind = 7.78 – 6.93 = 0.85

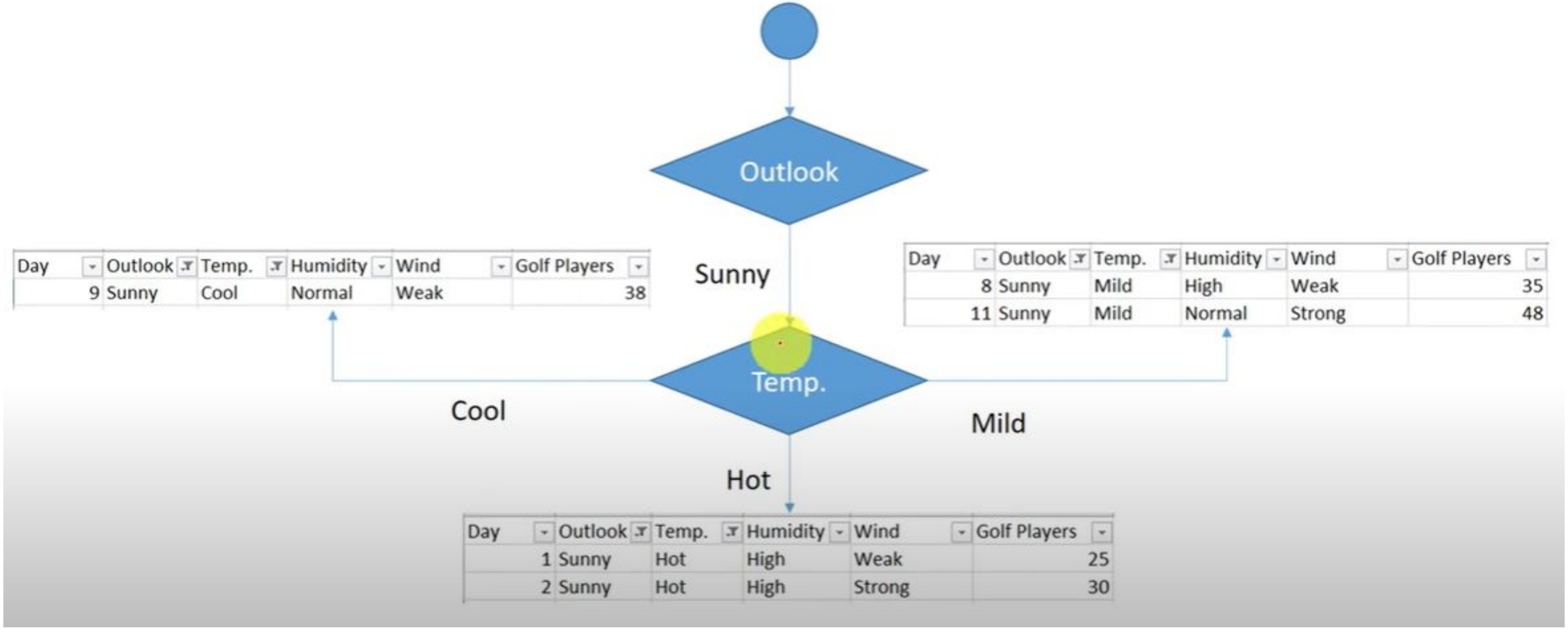
Feature	Standard Deviation Reduction
Temperature	4.18
Humidity	3.33
Wind	0.85



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44



Day	Outlook	Temp.	Humidity	Wind	Golf Players
9	Sunny	Cool	Normal	Weak	38

Day	Outlook	Temp.	Humidity	Wind	Golf Players
8	Sunny	Mild	High	Weak	35
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30

Rainy Outlook

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

We need to find standard deviation reduction values for the rest of the features in same way for the sub data set above.

Standard Deviation for “rainy outlook”=10.87

Rainy outlook and temperature

Temperature	Standard deviation for golf players	instances
Cool	14.50	2
Mild	7.32	3

Weighted standard deviation for rainy outlook and temperature = (2/5)x14.50 + (3/5)x7.32 = 10.19

Standard deviation reduction for rainy outlook and temperature = 10.87 – 10.19 = 0.67

Rainy outlook and humidity

Humidity	Standard deviation for golf players	instances
High	7.50	2
Normal	12.50	3

Weighted standard deviation for rainy outlook and humidity = (2/5)x7.50 + (3/5)x12.50 = 10.50

Standard deviation reduction for rainy outlook and humidity = 10.87 – 10.50 = 0.37

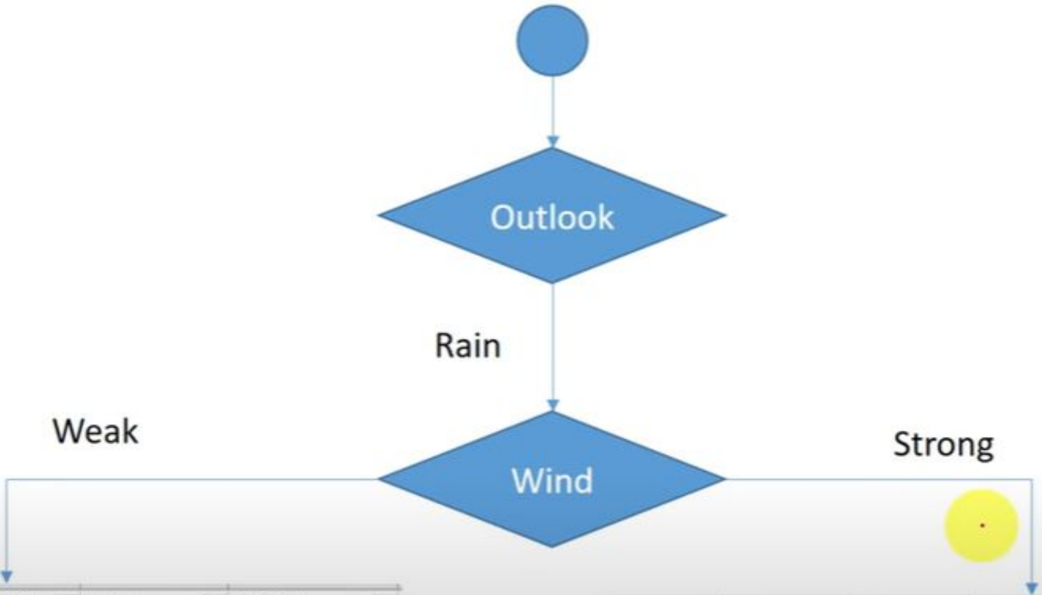
Rainy outlook and wind

Wind	Standard deviation for golf players	instances
Weak	3.09	3
Strong	3.5	2

Weighted standard deviation for rainy outlook and wind = (3/5)x3.09 + (2/5)x3.5 = 3.25

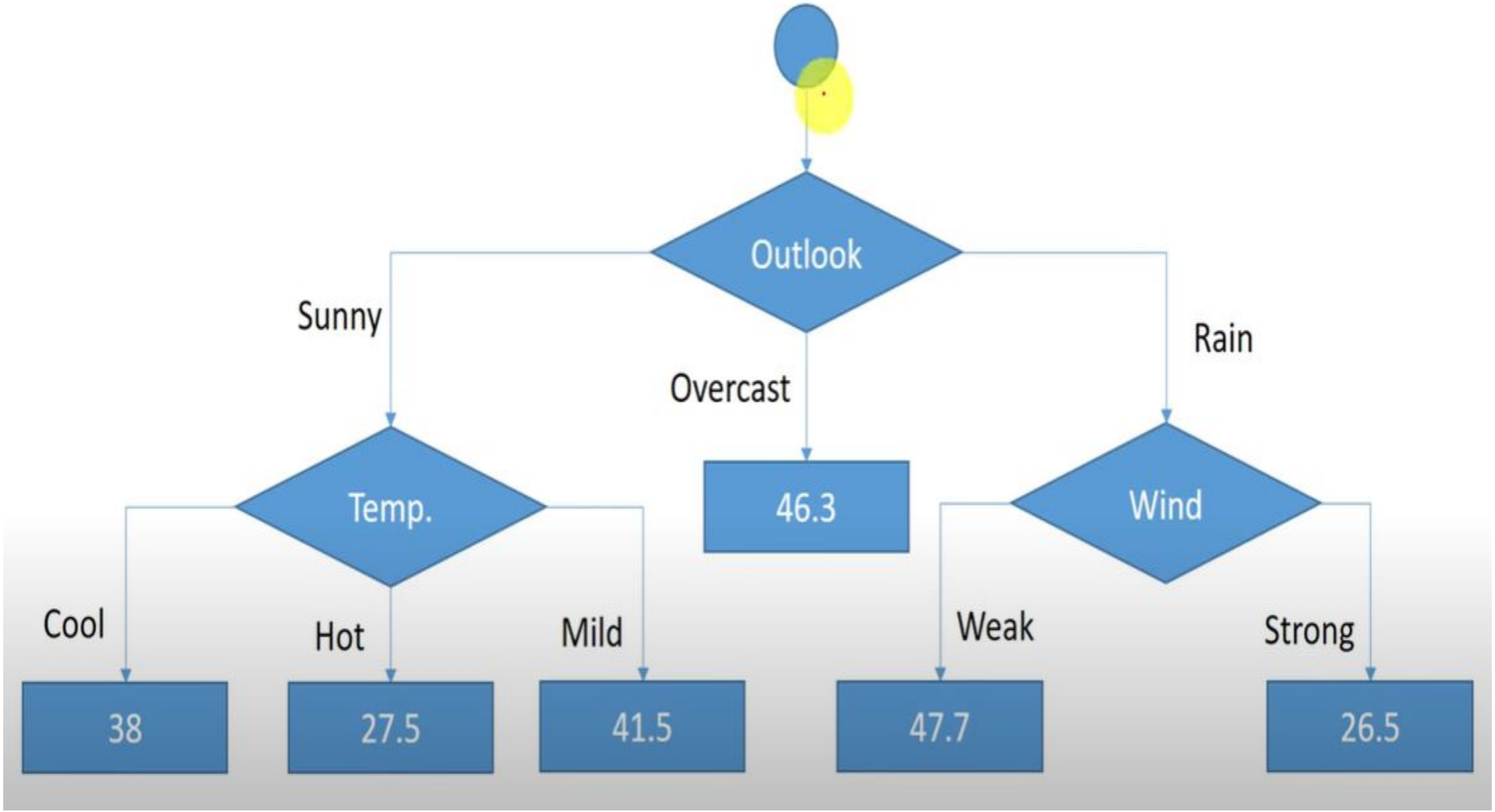
Standard deviation reduction for rainy outlook and wind = 10.87 – 3.25 = 7.62

Feature	Standard deviation reduction
Temperature	0.67
Humidity	0.37
Wind	7.62



Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
10	Rain	Mild	Normal	Weak	46

Day	Outlook	Temp.	Humidity	Wind	Golf Players
6	Rain	Cool	Normal	Strong	23
14	Rain	Mild	High	Strong	30


Outlook
Sunny
Overcast
Rain
Temp.
46.3
Wind
Cool
Hot
Mild
Weak
Strong
38
27.5
41.5
47.7
26.5

References

Decision Tree CART - Machine Learning Fun and Easy - YouTube

CART (Classification And Regression Tree) in Machine Learning - GeeksforGeeks

Regression Decision Tree Solved Example Regression Trees in Machine Learning by Mahesh Huddar - YouTube

Machine Learning - Standard Deviation (tutorialspoint.com)