MACHINE LEARNING ALGORITHM

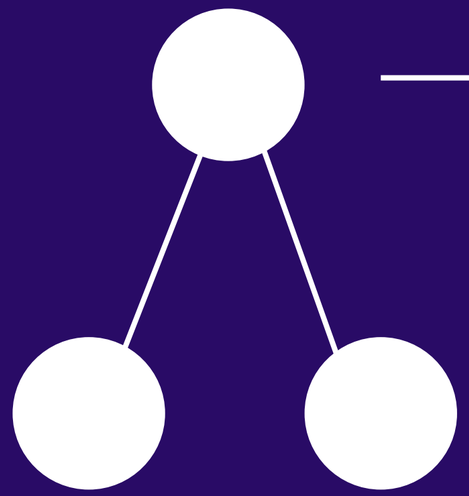
DECISION TREE CLASSIFICATION

CHI-SQUARE AUTOMATIC INTERACTION DETECTION (CHAID)

PRESENTED BY :
RACHITA C
BATCH 122

What is Decision Tree?

- Decision tree is a divide and conquer problem solving stratergy.
- Non-Parametric algorithm



⟶ Root Node = Has dataset

⟶ Branches = Splitting is done through intelligent stratergies

CHAID	CART	C 4.5
Chi-square	Gini index; Twoing criteria	Entropy info-gain

Why Decision Tree?

- These are predictive model with higher accuracy.
- Simple to understand.
- It will automate the process so that any distribution can be solved.

In the decision tree

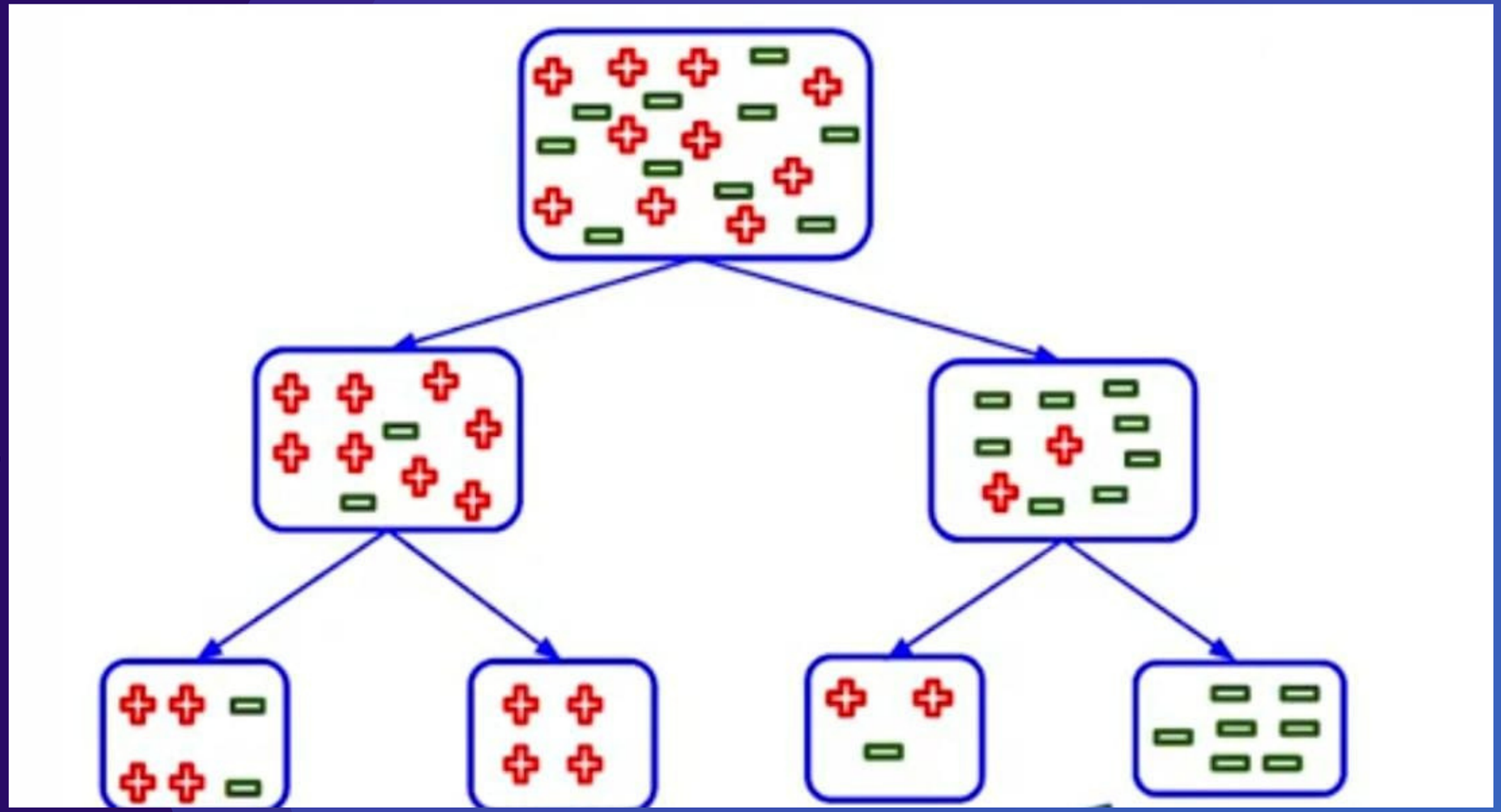
- Collections of record => Training Set
- Each record has set of attributes and one of them is class.
- We find a model for the class attribute as a function of values of other attributes.

Example: Anomaly Detection can be written as follows

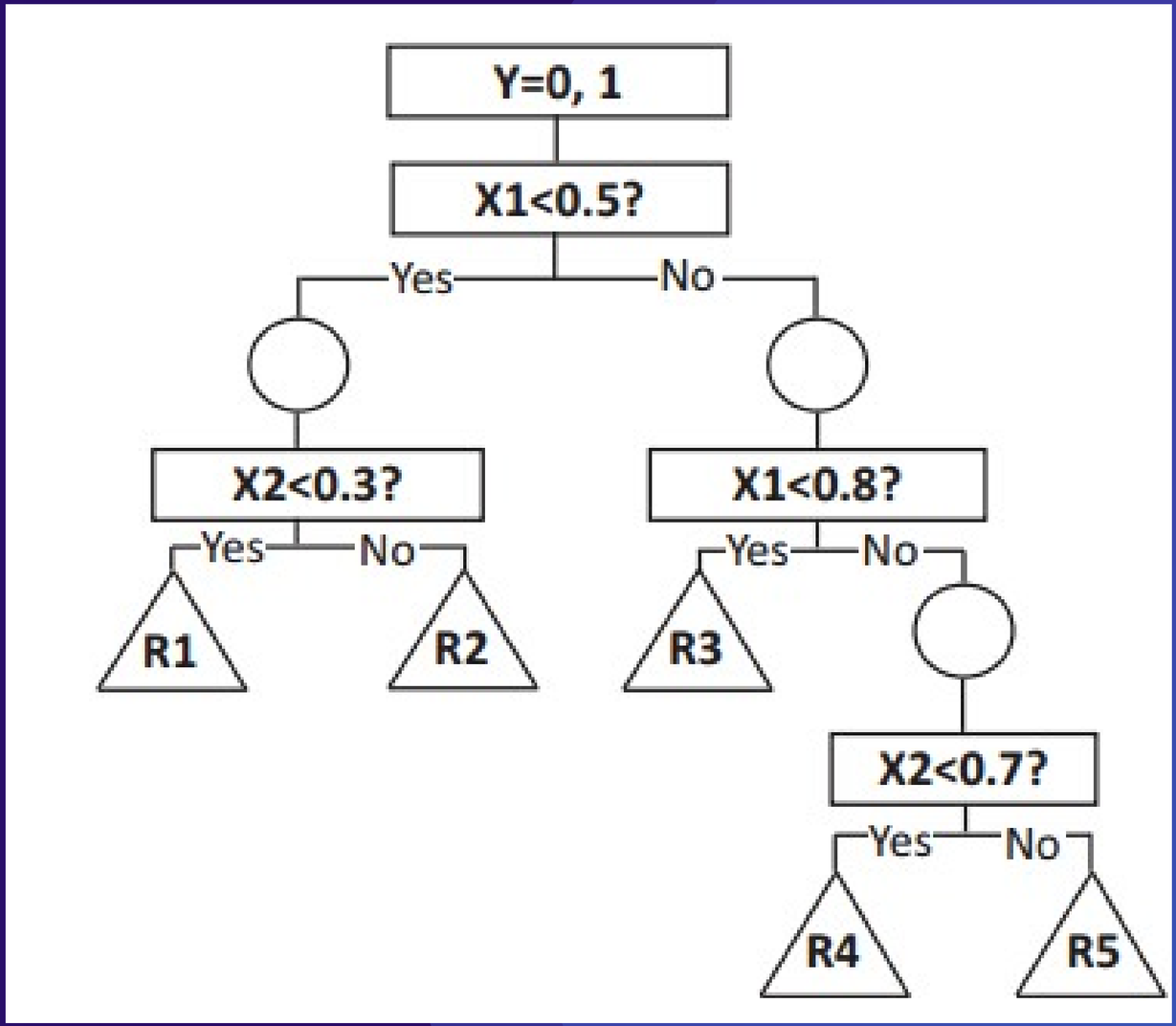
Anomaly = f(Amount, Time, Location)

Terminologies Related to Decision Tree

- Root Node
- Splitting
- Decision Node
- Leaf / Terminal Node
- Branch / Sub Tree
- Parent and Child Node
- Depth of tree



Sample decision tree based on binary target variable Y



What is Chi-Square?

It's a measurement metric to find the important feature used to find distance between the child and the parent nodes

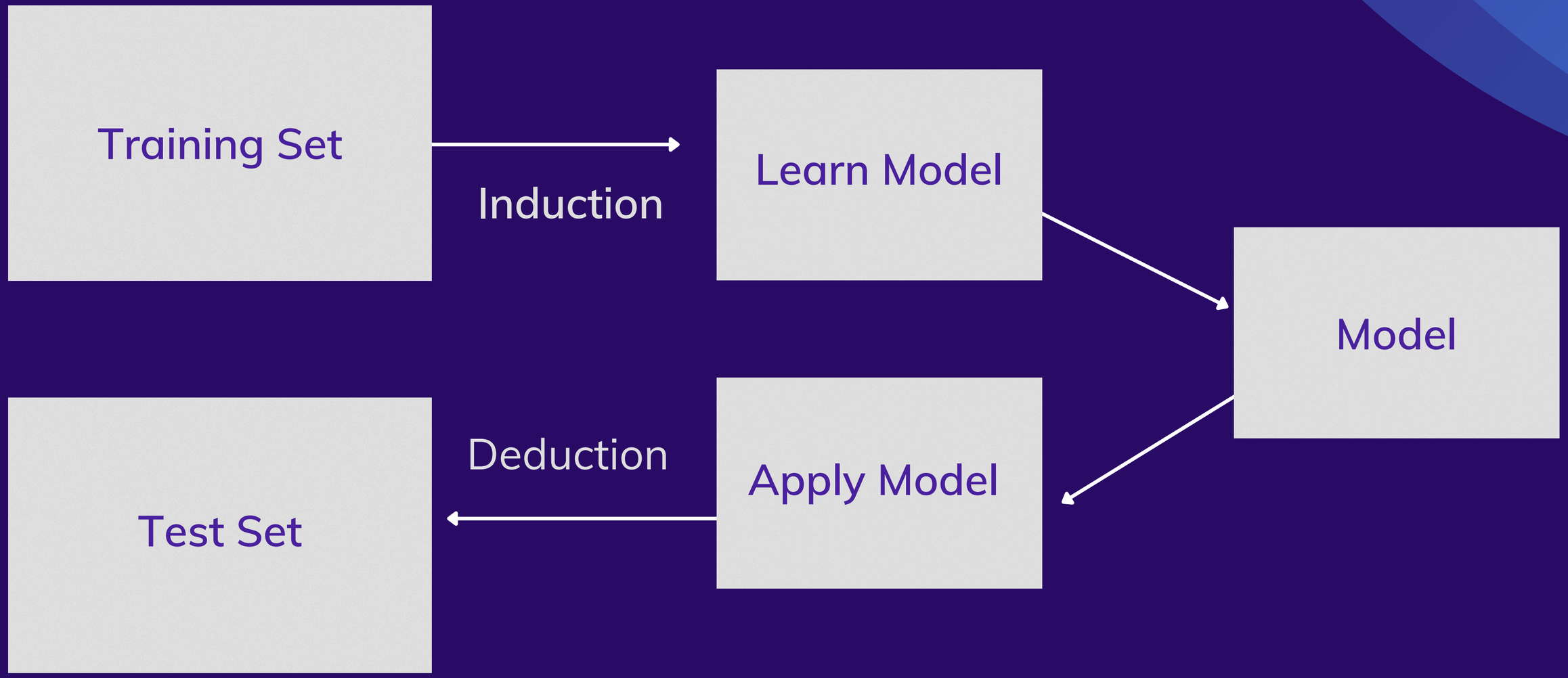
It's given by

Chi-Square = $\sqrt{\frac{(Y - Y`)}{Y`}^2}$

Where

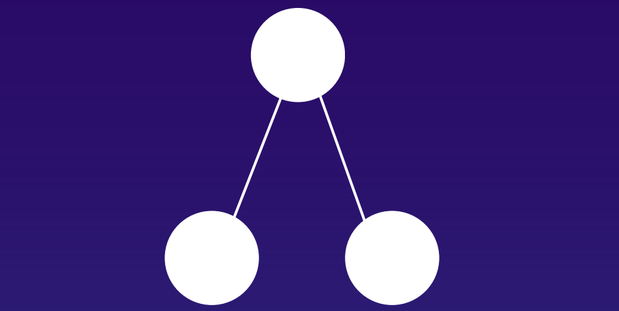
Y = Actual

Y` = Predicted / Expected


Training Set
Induction
Learn Model
Model
Test Set
Deduction
Apply Model

Tree Induction

Split the record based on an attribute test that optimizes certain criterion



Some of the issues to be addressed:

- How are attributes selected in decision trees?
- How to determine the best split?
- When to stop the split?

Tree Induction

- Split can be : two-way split and Multiway split
- Split on continuous attribute are of two types: Discrete and Binary
- The optimization happens in such a way that the difference in percentage of distribution should be high

CO	**6**	**co**	**4**
c1	**4**	**c1**	**6**

CO	**1**	**C0**	**8**	**CO**	**1**
C1	**3**	**C1**	**0**	**C1**	**7**

1st --> 2:2
2nd --> 2:8:6

- Nodes with homogeneous class distributions are preffered

CO	**5**
c1	5

5-5 = 0
High degree of impurity

CO	**9**
c1	**1**

9-1 = 8
Low degree of impurity

1st --> Non-Homogeneous
2nd --> Homogeneous

- Splitting should be stopped when all records belong to same class or similar attribute value.

Implementation of Algorithm

We have a well known Golf Dataset. The main objective of this classification activity is to decide if or not to go out to play Golf sport.

Reference: " Implement Of Decision Tree Using Chi_Square Automatic Interaction Detection " by Analitics Vidya

Day	Outlook	Temp.	Humidity	Wind	Decision
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

1) Humidity feature which has a split of High and Normal

	yes	No	Total	Expected	Chi-square Yes	Chi-square No
High	3	4	7	3.5	0.267	0.267
low	6	1	7	3.5	1.336	1.336

Chi-square yes for high humidity is √((3– 3.5)^2 / 3.5) = 0.267
whereas actual is 3 and expected is 3.5.
So, the chi-square value of the humidity feature is
= 0.267 + 0.267 + 1.336 + 1.336
= 3.207

Chi-Square = $\sqrt{\frac{(Y - Y`)^2}{Y`}}$

Where
Y = Actual
Y` = Predicted / Expected

2) Wind feature which has a split of Weak and Strong

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	5	2	7	3.5	0.802	0.802
Strong	3	3	6	3	0.000	0.000

Herein, the chi-square test value of the wind feature is

= 0.802 + 0.802 + 0 + 0

= 1.604

3) Temparature feature which has a split of Hot, Mild and Cool

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Hot	2	2	4	2	0	0
Mild	4	2	6	3	0.577	0.577
Cool	3	1	4	2	0.707	0.707

Herein, the chi-square test value of the temperature feature is

= 0 + 0 + 0.577 + 0.577 + 0.707 + 0.707

= 2.569

4) Outlook feature which has a split of Sunny, Outcast and Rain

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Sunny	2	3	5	2.5	0.316	0.316
Overcast	4	0	4	2	1.414	1.414
Rain	3	2	5	2.5	0.316	0.316

Herein, the chi-square test value of the outlook feature is

= 0.316 + 0.316 + 1.414 + 1.414 + 0.316 + 0.316

= 4.092

The outlook column has the most elevated and highest chi-square value.

Feature	Chi-square value
Outlook	4.092
Temperature	2.569
Humidity	3.207
Wind	1.604

Sunny

Rain

overcast

Day	outlook	Temp.	Humidity	wind	decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	outlook	Temp.	Humidity	wind	decision
4	Rain	Mild	High	Weak	Yes
5	Rain	cool	Normal	Weak	Yes
6	Rain	cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	outlook	Temp.	Humidity	wind	decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

On the image above, we've split the raw data based on the outlook classifications. In the sub informational dataset.

Both sunny and rain branches have yes and no decisions. And Outcast has all Yes in it. We will apply chi-square tests for these sub informational datasets.

1) Outlook is Sunny ---> Humidity, Wind and Temparature

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

1) Humidity feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
High	0	3	3	1.5	1.225	1.225
Normal	2	0	2	1	1	1

Chi-square value of humidity feature for sunny outlook is
= 1.225 + 1.225 + 1 + 1
= 4.449

2) Wind feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	1	2	3	1.5	0.408	0.408
Strong	1	1	2	1	0	0

Chi-square value of wind feature for sunny outlook is
= 0.408 + 0.408 + 0 + 0
= 0.816

3) Temperature feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Hot	0	2	2	1	1	1
Mild	1	1	2	1	0	0
Cool	1	0	1	0.5	0.707	0.707

So, the chi-square value of temperature feature for sunny outlook is
= 1 + 1 + 0 + 0 + 0.707 + 0.707
= 3.414

Therefore Humidity has the highest Chi-square value so its converted as a clild node.

Feature	Chi-square
Temperature	3.414
Humidity	4.449
Wind	0.816

Sunny

Rain

overcast

Day	outlook	Temp.	Humidity	wind	decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes

Day	outlook	Temp.	Humidity	wind	decision
4	Rain	Mild	High	Weak	Yes
5	Rain	cool	Normal	Weak	Yes
6	Rain	cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

High

Normal

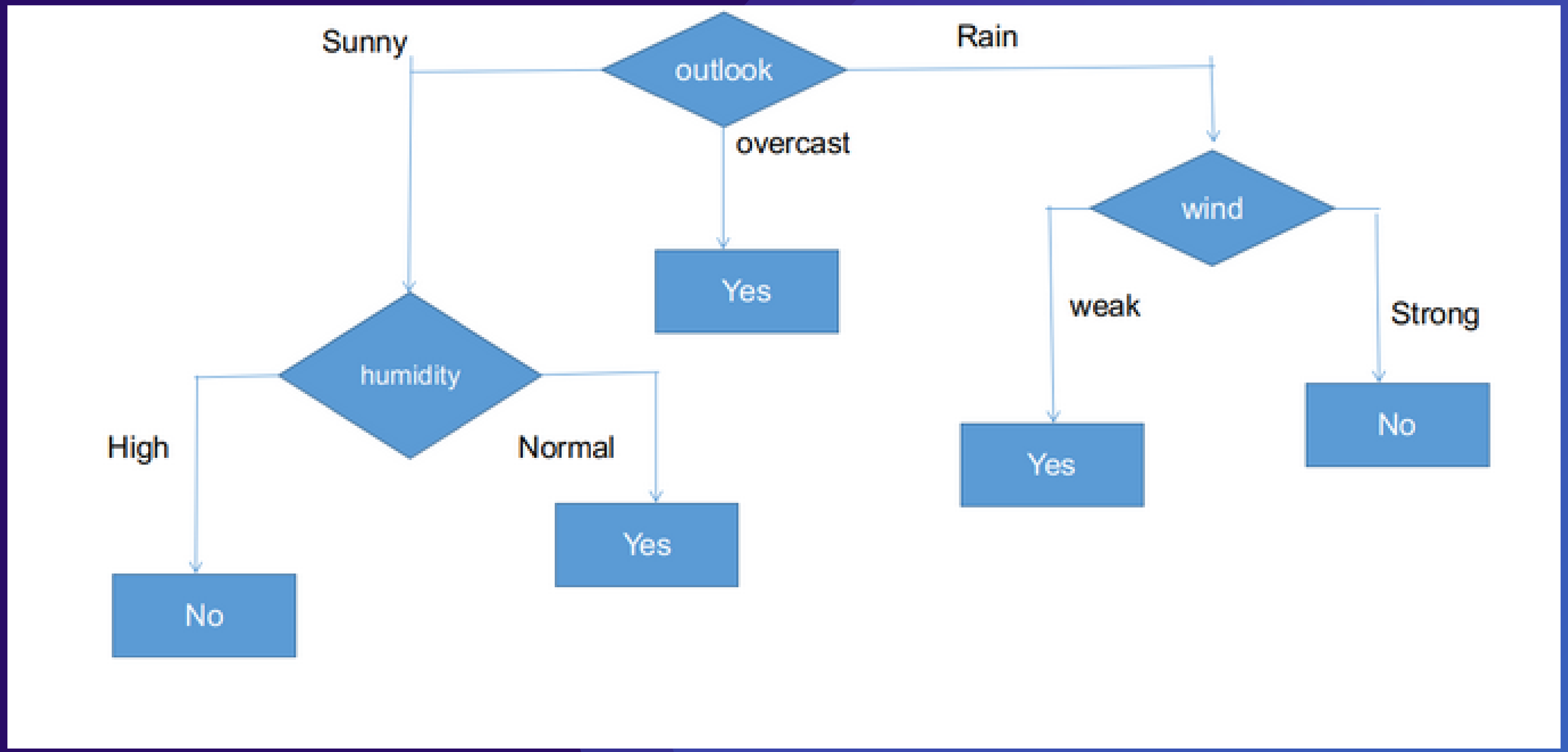
Day	outlook	Temp.	Humidity	wind	decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

Day	outlook	Temp.	Humidity	wind	decision
9	Sunny	cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Similarly do it with respect to Rain feature

Feature	Chi-squared
Temperature	0.816
Humidity	0.816
Wind	4.449

Therefore Wind has the highest Chi-square value so its converted as a child node.



The final form of the CHAID tree.

Pros

- Inexpensive to construct.
- Less Data Preparation and it is Robust to outliers.
- Extremely fast at classifying unknown records.
- Easy to interpret for small sized trees.
- Accuracy is comparable to other classification technique for many simple datasets.
- Non-parametric approach.

Cons

- Overfitting : Pruning
- They are largely unstable compared: A small change in data can lead to major change in the tree.
- Decision tree often involves higher time to train the model.

.

Thank You.