* Data mining - Process of extracting extra knowledge & hidden patterns from large amounts of data

Steps in data mining

1. Data selection - extracting relevant data from various sources
2. Data Preprocessing - Converting data to consistent form
   Removal of unnecessary info
3. Data Transformation - suitable format
4. Data mining - finding hidden patterns
5. Pattern evaluation
6. Presenting

* Data mining applications

(i) Customer segmentation
(ii) Risk management
(iii) Fraud detection

Data Mining

→ Types of attributes

① Categorical (Qualitative)

(i) Nominal :- They are just some kind of symbols used to distinguish one object from other.
eg - Zip codes, employee ID

(ii) Ordinal :- They define an ordering of objects apart from just distinguish
eg - grades,

② Quantitative attributes (having properties of numbers)

(i) Numeric -

(a) Interval : Unit of measurement exist. (+, - can be applied, not ÷ & x).
eg - temp. in C & F

(b) Ratio : x & ÷ also can be applied
eg - length, mass, age

(ii) Discrete :- It has a finite or countably infinite set of values. Binary attributes are special types of discrete attributes.
eg - Profession, zip code

(iii) Continuous : It has an infinite no. of states.
eg - Infinite no. of float values b/w 2 & 3.

* For asymmetric attributes, it is a special type of binary attribute where both the values are not equally important.

→ Precision, Bias & accuracy

(i) Precision - Closeness of repeated measurements to one another. It is measured by S.D of set of values.

(ii) Bias - Variation of measurements from quantity being measured. Measured by Diff b/w mean of set of values & known value of quantity being measured.

→ Outliers

* The data points that have characteristics or values that are diff from most of the other data objects in data set.
* Unlike noise, they can be sometimes be of interest. (in fraud detection)

→ Missing values

Ways to handle it:

(i) Eliminate data objects or attributes:-
Reliable if missing values is less. Very risky to eliminate attribute as it may be imp for analysis.

(ii) Estimate missing values: we can use the method of KNN (K nearest neighbour).

(iii) Ignore the missing value during analysis

→ Issues related to applications

1) Timeliness: If data provides a picture of ongoing phenomena, such as purchasing behaviour, this is relevant for only a limited time

2) Relevance: Common problem of sampling bias: When a sample does not contain diff types of objects in proportion to actual occurrence in population eg - survey responses are not by whole population

→ Data preprocessing

Topics -

(i) Aggregation
(ii) Sampling
(iii) Dimensionality reduction
(iv) Feature subset selection
(v) Feature creation
(vi) Discretization
(vii) variable transformation

(1) Aggregation

→ Combining of two or more objects into a single object.

→ How the aggregation is done is somem to look out for

→ Aggregation leads to smaller data sets which require less memory & time

Adv →

→ A group of objects generally have less variation than individual objects.

⇒ Data Quality

* Examples of data quality problems

(i) Noise & Outliers
(Regression, binning, clustering) (box plot & SD method)

(ii) missing values
(KNN method, remove)

(iii) duplicate data

(iv) Inconsistent values

(i) Noise :- Refers to modification of original values. eg - distortion of person's voice.

(2) Sampling

* Used because whole dataset is too expensive & time consuming to look at so we look at a sample
* Sample size should be large enough
* Using a sample should work as using the entire data set
* Sample is representative of dataset if it has a mean close to mat of original data.
* Sampling approach -

(1) Simple random sampling
→ Randomly selecting the objects
→ equal prob. of selecting any object

(2) Sampling without replacement
→ as each item is selected, it is removed from whole popⁿ

(3) Sampling with replacement
→ objects are not removed from popⁿ as they are selected
→ More simpler to analyze as prob. of selecting any obj remains constant throughout

* These approaches do not adequately represent all object types.

④ Stratified sampling :- Splitting the data into several partitions, draw random samples from each partition.

→ Equal no. of obj are drawn from each partition ensuring nothing is left out.

* Sample size should be large enough for accuracy but it eliminates purpose of sampling.
Small sample size leads to erroneous conclusions.

⑤ Progressive sampling

* → Proper sample size can be difficult to determine, hence we start with small sample size & progresses to large one which is a sample of sufficient size
→ When we need to look when we need to stop increasing sample size ✗ Note

③ Dimensionality reduction

* Data sets may have large no of features & attributes which are not necessary.
* Data mining algos work better on data set with less attributes

* Data can be more easily visualized.

* Curse of Dimensionality :-

As the dimension of data increases, it becomes more harder to analyze as chances of noisy & irrelevant data increases

* Techniques :-

(i) Principal component analysis (PCA)
Linear algebra techniques to find new attributes that are linear combination of original attributes & are orthogonal to each other

(ii) Singular value decomposition (SVD)

(4) Feature Subset Selection -

* Creating new attributes that are a combination or a subset of old attributes.

* Techniques :-

(i) Embedded approach - Feature selection occurs naturally as part of data mining algo. The algo itself decides the attributes
(ii) Filter approach - Features are selected before data mining algo is run through independent task
(iii) Wrapper approach - The data mining algo is used as black box to find best subset of attributes

(5) Feature creation

* To create new set of attributes from original attributes

* Techniques

(i) Feature extraction - The creation of a new set of features from original raw data

(6) Discretization

* To replace raw values of numeric attribute by interval levels.

* Numerous continuous attributes are replaced by small intervals.

* Techniques

(i) Equal width discretization : It divides the data into equal intervals

(ii) Equal freq. discretization : Same no. of objects into each interval.

(iii) K means : Clustering method, generally forms clusters of similar objects

(7) Binarization

* If there are n categorical values, then $\lceil \log_2 n \rceil$ digits are req. to represent it in binary form

(8) Variable transformation

* Techniques :-

(i) Normalization or standardizati- Attribute values are normalized by scaling their values so that they fall in specified range.

eg- If we have our data which is scattered, we can use Z-transformation so that our values lie b/w 0 & 1.

$$\text{Z value} = \frac{x - \bar{x}}{\sigma}$$

$\bar{x} \rightarrow$ mean

$\sigma \rightarrow$ S.D

But if we have outliers in our data, we use median instead of mean in our calculat of SD.

⇒

→ Examples of Proximity decision

(i) Simple Let x & y be two objects that consist of n binary attributes

f_{00} → no. of attributes where x is 0 & y is 0
f_{01} → " " " where x is 0 & y is 1
f_{10} → " " " where x is 1 & y is 0
f_{11} → " " " where x is 1 & y is 1

(i) Simple matching coeff

$$SMC = \frac{\text{no. of matching attributes}}{\text{no of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

(ii) Jaccard coeff

$$J = \frac{\text{no. of matching presence}}{\text{no. of attributes except 00}} = \frac{f_{11}}{f_{10} + f_{11} + f_{01}}$$

(ii) Cosine similarity

$$\cos(x,y) = \frac{x \cdot y}{|x|\,|y|}$$

eg- $x = (3,2,0,5,0,0,0,2,0,0)$
$y = (1,0,0,0,0,0,0,1,0,2)$

$x \cdot y = 3\times1 + 2\times0 + 0\times0 + 5\times0 + 0\times0 + 0\times0 + 0\times0 + 2\times1 + 0\times0 + 0\times2$
$= 5$

$|x| = \sqrt{3^2 + 2^2 + 5^2 + 2^2 + \dots} = 6.48$

$|y| = \sqrt{1^2 + 0^2 + \dots + 1^2 + 2^2} = 2.24$

$\cos(x,y) = 0.31$

Ch-4

Classification :

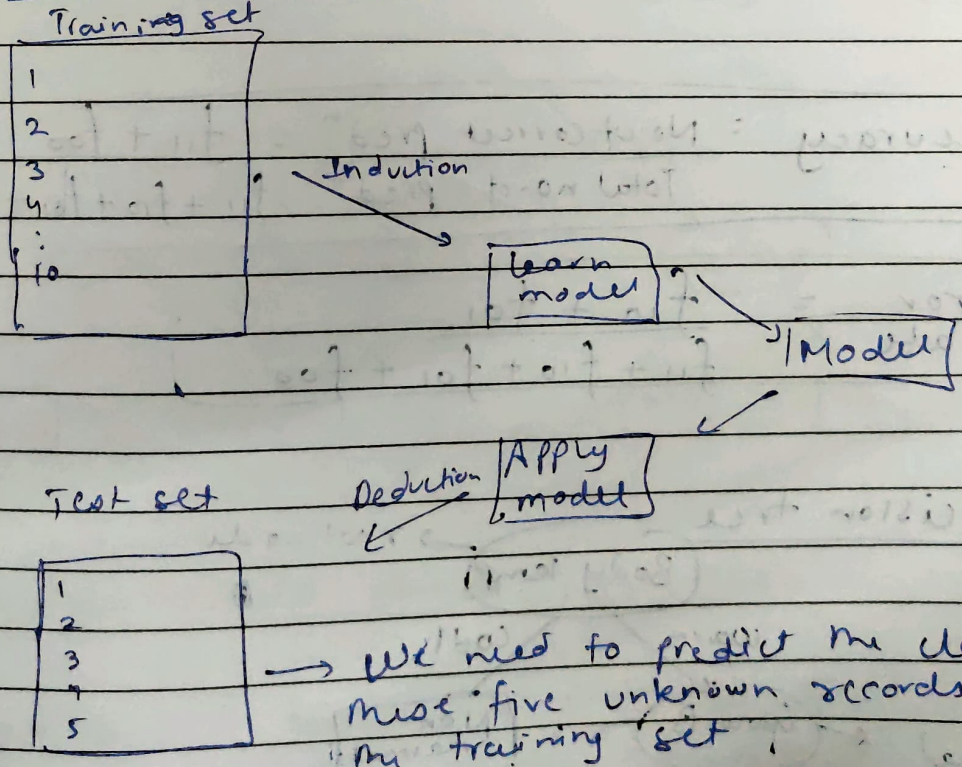
→ Definition

* Task of assigning objects to one of the predefined categories
* It is a task of mapping an input attribute set x into its class label y.
* It is usually done by a technique called decision tree induction

eg- classifying the species by looking at various attributes like Body temp, skin cover, etc into classes like mamal, reptile, bird, fish etc.

eg- categorizing news as finance, weather, entertain, sports

→ General approach



→ We need to predict the classes of those five unknown records using the training set.

(Predictive algo)

Evaluation of performance of classification model is based on number of test records correctly & incorrectly predicted by model. These numbers are represented in confusion matrix.

		Predicted class	
		Class = 1	Class = 0
Actual class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

No. of correct predictions = $f_{11} + f_{00}$

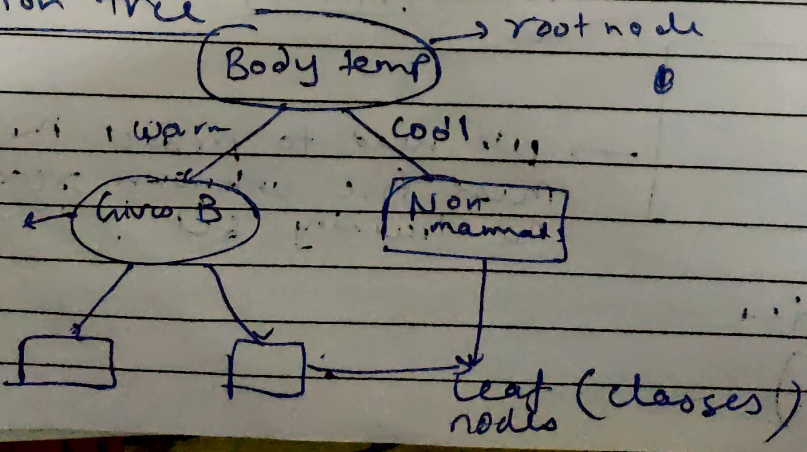
" " incorrect " = $f_{10} + f_{01}$

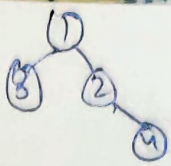
where f_{01} = no. of records from class 0 incorrectly predicted as class 1

* Accuracy = $\frac{\text{No. of correct pred}^n}{\text{Total no. of pred}^n} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$

* Error rate = $\frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$

→ Decision tree





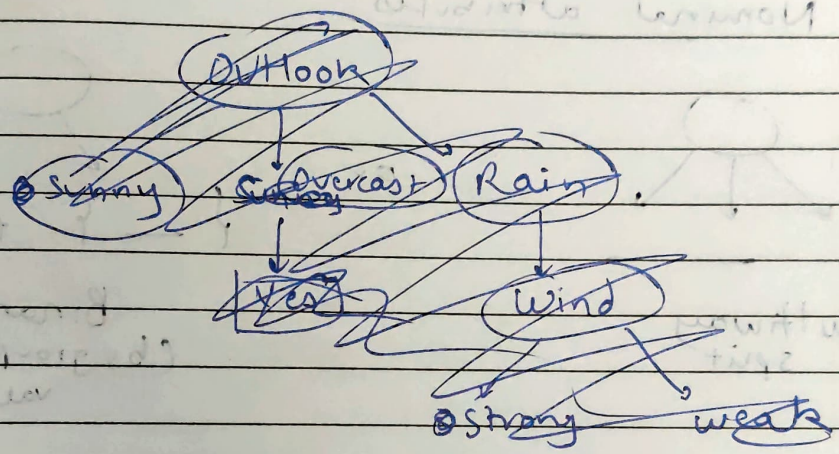
Now, on the basis of decision tree, we can classify the new objects that come.

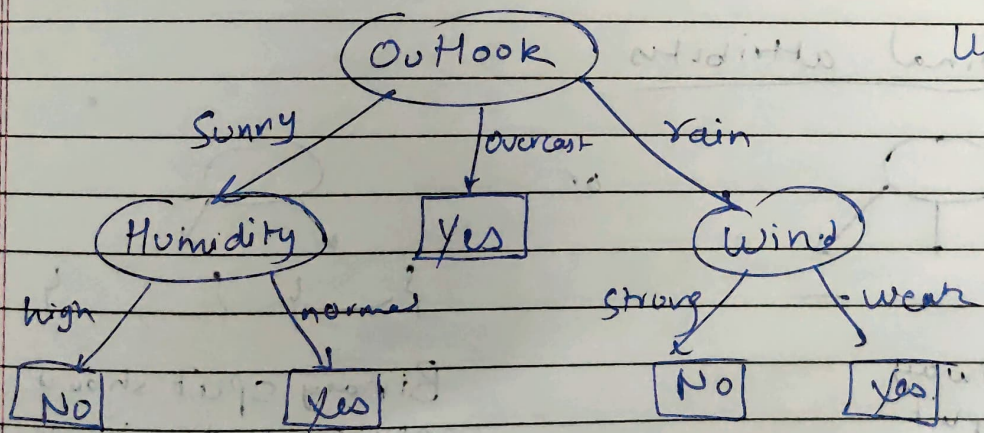
* How to build a decision tree

Hunt's algo :-

We have to build an optimal decision tree starting from an attribute which gives the most purer results as compared to other

eg-





We started with OutHook attribute

* Design issues of decision tree

→ How should training records be split

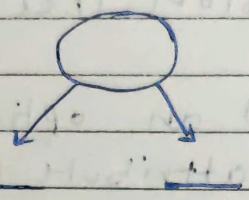
→ How should the splitting procedure stop.

Anushaka
Gupta

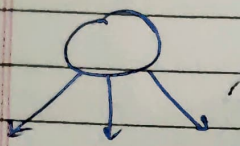
classmate
Date
Page

Methods for expressing Attribute test conditions

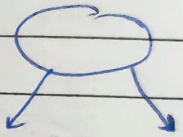
(i) Binary attributes — Two outcomes



(ii) Nominal attributes



or

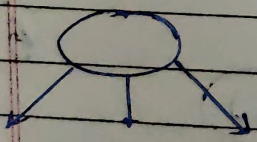


{ — } { — , — }

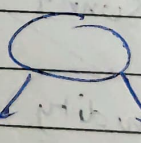
Multiway split

Binary split (by grouping attribute values)

(iii) Ordinal attributes



or



{ — } { — , — }

Multiway split

Binary split should ensure order is preserved
eg- small cannot be combined with large

(iv) Continuous attributes

Binary / Multiway Split

or

Through discretization, we can make intervals

★ Decision Tree using Gini index

Example

Weekend	Weather	Parents	Money	Decision
W_1	Sunny	Yes	Rich	Cinema
W_2	Sunny	No	Rich	Tennis
W_3	Windy	Yes	Rich	Cinema
W_4	Rainy	Yes	Poor	Cinema
W_5	Rainy	No	Rich	Stay in
W_6	Rainy	Yes	Poor	Cinema
W_7	Windy	No	Poor	Cinema
W_8	Windy	No	Rich	Shopping
W_9	Windy	Yes	Rich	Cinema
W_{10}	Sunny	No	Rich	Tennis

To decide root node

① Gini index for weather → 3 possibilities

→ Gini index of overall sample

6C + 2T + 1ST + 1SH

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

$$= 0.58$$

→ Gini index of money attribute

7 Rich + 3 Poor

3 Cinema, 3 Cinema, 2 Tennis, 1 Stay, 1 Shop

Money = Rich $= 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{2}{7}\right)^2 + \left(\frac{1}{7}\right)^2 + \left(\frac{1}{7}\right)^2\right]$

$= 0.694$

Money = Poor $= 1 - \left[\left(\frac{3}{3}\right)^2\right] = 0$

Weighted index $= \frac{7}{10} \times 0.694 + 0 = \boxed{0.486}$

→ Gini index of Parents attribute

5 Yes + 5 No

5 Cinema, 2 Tennis, 1 Cinema, 1 Stay, 1 Shop

Parents = Yes $= 1 - \left[\left(\frac{5}{5}\right)^2\right] = 0$

Parents = No $= 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$

$= 0.72$

Weighted index $= 0 + \frac{5}{10} \times 0.72 = \boxed{0.36}$

→ Gini index of weather attrin

3 Sunny + 3 Rainy + 4 Windy

1 Cinem | 2 Tennis | 2 Cinem | 1 Stayin | 3 Cinem | 1 shop

Weath = Sunny $= 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$

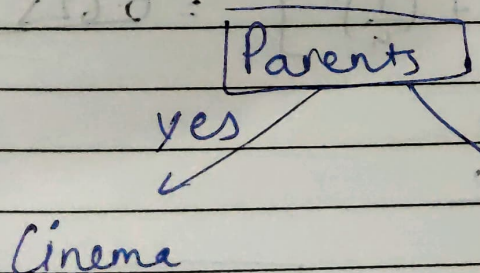
Weath = Rainy $= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$

Weath = Windy $= 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$

Weighted - $0.444 \times \frac{3}{10} + 0.444 \times \frac{3}{10} + 0.375 \times \frac{4}{10}$

$= \boxed{0.416}$

↑ So, we select Parent as root node



→ Gini index of Parents | weather

2 Sunny + 2 Windy + 1 Rainy

↓ | ↓

2 Tennis : 1 Cinema, 1 Shop in | Stay in

⇒ $1 - \left(\frac{2}{2}\right)^2 = 0$

⇒ $1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = \frac{1}{2} = 0.5$

⇒ $1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$

Weighted = $0 + 0 + \frac{2}{5} \times 0.5 = \boxed{0.2}$

→ Gini index of Parents | Money

4 Rich + 1 Poor

↓ | | | ↓

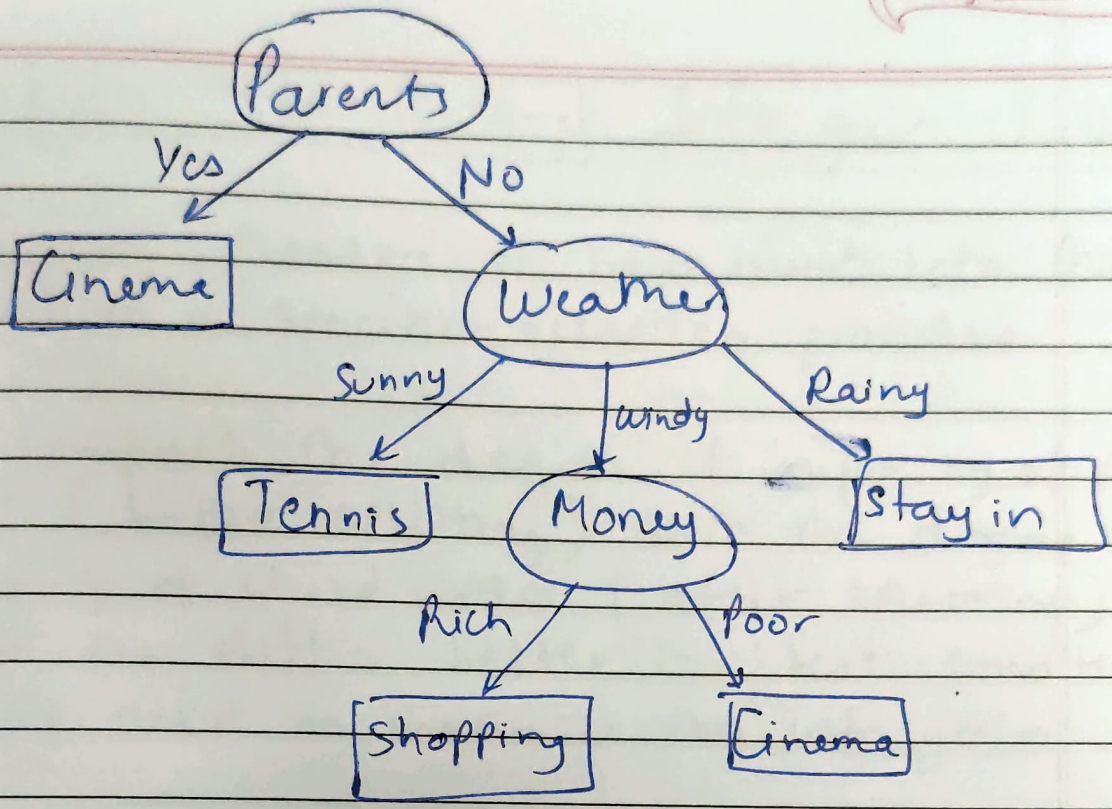
2 Tennis 1 Stay 1 Shop 1 Cinema

⇒ $1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.625$

⇒ $1 - \left[\left(\frac{1}{1}\right)^2\right] = 0$

Weighted = $0 + \frac{4}{5} \times 0.625 = \boxed{0.5}$

So, we select weather.



→ Decision tree through Info. gain & Entropy

* Info gain: Measure of how much info. the answer to a specific question provides

We compare the degree of impurity of parent node (before splitting) with the degree of impurity of child nodes (after splitting). Larger the diff. better the test condition. So, we need to maximize the info gain.

* Entropy - measure of how much uncertainty is there in info.

* Info gain ↑ → Entropy ↓

* Formulas :-

Info gain : $$I(p_i, n_i) = -\frac{p}{s}\log_2\frac{p}{s} - \frac{n}{s}\log_2\frac{n}{s}$$

S = Total sample size

Entropy : $$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{s} I(p_i, n_i)$$

Gain : Info gain − Entropy

Example

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

Info gain : 9 Yes + 5 No

$$= -\left[\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right]$$

$$= -\left[\frac{9}{14}(3.16 - 3.80) + \frac{5}{14}(2.32 - 3.80)\right]$$

$$= -\left[(0.64)(-0.64) + (0.357)(-1.48)\right]$$

$$= \left[-0.4096 - 0.52836\right]$$

$$= \boxed{0.937} \neq \boxed{0.940}$$

Entropy of Outlook

	Yes	No	Total	Prob
Sunny	3	2	5	5/14
Overcast	4	0	4	4/14
Rain	3	2	5	5/14

$$I(\text{Outlook, Sunny}) = -\left[\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right]$$

$$= -\left[(0.6)(1.58 - 2.32) + 0.4(1 - 2.32)\right]$$

$$= -\left[(0.6)(-0.74) + (0.4)(-1.32)\right]$$

$$= -\left[-0.444 - 0.528\right]$$

$$= 0.972$$

$I(\text{Outlook, Overcast}) = 0$

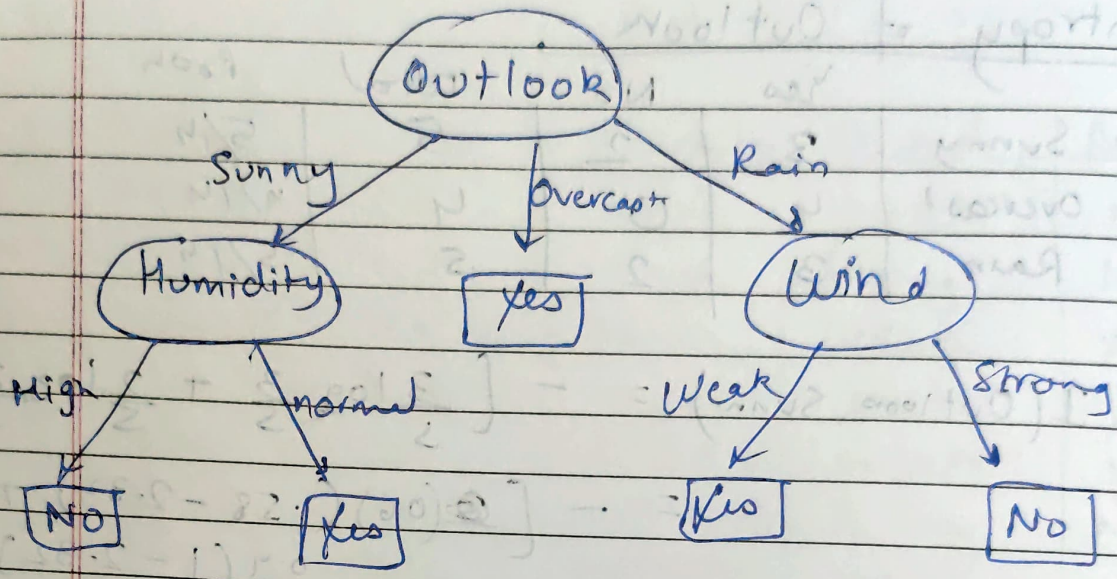
$I(\text{Rain}) = 0.972$

$$\text{Entropy} = 0.972 \times \frac{5}{14} + 0 + 0.972 \times \frac{5}{14}$$

$$= 0.70$$

Gain of Outlook $= 0.940 - 0.70 = 0.24$

Gain of Temp = 0.029
Gain of Humidity = 0.151
Gain of Wind = 0.048



* Impurity measures like entropy & Gini index tend to favour attributes that have large no. of distinct values.

* Customer ID cannot be used as root node because it is unique for every customer & we will not be able to use decision tree

* $$\text{Gain ratio} = \frac{\text{Info gain}}{\text{Split info}}$$

→

Evaluating performance of classifier

(1) Hold out method

* Original data is partitioned into two sets - training set & test set.
* Model is imposed on training set & tested on test set
* Proportion of data for training & test set can be 50-50 or 75-25.
* Accuracy of model on test set needs to be calculated
* Limitations :-
  (i) Fewer records are there for training set which affects accuracy of model
  (ii) Smaller the training set, larger the variance of model; larger the training set, we cannot judge accuracy from smaller test set.

(2) Random subsampling

* Holdout method repeated several times to improve estimation of classifier's performance is called random subsampling

(3) Cross validation

* Suppose we partition data into two equal subsets. First, we choose one of them for training, other for testing then, we swap the roles. The total error is sum of error for both runs.
* K-fold cross validation method generalises this by segmenting data into K equal partitions.