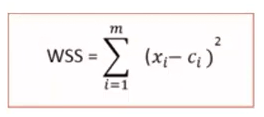
**Selecting number of clusters K**

**The Elbow Method**

To evaluate the performance of our k-means algorithm we can take a look at the inertia or objective function value. This is essentially the sum of squared distances our data points are away from their cluster centroid. By looking at different inertia values for different numbers of clusters (K)

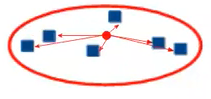
Within Cluster Sum of Squares (WCSS)

WCSS is defined as the sum of the squared distance between each member of the cluster and its centroid.



For each k value, we will initialize k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster center.

Inertia\_float Sum of squared distances of samples to their closest cluster center.



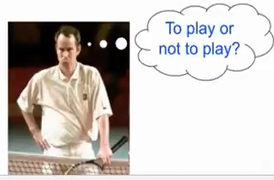
**Intra cluster distance**

This is what inertia evaluates: It tells us how far the points within a cluster are. So, inertia actually calculates the sum of distances of all the points within a cluster from the centroid of that cluster

We calculate this for all the clusters and the final inertial value is the sum of all these distances This distance within the clusters is known as intracluster distance so inertia gives us the sum of intracluster distances:

**Gini and Entropy**

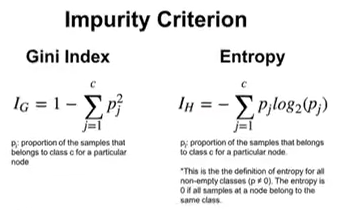
**Decision Tree**

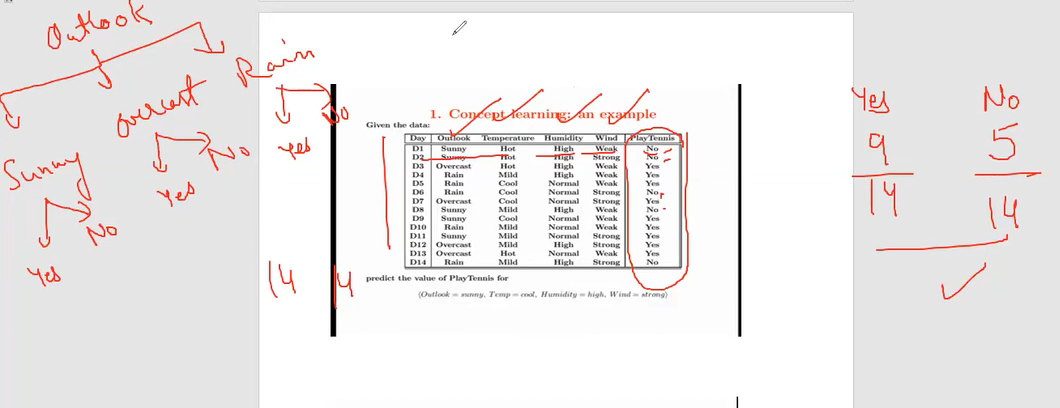


**Pure and impure dataset:-**

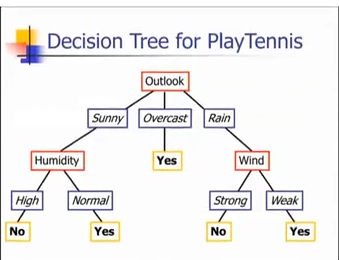
If your dataset has y label as like (0,0,0,0,1,1) this is impure we can call our dataset has randomness and if you dataset has only one type of value like (0,0,0,0,0) or (1,1,1,1,1) In y label that is a a pure dataset.But remember one thing we don’t need pure dataset .If our dataset is impure our machine could learn different type of situations ,for pure dataset it would not possible.

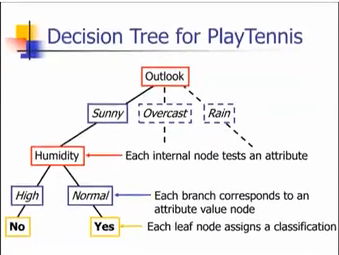
Entropy can judge a y column value with 0 to 1.If our entropy show us 1 then I should understand our dataset is balanced means our y column like(0,0,0,0,1,1,1,1) or(0,0,0,1,0,1,1,1).

****

****

****

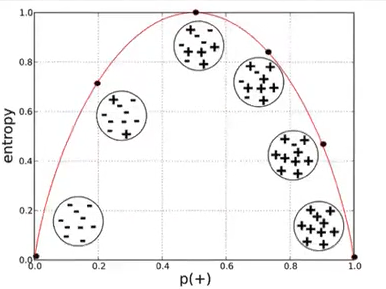
****

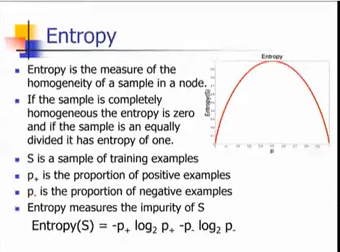
****

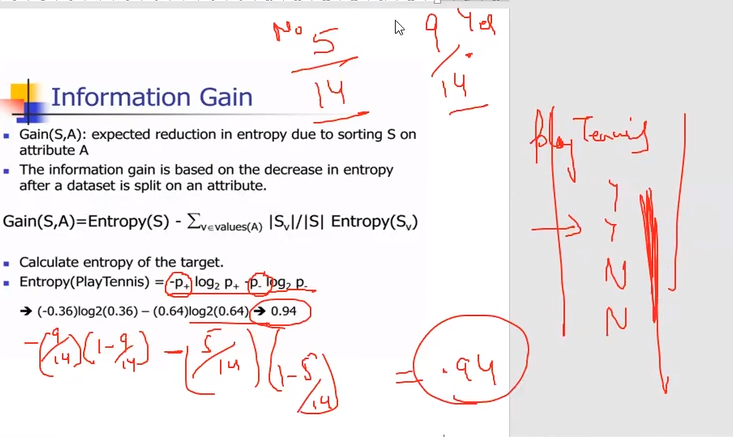
**Entropy**, as it relates to **machine learning**, is a measure of the randomness in the information being processed.

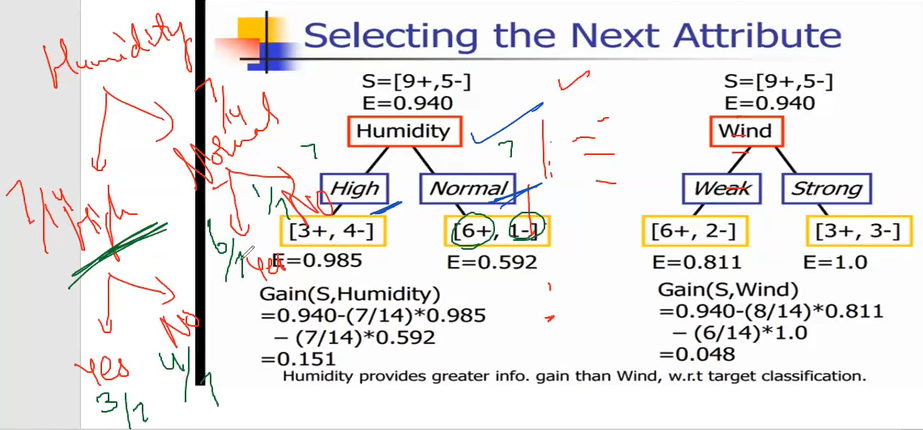
In case of diabetes dataset if all patients has no diabetes. Then no information is gained it is 0. Or if all people diabetes then also information gain is 0.But when

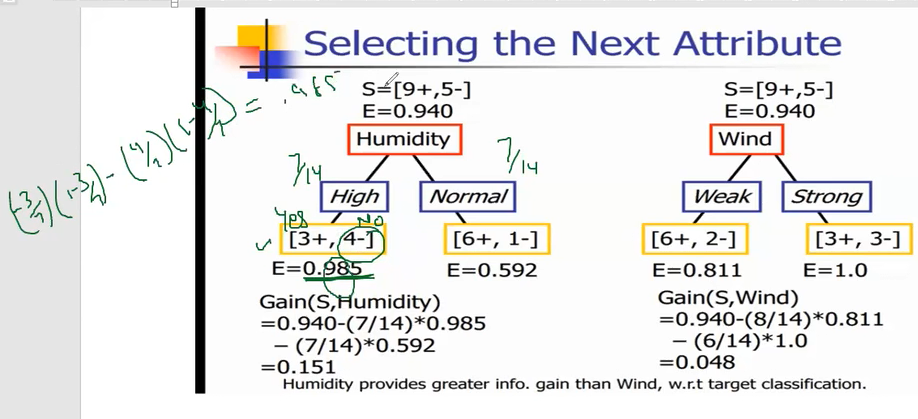
50% people has diabetes then entropy is 1.

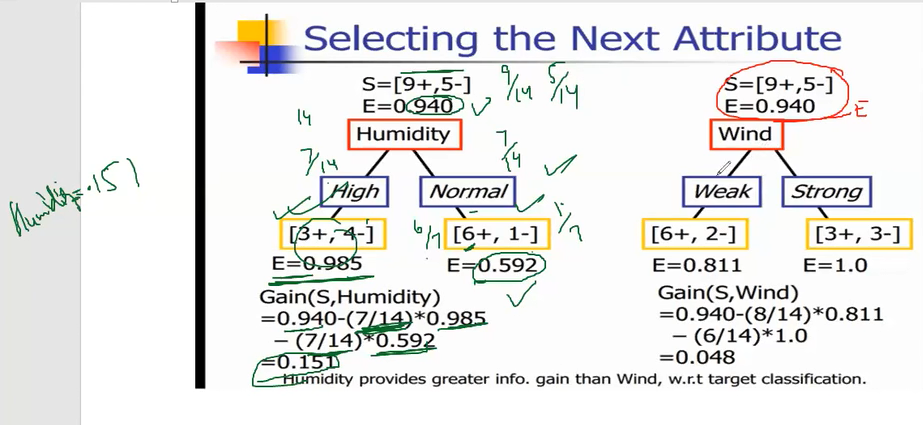


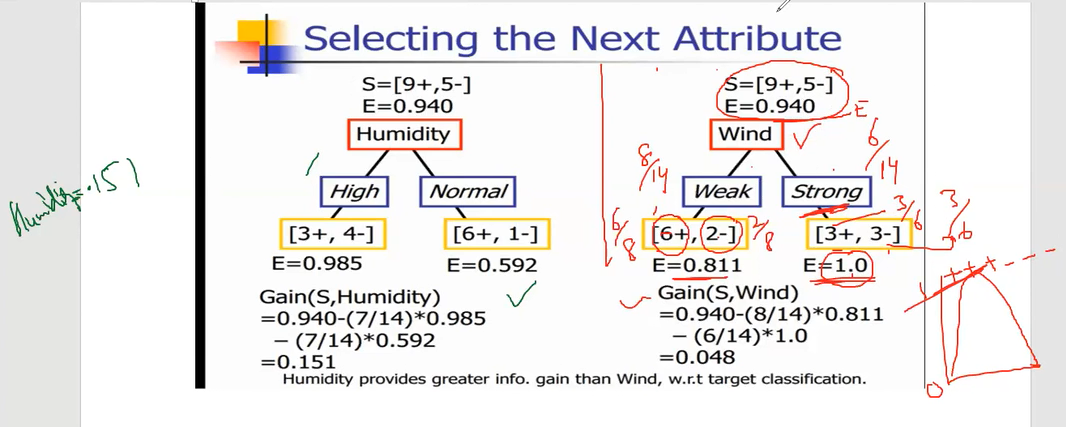


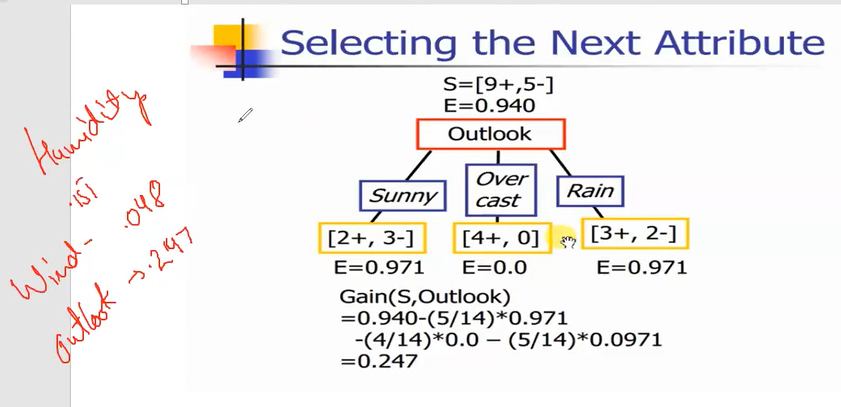




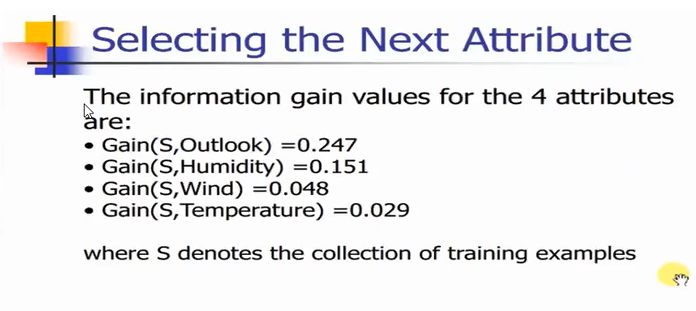


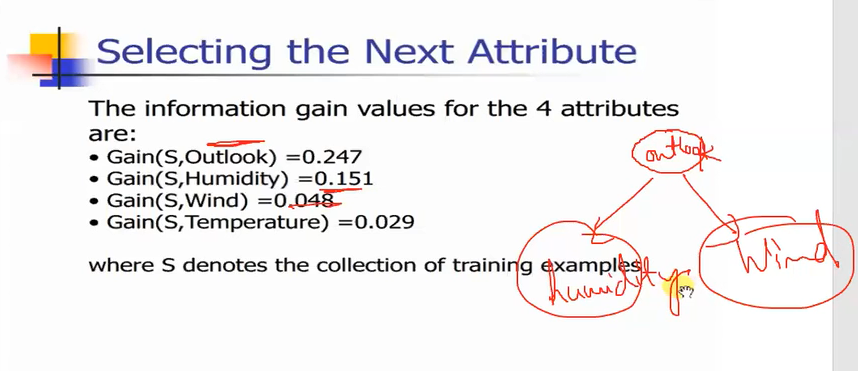


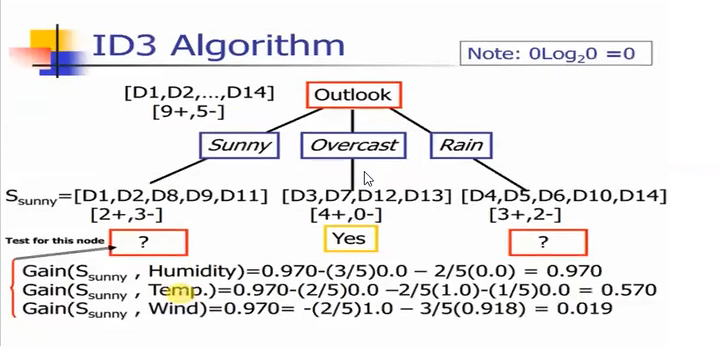


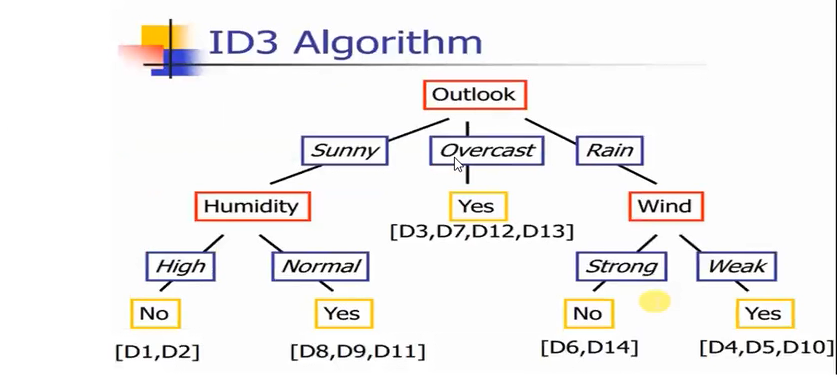


Here, overcast is 0.0 as it is very homogeneous sample.

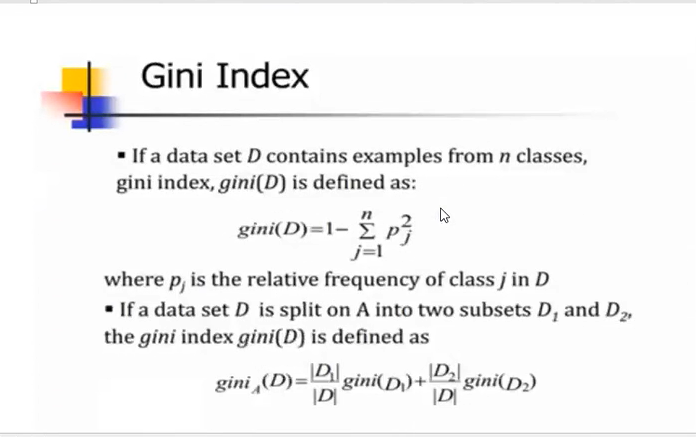




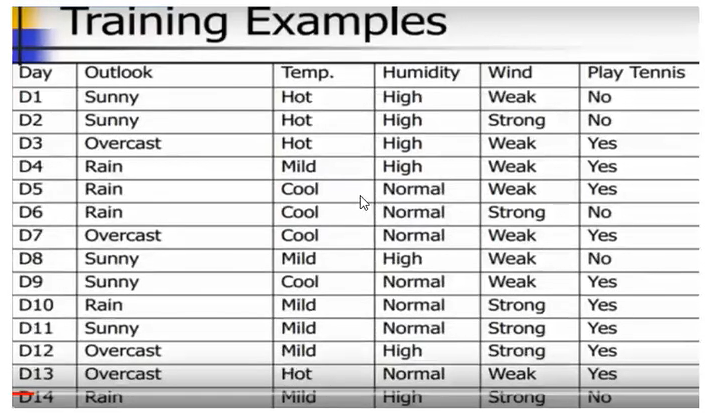




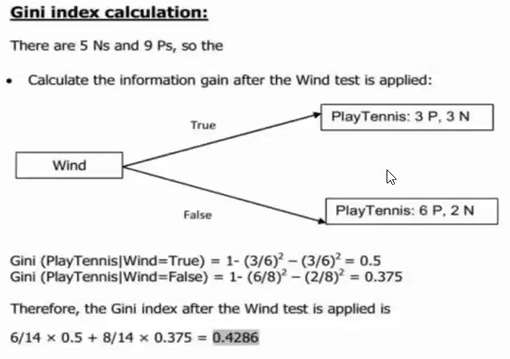
**Gini Index**

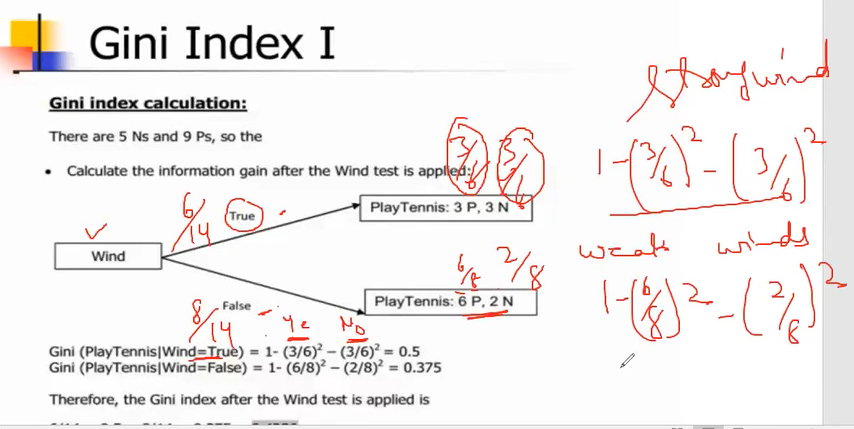


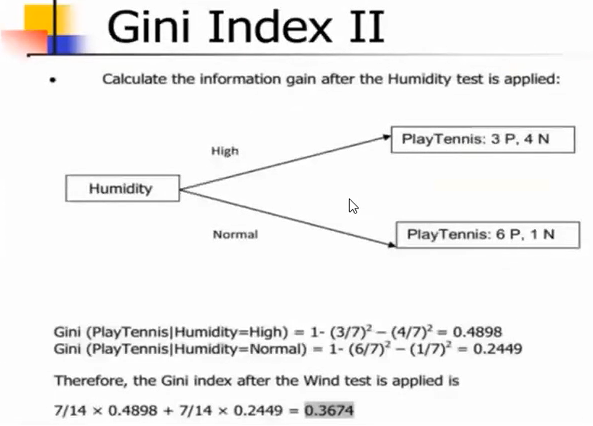




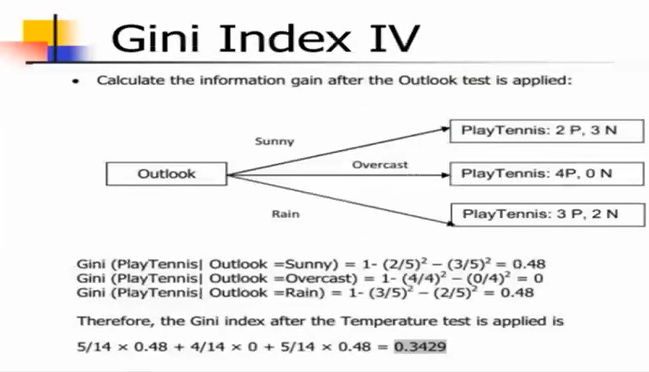
**Gini Index I**

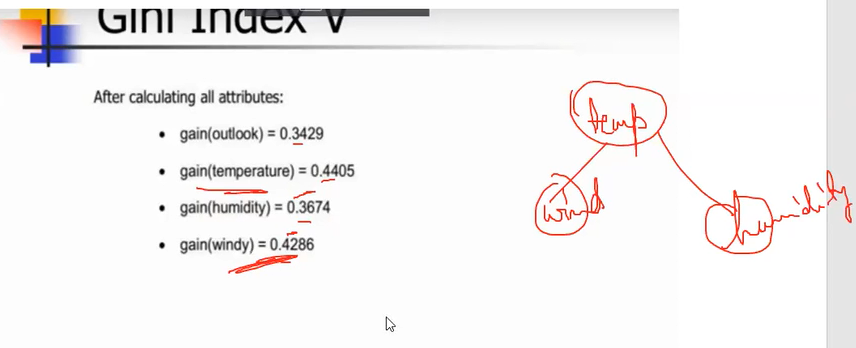












Here we are getting temperature as a highest gain 0.4405, hence

root node will be temperature.