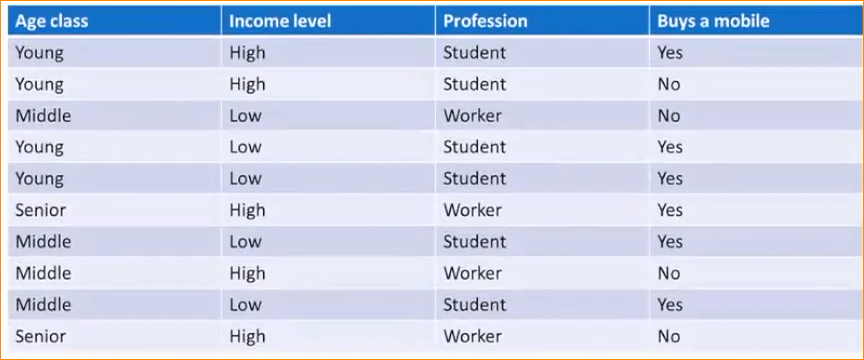
**Supervise Learning**

* Using when the dataset has classes/label(Ex- You have deiffernt variables like house area, number of rooms, number of bathrooms and output variable is house price), that dataset call label dataset.
* So if we have categorical variable we use classification. (Ex – True or false, positive or negative or neutral)

Sample dataset used in classification



This is structured data set.

Algorithms can work only structured data.

In here we have to predict person buy a mobile phone or not?

Dependent variable is – buy a mobile phone

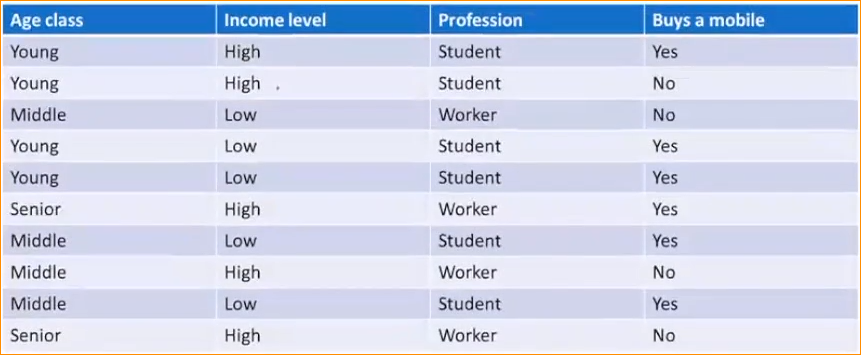
Independent variable – Age class, Income level, Profession

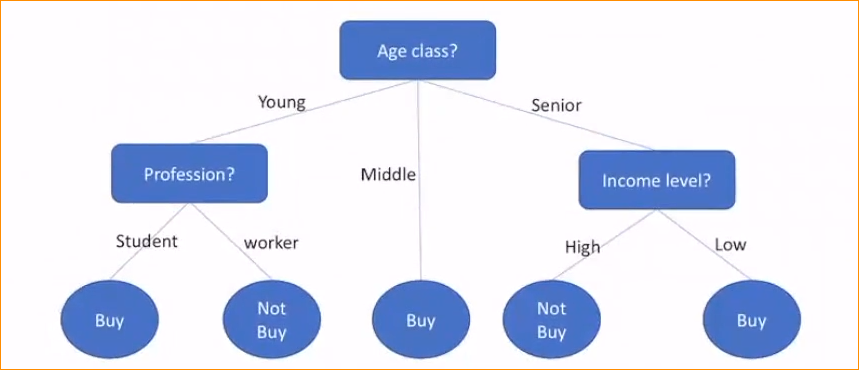
**What is Decision Trees ?**

* A tree structure starting with some observation about an item to conclusions.
* Make a prediction on the basis of a series of decisions along the branch of a tree
* Representing possible paths.
* In here we make set of rules and check one by one conditions and make a prediction.

**Sample Decision Tree**

In this decision tree made this dataset-





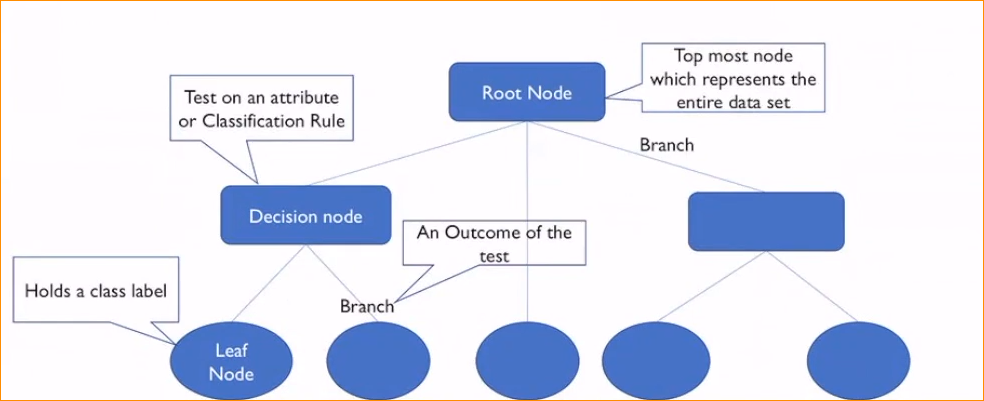
Level 02

Level 01

(Root node)

* In here select on input feature in one level
* This DT contains 2 main terms – nodes and leaves (Green)
* Variable represents as node (Red)
* Leave – that class labeled variable that we are going to predict.(Dependent variable)

**Decision Tree structure**

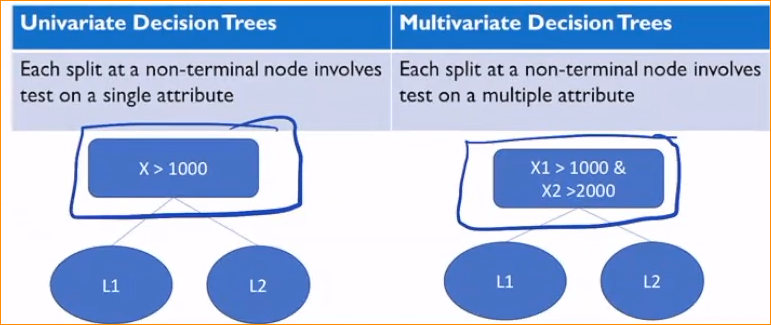


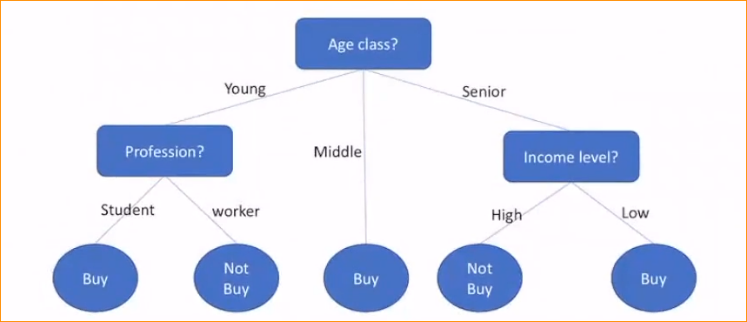
Base on a number of input features, decision trees divide in to two parts.

1. Univariate Decision Trees
2. Multivariate Decision Trees

If you only going to use one variable or one node. That is call **Univariate Decision trees.**

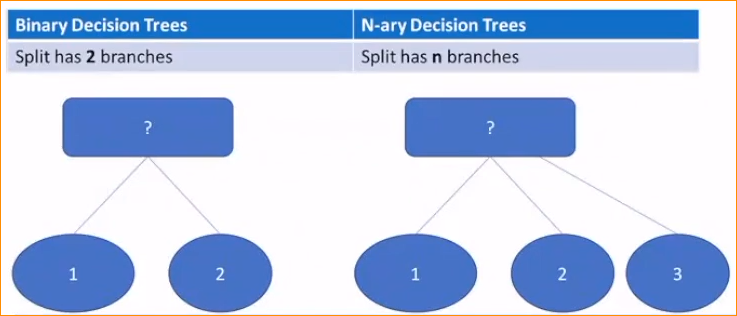
In one node if you are going to use many input features that we call a **Multivariate Decision tree.**

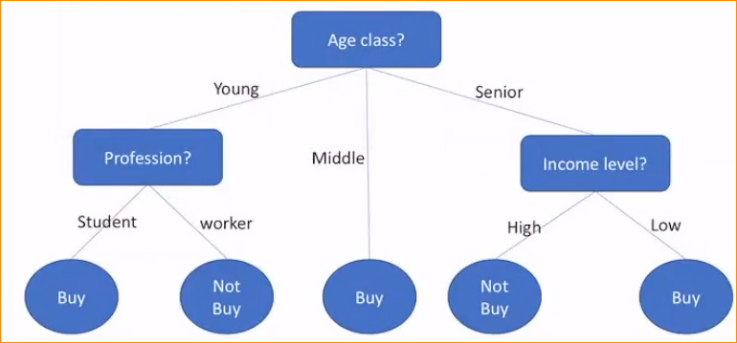


In here we using only one input feature. So this is **Univariate Decision tree.**

Based on number of children that base on Decision trees divided in to two parts.

1. Binary Decision Trees
2. N-ary Decision Trees



here we are using 3 class labels under that variable (Young, middle or Senior)

if you consider a profession there are two classes. Income level also.

But this is n Array Decision tree.

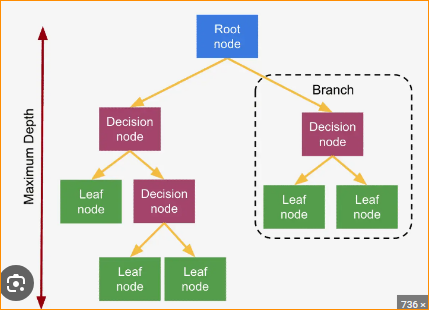
How to Construction of a Decision Tree?

* Select which is more suitable input variable to be selected?
* There are different methods

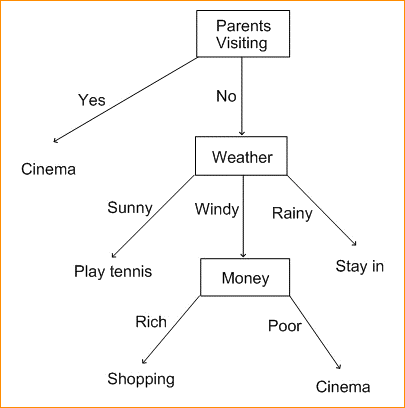
**Attribute Selection Measure**

Determine the most appropriate attribute to partition/spit samples at each node.

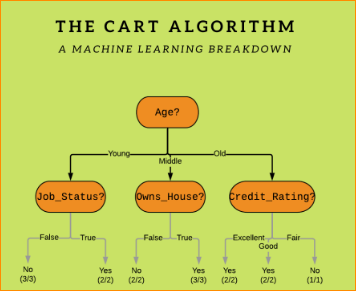
1. Information gain (used in ID3 algorithm)



1. Gain Ratio (used in C4.5 Algorithm)

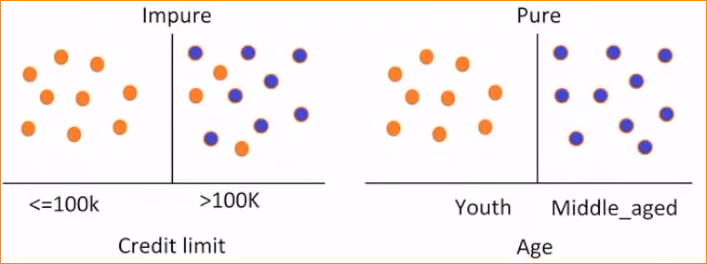


1. Gini Index (used in CART algorithm)



**Entropy**

Which test is more information to correctly classify samples? Credit limit or Age?



This entropy concept is going to consider , when you have some data how pure that data set is.

In here

Orange – not buy a mobile phone

Purple – buy a mobile phone

Now check pure or impure of the data set

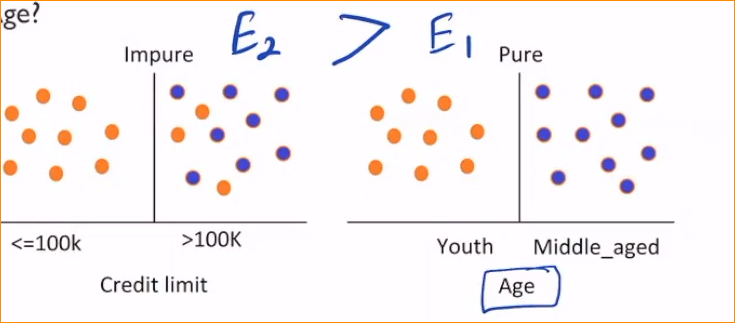
Pure – in Age category that youth people always not need to buy a mobile phone. But middle age always buy a mobile phone.

Impure – in the credit limit, in high income level (grater than 100k)not always buys a mobile phone. But low income level the don’t want to buy.

That mean entropy gives base idea about how pure or impure of our data set.

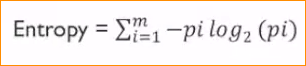
So when you have higher data set – the entropy is high.

***But when you come to decision tree. The always best prefer variable with has lower entropy.***



In here Entropy of Credit limit higher than Age. That’s why? Data is pure.

* The expected information needed to classify a given partition(D) with minimum impurity.



Where





**Information Gain –**

Amount of information gained about a random variable or signal from observing another random variable.



Get different between entropy of our entire dataset

And the weighted Avarage of the entropy of the given input variable.

Example in the Calculate entropy and build a decision tree in the next document.