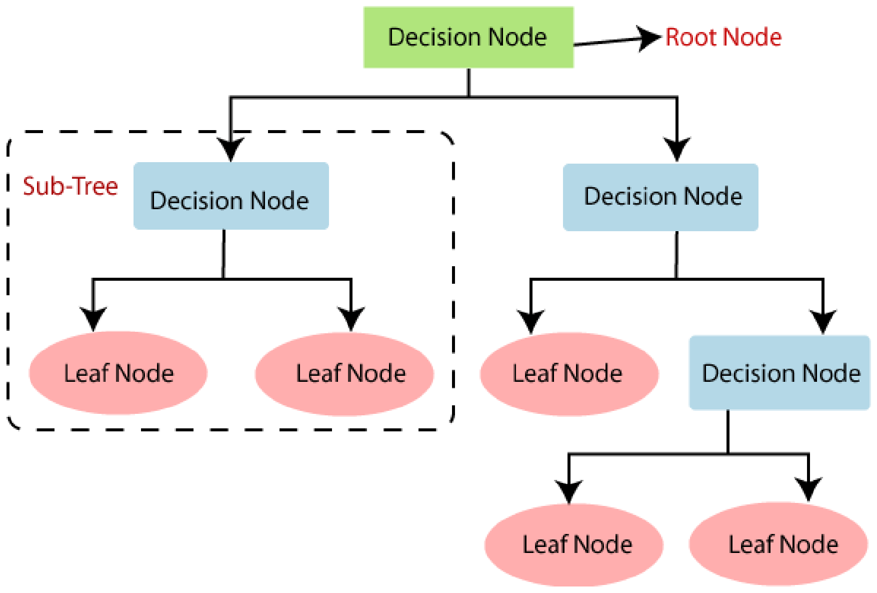
- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.



Decision Tree Terminologies

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

Example:

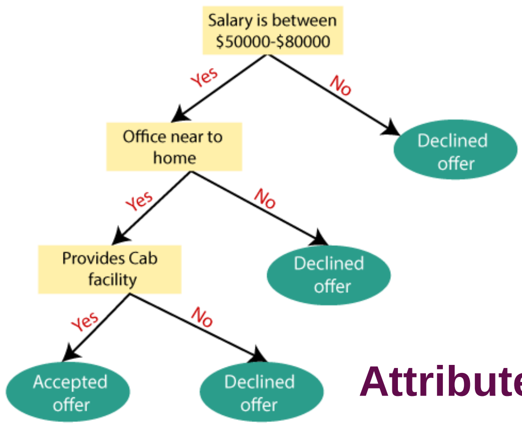
Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not.

So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM).

The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels.

The next decision node further gets split into one decision node (Cab facility) and one leaf node.

Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.**

By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**

- **Gini Index**

1. Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

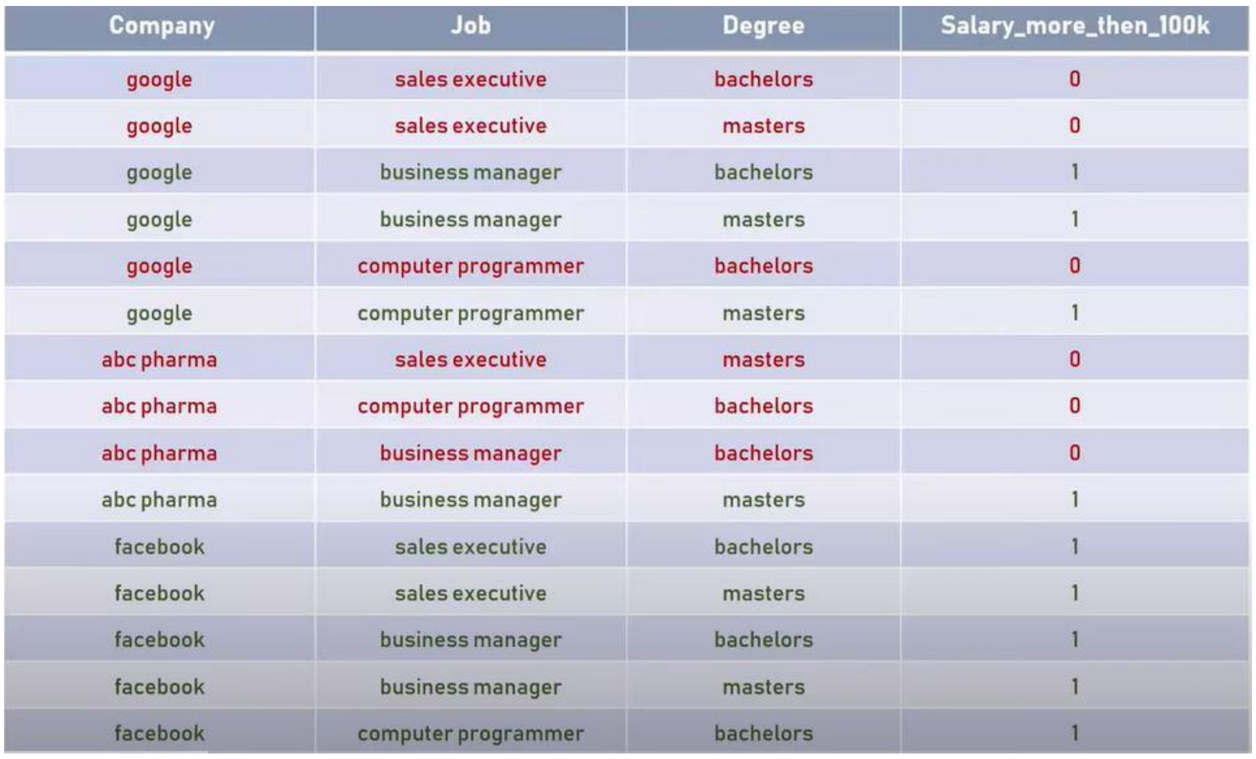
```
Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)
```

Where,

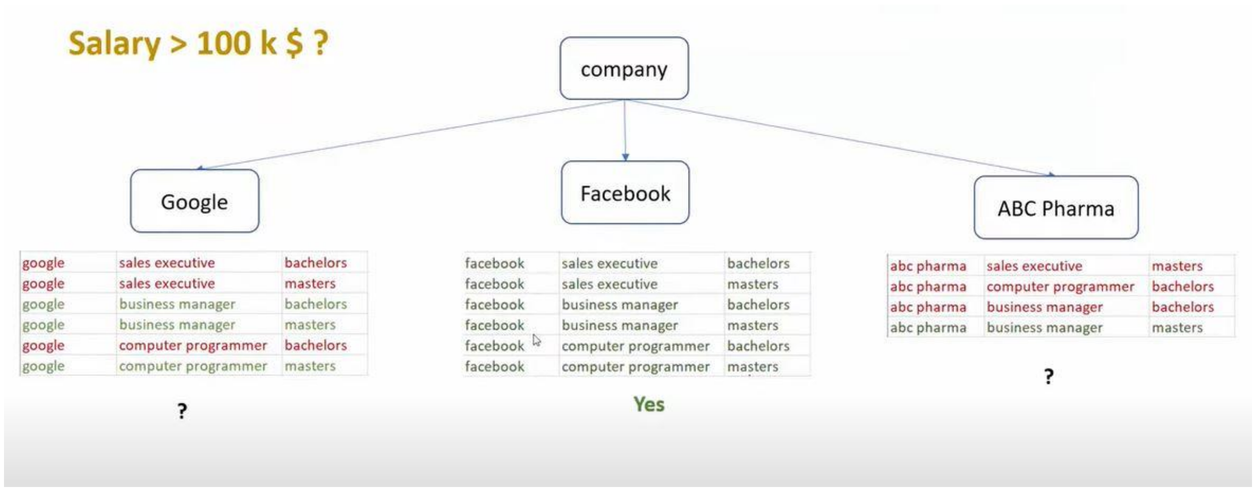
- **S= Total number of samples**
- **P(yes)= probability of yes**
- **P(no)= probability of no**

Example of Graphical Representation

Data



Company	Job	Degree	Salary_more_then_100k
google	sales executive	bachelors	0
google	sales executive	masters	0
google	business manager	bachelors	1
google	business manager	masters	1
google	computer programmer	bachelors	0
google	computer programmer	masters	1
abc pharma	sales executive	masters	0
abc pharma	computer programmer	bachelors	0
abc pharma	business manager	bachelors	0
abc pharma	business manager	masters	1
facebook	sales executive	bachelors	1
facebook	sales executive	masters	1
facebook	business manager	bachelors	1
facebook	business manager	masters	1
facebook	computer programmer	bachelors	1

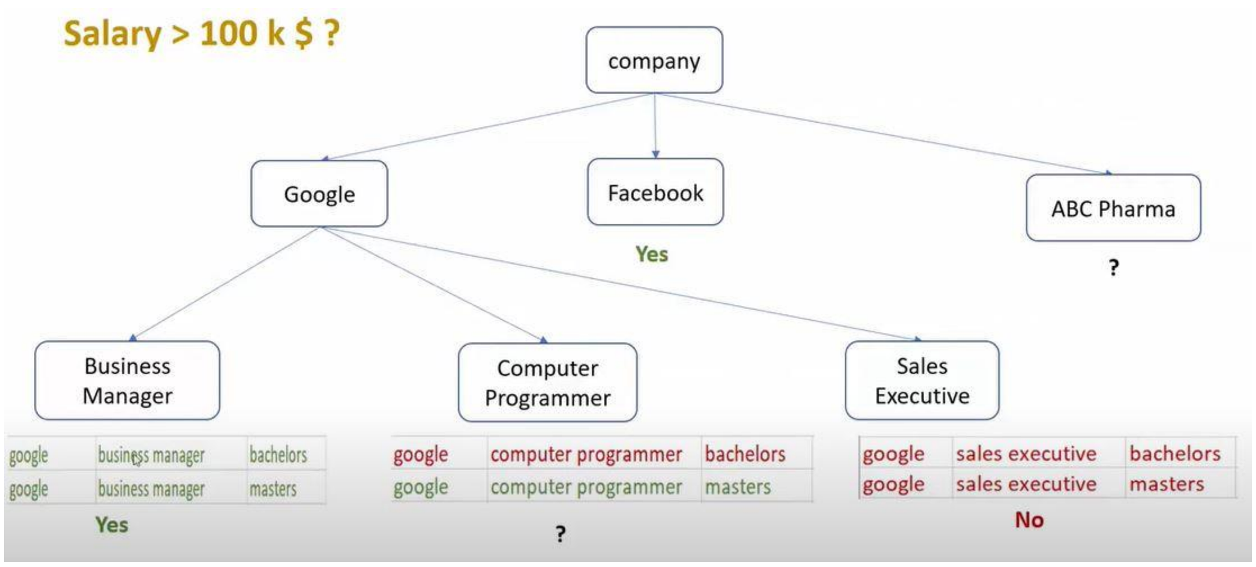


google	sales executive	bachelors
google	sales executive	masters
google	business manager	bachelors
google	business manager	masters
google	computer programmer	bachelors
google	computer programmer	masters

facebook	sales executive	bachelors
facebook	sales executive	masters
facebook	business manager	bachelors
facebook	business manager	masters
facebook	computer programmer	bachelors
facebook	computer programmer	masters

abc pharma	sales executive	masters
abc pharma	computer programmer	bachelors
abc pharma	business manager	bachelors
abc pharma	business manager	masters

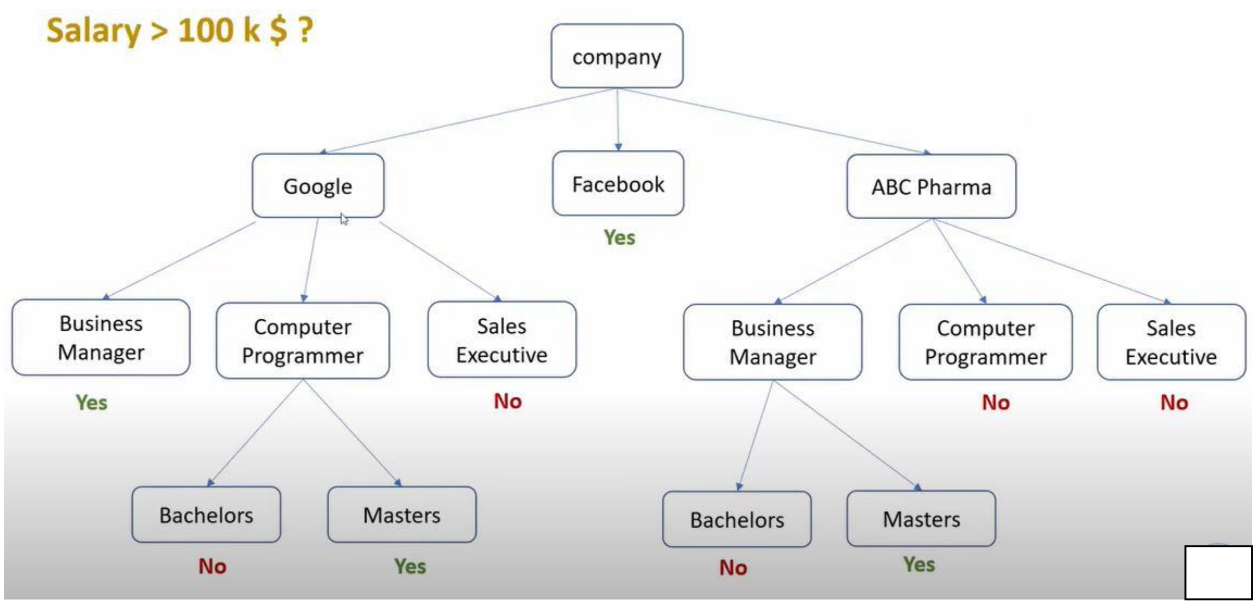
CREATIVE INSTITUTE DATA SCIENCE



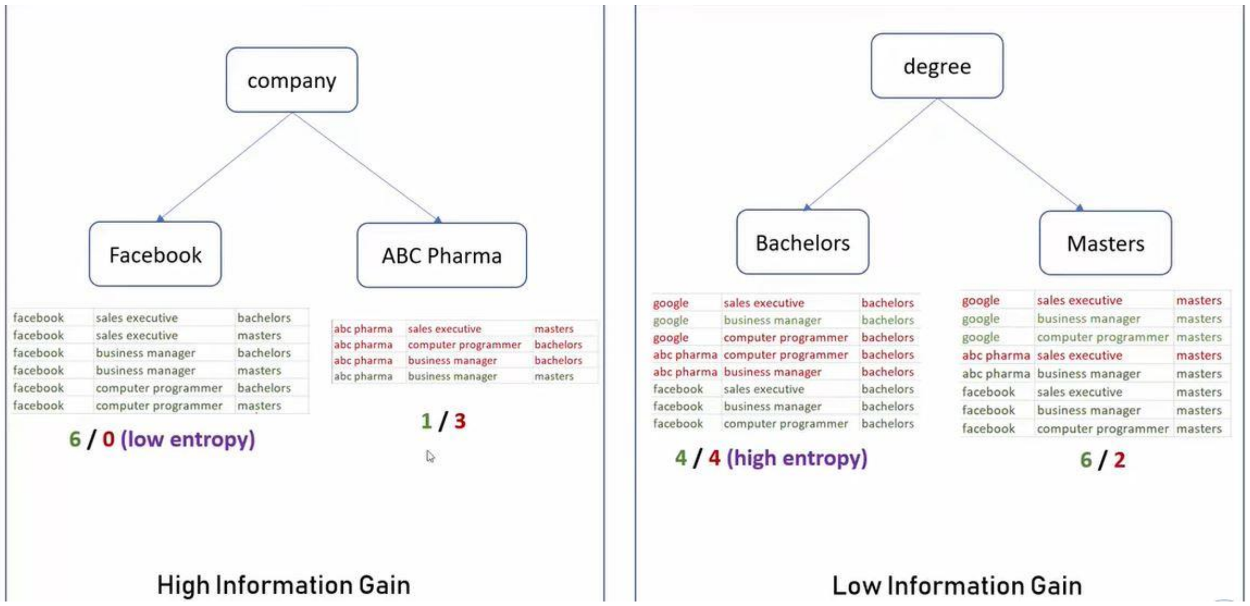
google	business manager	bachelors
google	business manager	masters

google	computer programmer	bachelors
google	computer programmer	masters

google	sales executive	bachelors
google	sales executive	masters



How do you select feature ordering?



facebook	sales executive	bachelors
facebook	sales executive	masters
facebook	business manager	bachelors
facebook	business manager	masters
facebook	computer programmer	bachelors
facebook	computer programmer	masters

6 / 0 (low entropy)

abc pharma	sales executive	masters
abc pharma	computer programmer	bachelors
abc pharma	business manager	bachelors
abc pharma	business manager	masters

1 / 3

High Information Gain

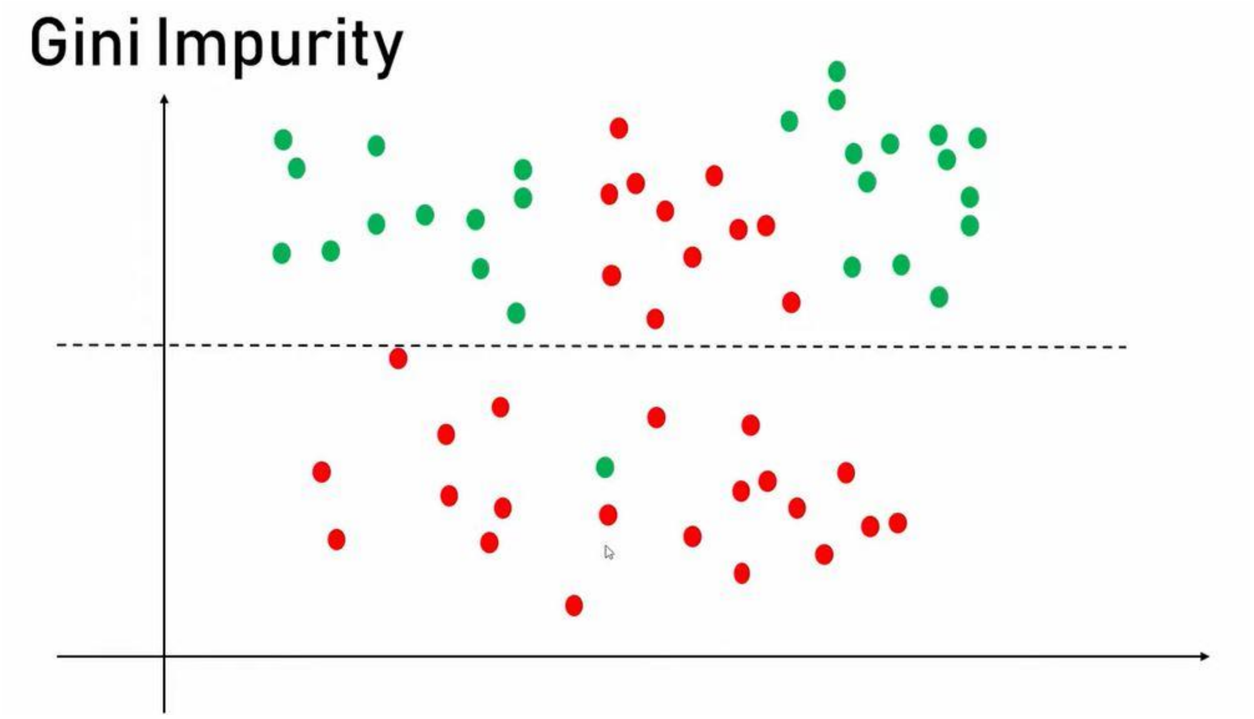
google	sales executive	bachelors
google	business manager	bachelors
google	computer programmer	bachelors
abc pharma	computer programmer	bachelors
abc pharma	business manager	bachelors
facebook	sales executive	bachelors
facebook	business manager	bachelors
facebook	computer programmer	bachelors

4 / 4 (high entropy)

google	sales executive	masters
google	business manager	masters
google	computer programmer	masters
abc pharma	sales executive	masters
abc pharma	business manager	masters
facebook	sales executive	masters
facebook	business manager	masters
facebook	computer programmer	masters

6 / 2

Low Information Gain



Decision Tree Classifier

Now import library and read the dataset

```
import pandas as pd
df = pd.read_csv("salaries.csv")
df.head()
```

now remove the target feature from the dataset

```
df2 = df.drop('salary_more_then_100k',axis='columns')
```

```
target = df['salary_more_then_100k']
```

now change the datset into label coding using scikit learn

```
from sklearn.preprocessing import LabelEncoder
le_company = LabelEncoder()
le_job = LabelEncoder()
le_degree = LabelEncoder()
```

now convert this data into labeld data

```
df2['company_n'] = le_company.fit_transform(df2['company'])
df2['job_n'] = le_job.fit_transform(df2['job'])
df2['degree_n'] = le_degree.fit_transform(df2['degree'])
```

now view the new dataset of label coding

```
df2
```

now remove the unused data from datset

```
df3 = df2.drop(['company','job','degree'],axis='columns')
df3
```

now also view the target feature

```
target
```

now build the model using scikit learn library of tree

```
from sklearn import tree
model = tree.DecisionTreeClassifier()
```

now fit the model

```
model.fit(df3, target)
```

now find the score of this model

```
model.score(df3,target)
```

now ansert the question based on model

- Is salary of Google, Computer Engineer, Bachelors degree > 100 k ?

```
model.predict([[2,1,0]])
```

- Is salary of Google, Computer Engineer, Masters degree > 100 k ?

```
model.predict([[2,1,1]])
```

Excericise

Build decision tree model to predict survival based on certain parameters

this file using following columns build a model to predict if person would survive or not,

1. Pclass
2. Gender
3. Age
4. Fare

Calculate score of your model