MODULE-II

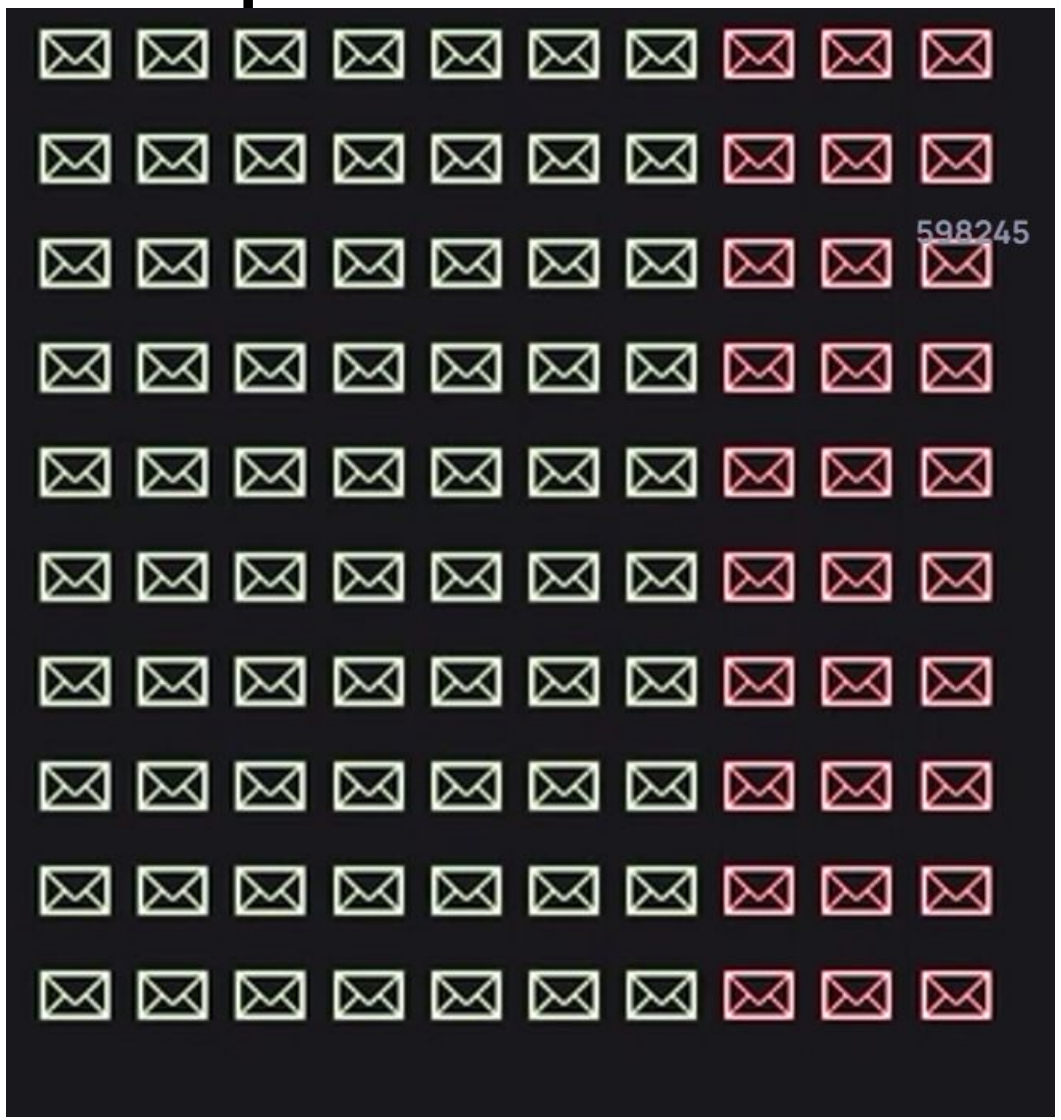
Understanding Naïve Bayes

- Naïve Bayes algorithm is a supervised learning algorithm based on the **Bayes theorem**. It is used for solving classification problems.
- **It is a probabilistic classifier, which means it predicts based on the probability of an object**.

Example: Email spam classification problem

- Often we see the following words in spam mails
  - Free
  - Lottery
  - Discount
  - Hurry up

Sample dataset

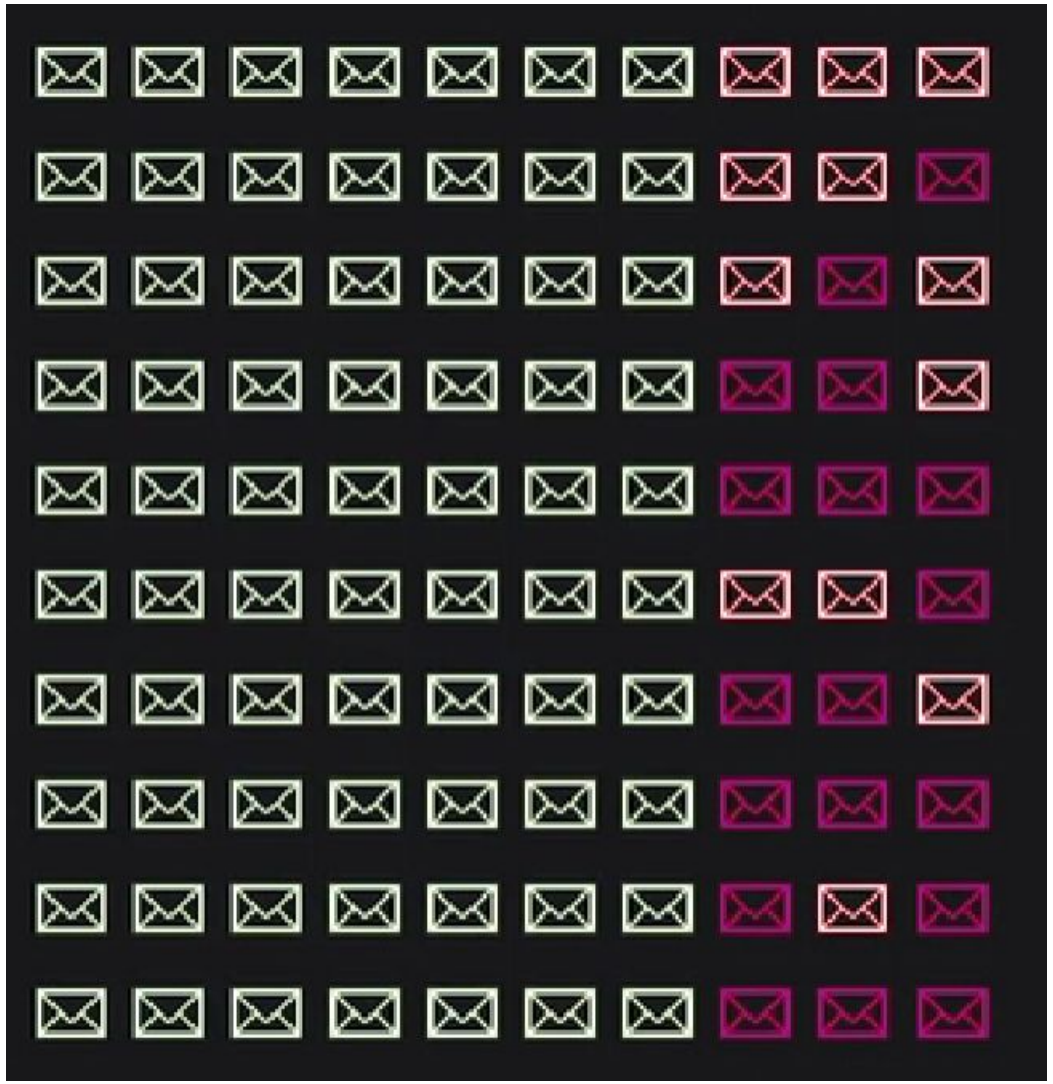


100 emails are there

30 emails are spam

P(spam)=30%

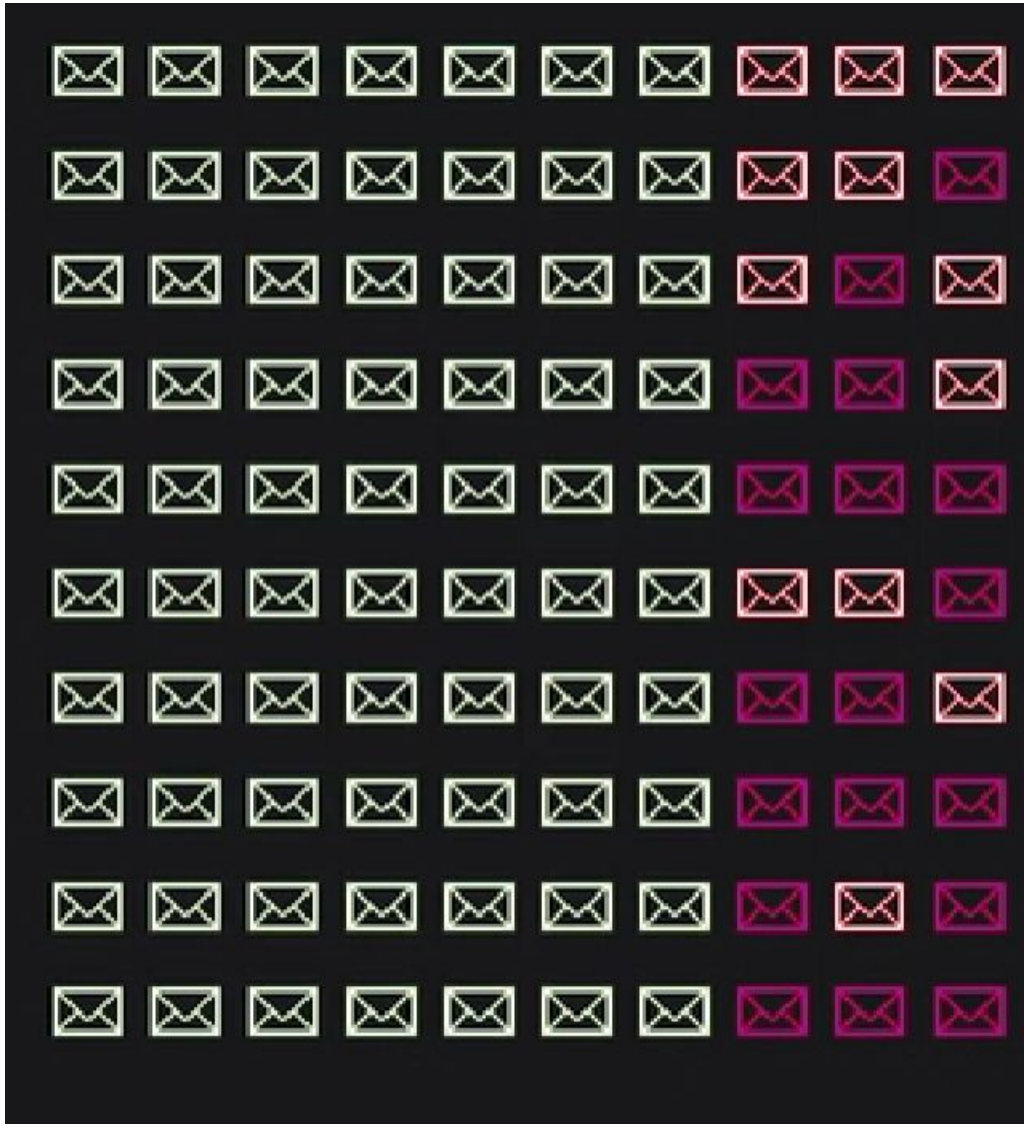
What is the probability that the email is spam and contains the word ‘free’?



$$p(free \mid spam) = 40\%$$

12/30=40%

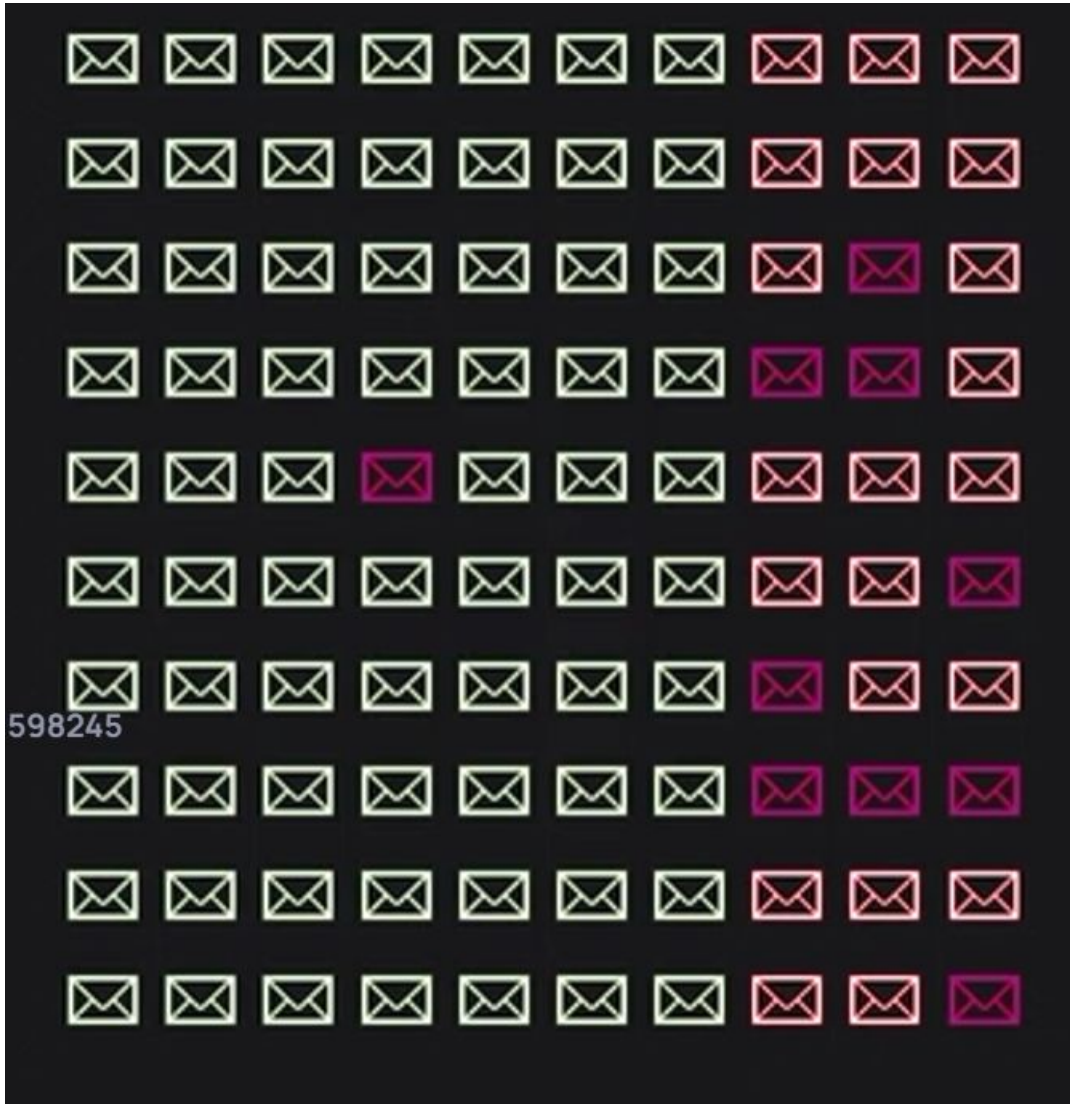
Second event ‘spam’ is already occurred. We are finding the probability of occurring ‘free’ word in spam mails.



$$p(lottery \mid spam) = 60\%$$

The probability that the email is spam and contains the word "lottery"

18/30=60%



$p(free\_lottery) = 10\%$

The probability of an email containing the words "free" and "lottery"

10/100=10%

What is the probability that an email is spam given that it contains the word free and lottery?

$p(spam) = 30\%$

$p(free\_lottery) = 10\%$

598245

$p(free \mid spam) = 40\%$

$p(lottery \mid spam) = 60\%$

$p(free\_lottery \mid spam) = 0.4 * 0.6$

$= .24 = 24\%$

$$p(spam|free\_lottery) = \frac{p(free\_lottery|spam) * p(spam)}{p(free\_lottery)}$$

$$= \frac{0.24 * 0.3}{0.1}$$

$= 0.72$

$= 72\ \%$

Naïve Bayes Theorem

It is called Naïve Bayes because it makes a naïve assumption that all features (such as $p(free)$ or $p(lottery)$) are independent of each other

Why is it called Naïve Bayes?

The Naïve Bayes algorithm comprised of **two words Naïve and Bayes**, which can be described as:

- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

i.e. changing the value of one feature, does not directly influence or change the value of any of the other features used in the algorithm.

Example: If the fruit is identified based on color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple.

Hence each feature individually contributes to identifying that it is an apple without depending on each other.

- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes Theorem

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**,
- It is used to **determine the probability of a hypothesis with prior knowledge.**
- It depends on the **conditional probability.**
- The formula for Bayes' theorem is given as:

$$P(A/B)=\frac{P(B/A)*P(A)}{P(B)}$$

- P(A | B): Probability of event A occurring given that B is true (posterior probability).
- P(B | A) : Probability of event B occurring given that A is true (likelihood).
- P(A): Probability of event A occurring on its own (prior probability).
- P(B): Probability of event B occurring on its own (normalizing constant).

Naive Bayes is a **classifier**. That means:

- You're given a set of features (like words in an email).
- You want to predict the **class** (like "spam" or "not spam").
- So the goal is to compute:
- **P(Class | Features)** — the probability of a class given the observed features.
- **P(B|A)** tells us how likely the features are given a class but it doesn't help us directly decide which class is most likely for a new observation.
- We need **P(A|B)** to make that decision.

Consider a set of patients coming for treatment in a certain clinic.
Let A denote the event that a “patient has a liver disease” and B the event that the “patient is an alcoholic”. It is known from experience that 10% of the patients entering the clinic have liver diseases and 5% of the patients are alcoholic. Also among those patients diagnosed with liver diseases, 7% are alcoholics. Given that patient is an alcoholic, what is the probability that he will have liver diseases?

Solutions
A –the patient has a liver disease
B – the patient is an alcoholic
P(A)=10%=0.10
P(B)=5%=0.05
P(B|A)=7% =0.07
P(A/B)=(P(B|A).P(A))/P(B)
(0.07*0.1)/0.05=0.14

Advantage	Description
Fast and Efficient	Training and prediction are quick—even with large datasets.
Simple to Implement	Easy to understand and code
Works Well with Categorical Data	Perfect for datasets with discrete features.
Performs Well with Small Data	Doesn’t need massive datasets to make accurate predictions.
Handles Missing Values	Can skip missing features during prediction without major issues.
Robust to Irrelevant Features	Unrelated features don’t heavily impact performance.

Disadvantages

- **Strong Independence Assumption** It assumes all features are independent given the class, which is rarely true in real-world data.
- **Poor Performance with Correlated Features** If features are related (e.g., age and income), Naive Bayes may miscalculate probabilities.
- **Zero-Frequency Problem** If a feature value wasn’t seen in training, it assigns zero probability

Using these probabilities estimate the probability values for the new instance – (Color=Green, legs=2, Height=Tall, and Smelly=No).

No	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

New Instance

(Color=Green, legs=2, Height=Tall, and Smelly=No)

To estimate the probability values for the **new instance**
(Color=Green, Legs=2, Height=Tall, Smelly=No)
using the Naive Bayes classifier,we can follow these steps:

1.Calculate Prior Probabilities:

Calculate the prior probability for each species, P(M) and P(H).

$$P(M) = \frac{4}{8} = 0.5 \quad P(H) = \frac{4}{8} = 0.5$$

2. Calculate Likelihood Probabilities:

Calculate the likelihood of each feature given the species,
P(Color=White|M) P(Color=White|H), P(Color=Green|M) P(Color=Green|H)
P(Legs=2|M), P(Legs=2|H), P(Legs=3|M), P(Legs=3|H),
P(Height=Tall|M), P(Height=Tall|H), P(Height=Short|M), P(Height=Short|H),
P(Smelly=Yes|M), P(Smelly=Yes|H), P(Smelly=No|M),P(Smelly=No|H),

COLOR	M	H
White	2/4	¾
Green	2/4	¼

LEGS	M	H
2	1/4	4/4
3	3/4	0/4

HEIGHT	M	H
Short	3/4	2/4
Tall	1/4	2/4

SMELLY	M	H
YES	¾	¼
No	1/4	3/4

3. Compute Posterior Probabilities

$$P(M) = \frac{4}{8} = 0.5 \quad P(H) = \frac{4}{8} = 0.5$$

COLOR	M	H
White	2/4	¾
Green	2/4	¼

LEGS	M	H
2	1/4	4/4
3	3/4	0/4

HEIGHT	M	H
Short	3/4	2/4
Tall	1/4	2/4

SMELLY	M	H
YES	¾	¼
No	1/4	3/4

$$p(M|New\ Instance) = p(M) * p(Color = Green|M) * p(Legs = 2|M) * p(Height = tall|M) * p(Smelly = no\ |M)$$

==

=0.5*(2/4)*(1/4)*(1/4)*(1/4)=1/256=0.00390625

$$p(H|New\ Instance) = p(H) * p(Color = Green|H) * p(Legs = 2|H) * p(Height = tall|H) * p(Smelly = no\ |H)$$

$$p(H|New\ Instance) = 0.5 * \frac{1}{4} * \frac{4}{4} * \frac{2}{4} * \frac{3}{4} = 0.047$$

$p(H|New\ Instance) > p(M|New\ Instance)$

Hence the new instance belongs to Speices H

4. Conclusion

Since P(H|New Instance) is higher than P(M|New Instance), the Naive Bayes classifier predicts that the new instance belongs to species

Exercise

Given the following dataset of customer information, use the Naive Bayes classifier to predict if a new customer will buy a product:

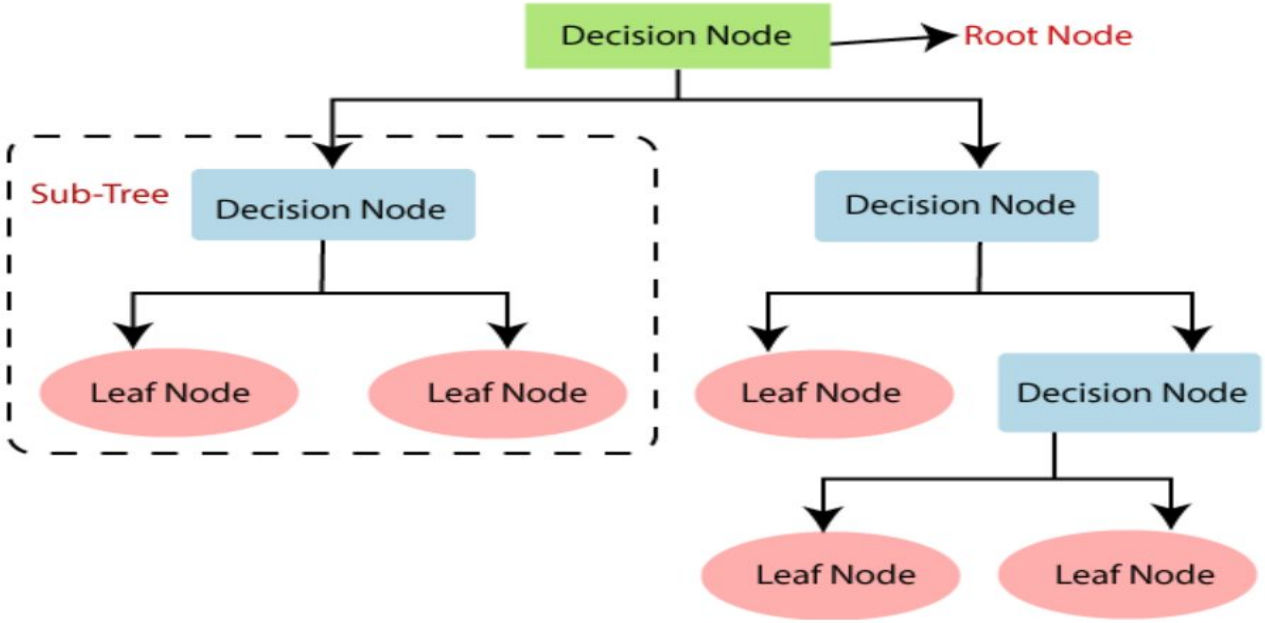
- **New Customer[**Age: Middle-aged, Income: Medium, Student Status: Yes]

Customer ID	Age	Income	Student Status	Bought Product
1	Young	High	Yes	No
2	Young	Medium	Yes	Yes
3	Middle-aged	High	No	Yes
4	Senior	Medium	Yes	No
5	Senior	Low	No	No
6	Middle-aged	Low	Yes	Yes
7	Young	Low	No	No
8	Young	Medium	No	Yes
9	Senior	High	Yes	Yes
10	Middle-aged	Medium	No	No

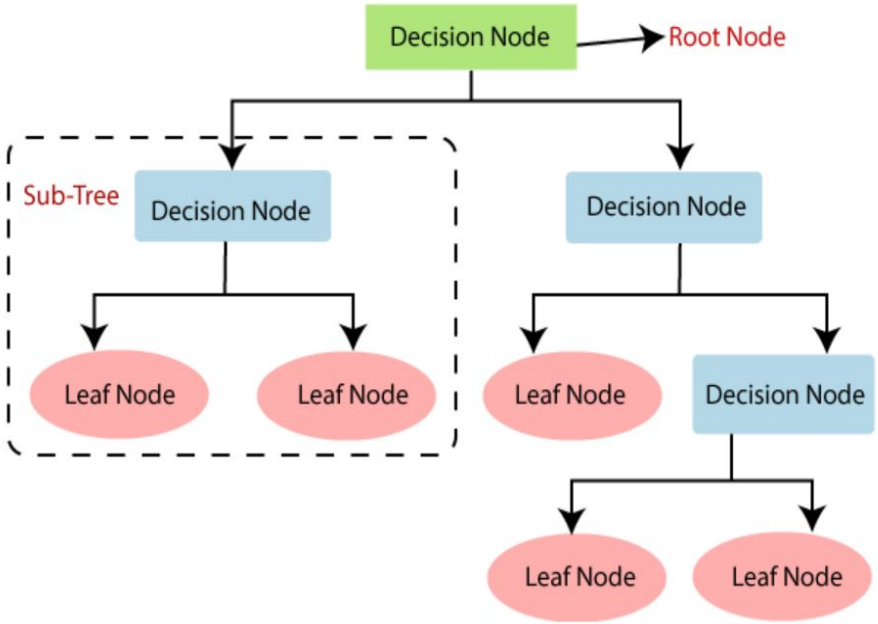
DECISION TREES

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 anv further branches.



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.

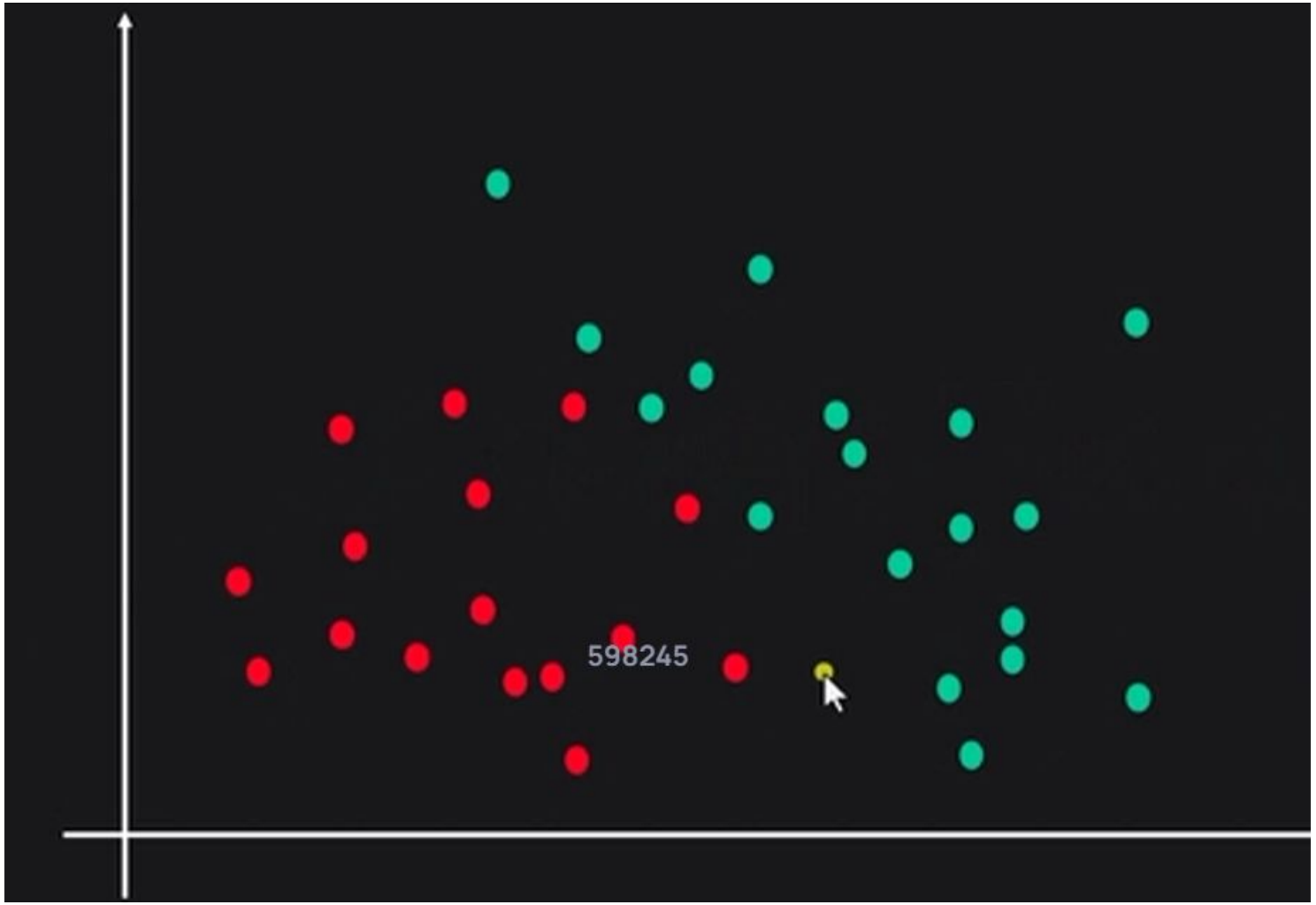
•**Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

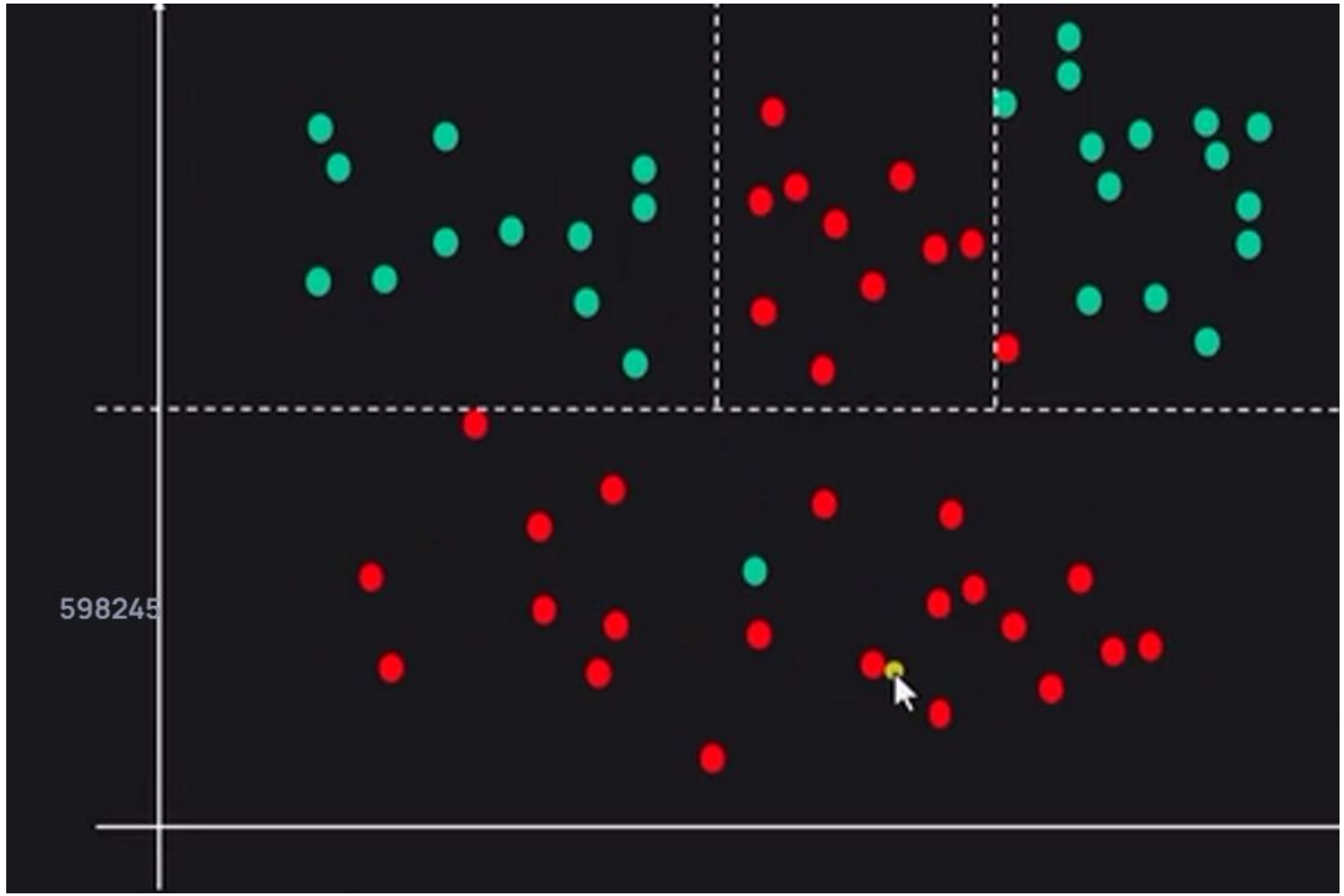
•**Branch/Sub Tree:** A tree formed by splitting the tree.

•**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.

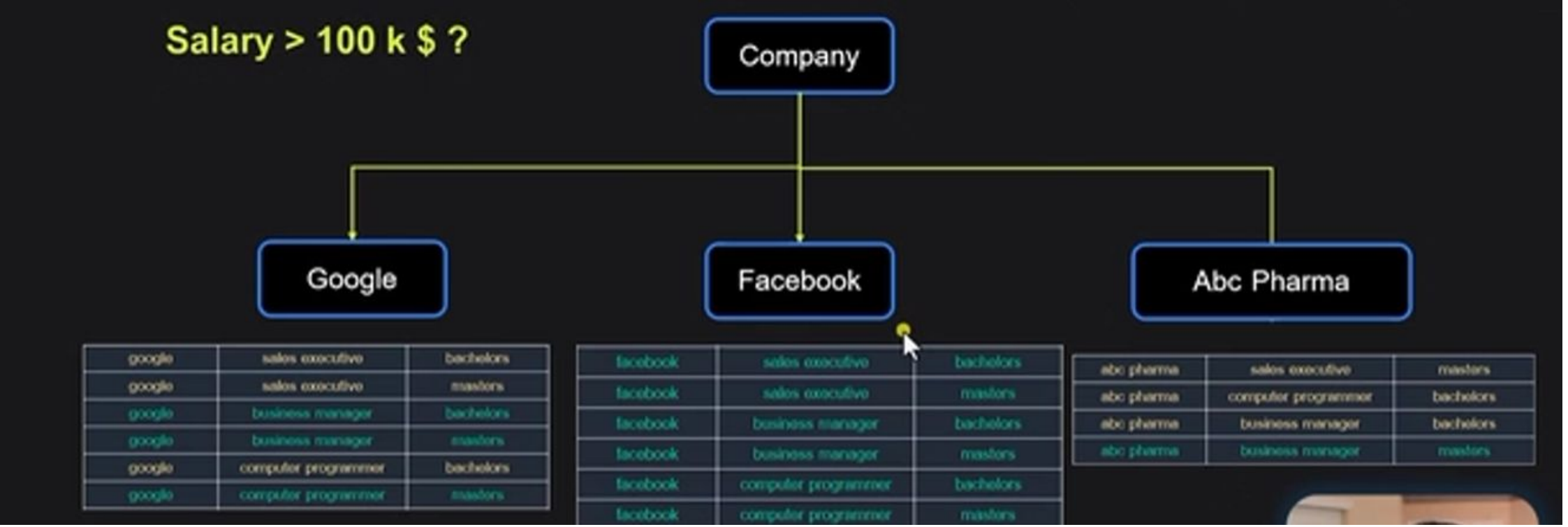
Example

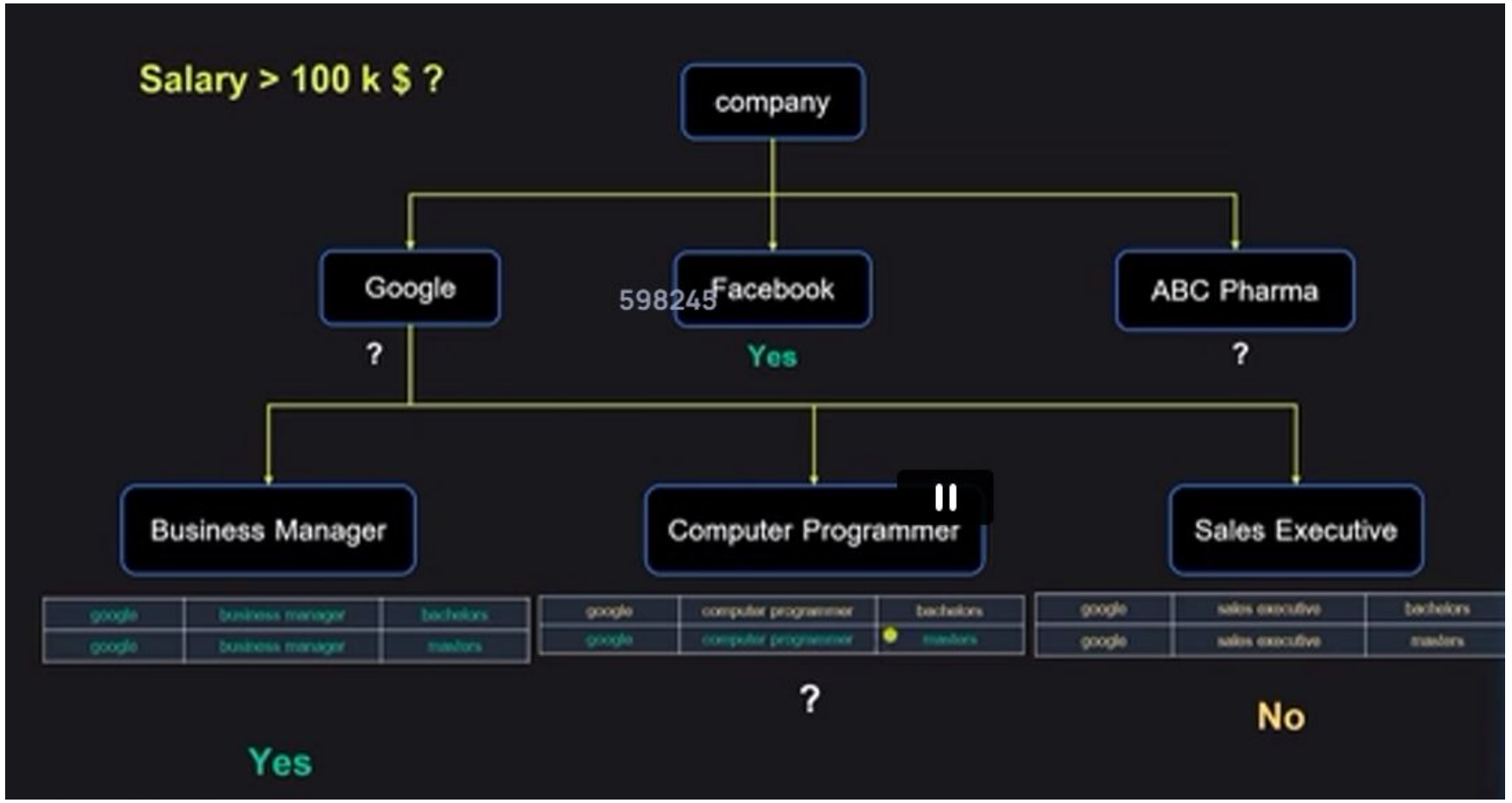



598245

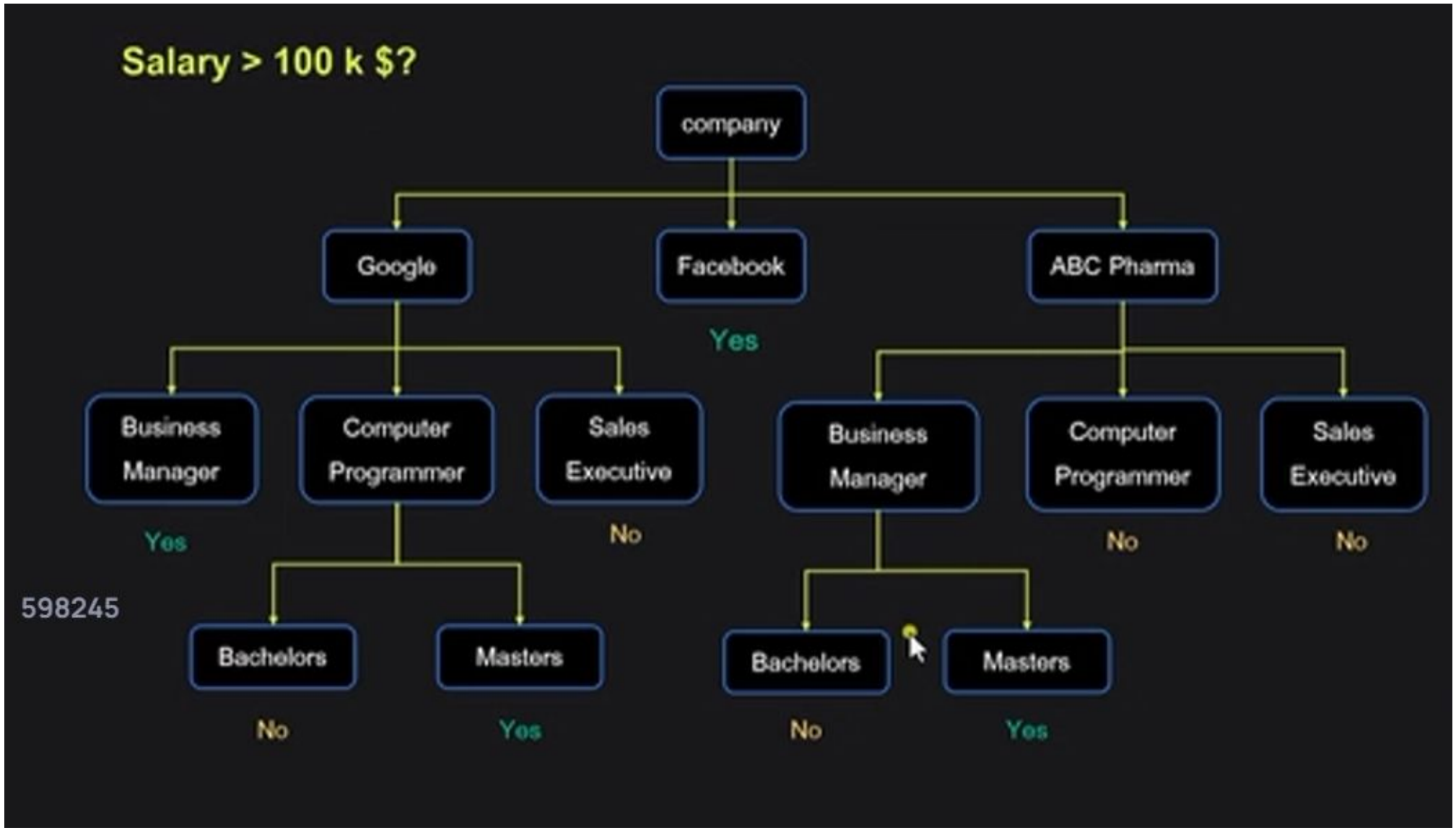
If Salary more than 100k or not? Build a Decision tree

Company	Job	Degree	Salary_more_th
google	sales executive	bachelors	0
google	sales executive	masters	0
google	business manager	bachelors	1
google	business manager	masters	598245 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
facebook	computer programmer	masters	1


Salary > 100 k $?
Company
Google
Facebook
Abc Pharma
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

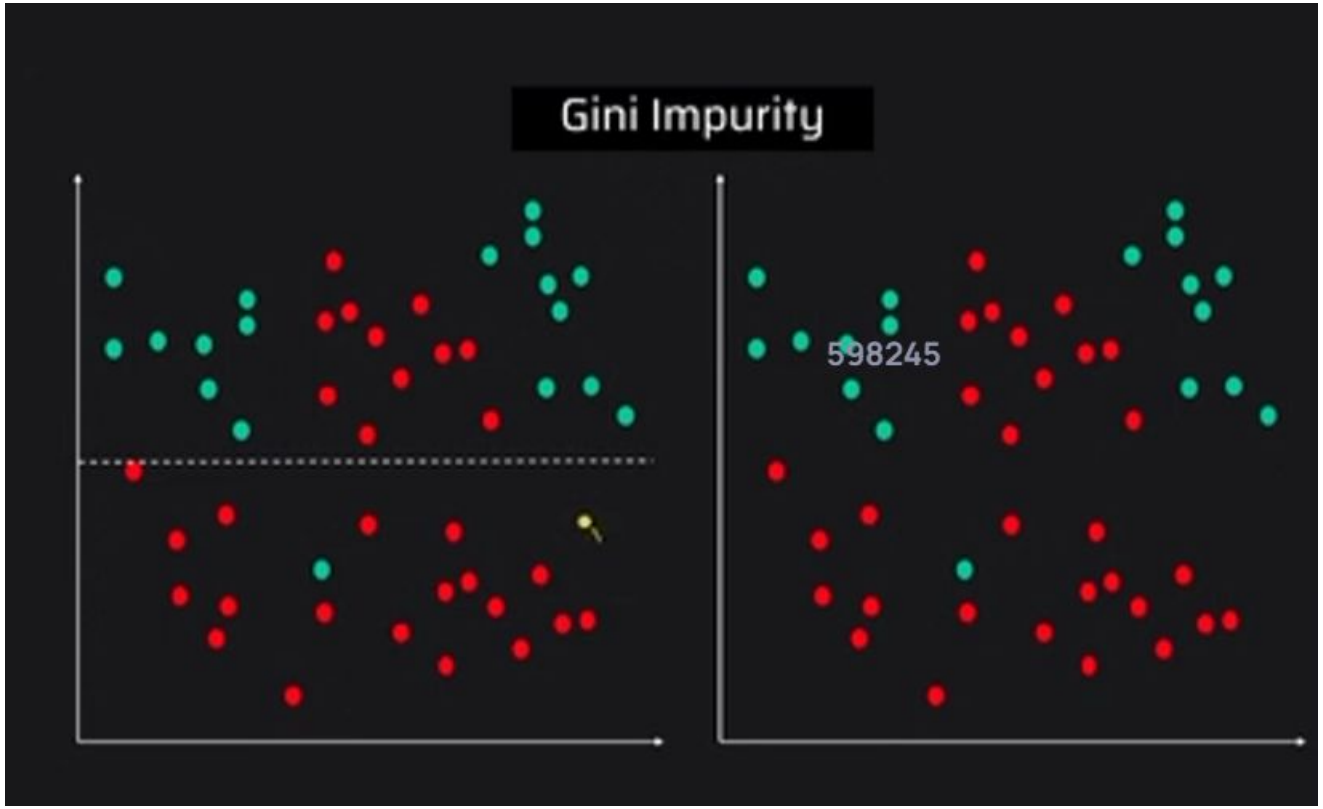


For Facebook, all employees got a salary greater than 100k
For Google and Abc Pharma we need to check remaining features
If business manager in Google, YES
If computer programmer , need to check the degree
If masters degree means YES else NO


Salary > 100 k $?
company
Google
Facebook
ABC Pharma
Yes
Business Manager
Computer Programmer
Sales Executive
Yes
No
Business Manager
Computer Programmer
Sales Executive
No
No
598245
Bachelors
Masters
No
Yes
Bachelors
Masters
No
Yes

Gini impurity

- Gini impurity is a key concept used in decision tree algorithms to measure how “mixed” the classes are in a node. It helps the algorithm decide the best feature to split the data on.



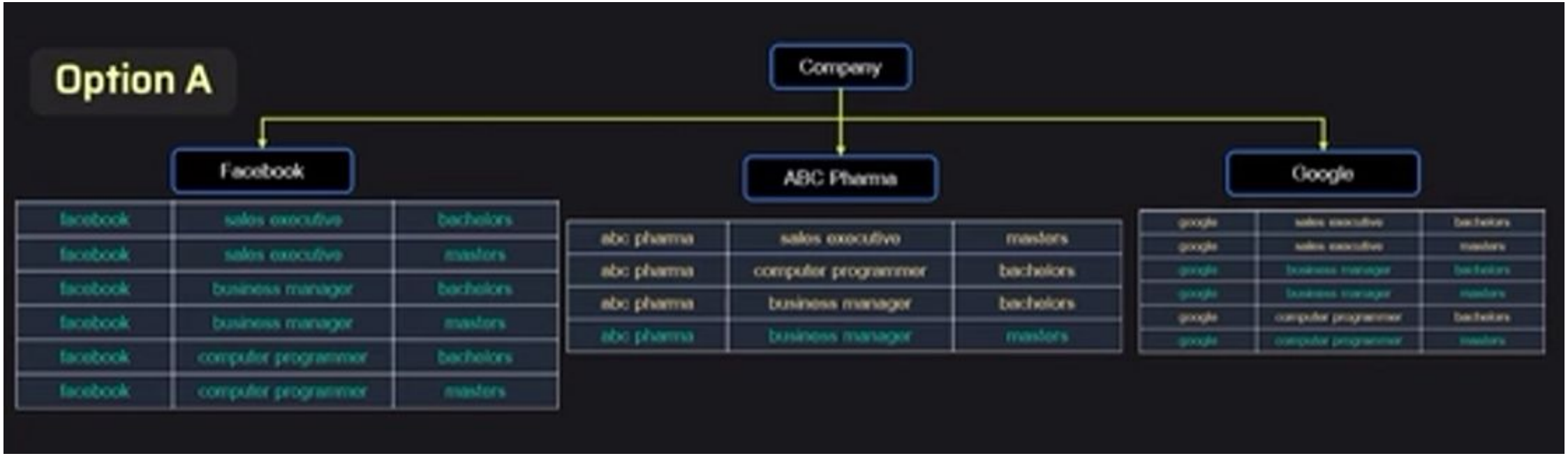
Impurity=40%

Impurity=2%

Impurity<45%
Impurity<52%
the impuriy of second split is more compare with the impurity of first split
So we will select the first split

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
facebook	computer programmer	masters	1

Option A: start with company



Option B: start with degree


Option A
Company
Facebook
ABC Pharma
Google
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
abc pharma sales executive masters
abc pharma computer programmer bachelors
abc pharma business manager bachelors
abc pharma business manager masters
1 / 3
google sales executive bachelors
google sales executive masters
google business manager bachelors
google business manager masters
google computer programmer bachelors
google computer programmer masters
3 / 3
Option B
Degree
598245
Bachelors
Masters
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
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
6 / 2


Option A
Company
Facebook
ABC Pharma
Google
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
abc pharma | masters
abc pharma | computer programmer | bachelors
abc pharma | business manager | bachelors
abc pharma | business manager | masters
1 / 3
google | sales executive | bachelors
google | sales executive | masters
google | business manager | bachelors
google | business manager | masters
google | computer programmer | bachelors
google | computer programmer | masters
3 / 3
$p(\text{salary} > 100k) = {}^{6}/_{6} = 1$
$p(\text{salary} \leq 100k) = {}^{0}/_{6} = 0$
$Gini = 1 - (1^2 + 0^2)$
$= 0$
$p(\text{salary} > 100k) = {}^{1}/_{4} = 0.25$
$p(\text{salary} \leq 100k) = {}^{3}/_{4} = 0.75$
$Gini = 1 - (0.25^2 + 0.75^2)$
$= 0.375$
$p(\text{salary} > 100k) = {}^{3}/_{6} = 0.5$
$p(\text{salary} \leq 100k) = {}^{3}/_{6} = 0.5$
$Gini = 1 - (0.5^2 + 0.5^2)$
$= 0.5$
$Total\ Impurity = ({}^{6}/_{16} \times 0) + ({}^{4}/_{16} * 0.375) + ({}^{6}/_{16} * 0.5) = 0.28$

598245

Option B

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

facebook	computer programmer	masters
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

6 / 2

$$p(\text{salary} > 100k) = {}^{4}/_{8} = 0.5$$

$$p(\text{salary} \leq 100k) = {}^{4}/_{8} = 0.5$$

$$Gini = 1 - (0.5^2 + 0.5^2)$$

$$= 0.5$$

$$p(\text{salary} > 100k) = {}^{6}/_{8} = 0.75$$

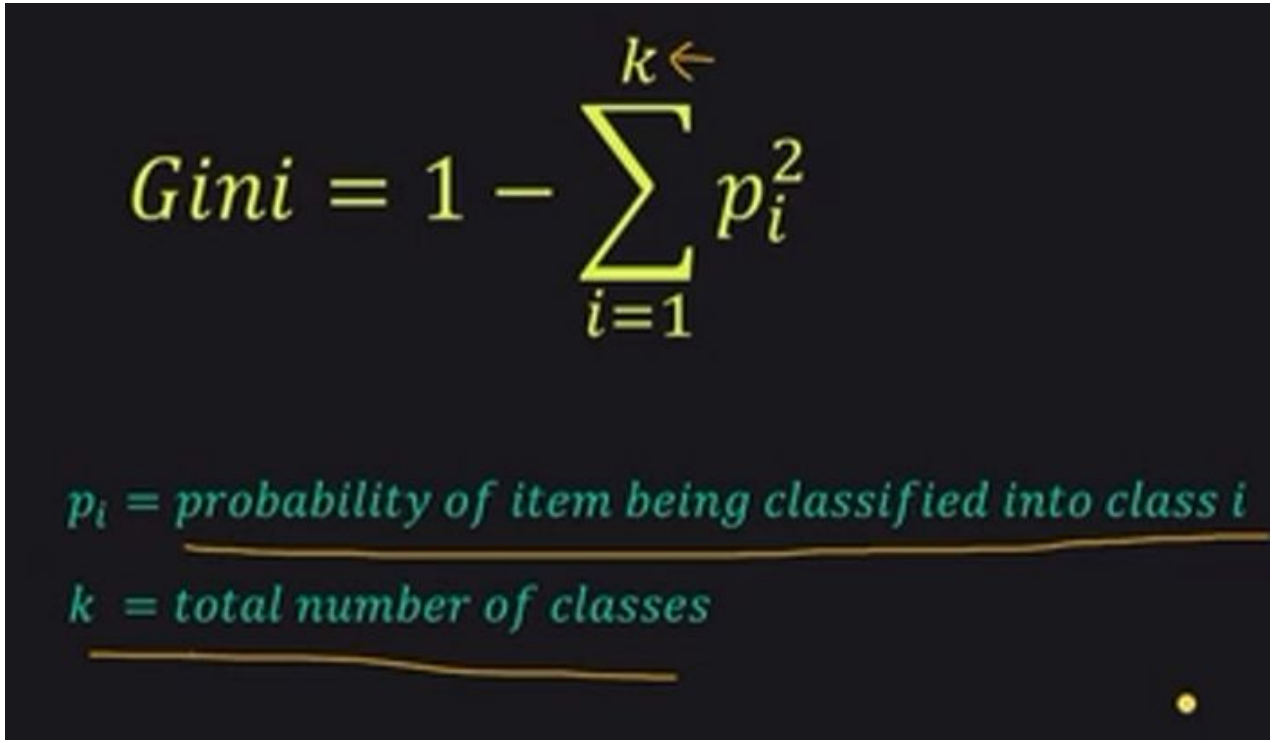
$$p(\text{salary} \leq 100k) = {}^{2}/_{8} = 0.25$$

$$Gini = 1 - (0.25^2 + 0.75^2)$$

$$= 0.375$$

$$Total\ Impurity = \left({}^{8}/_{16} \times 0.5\right) + \left({}^{8}/_{16} * 0.375\right) = 0.4375$$

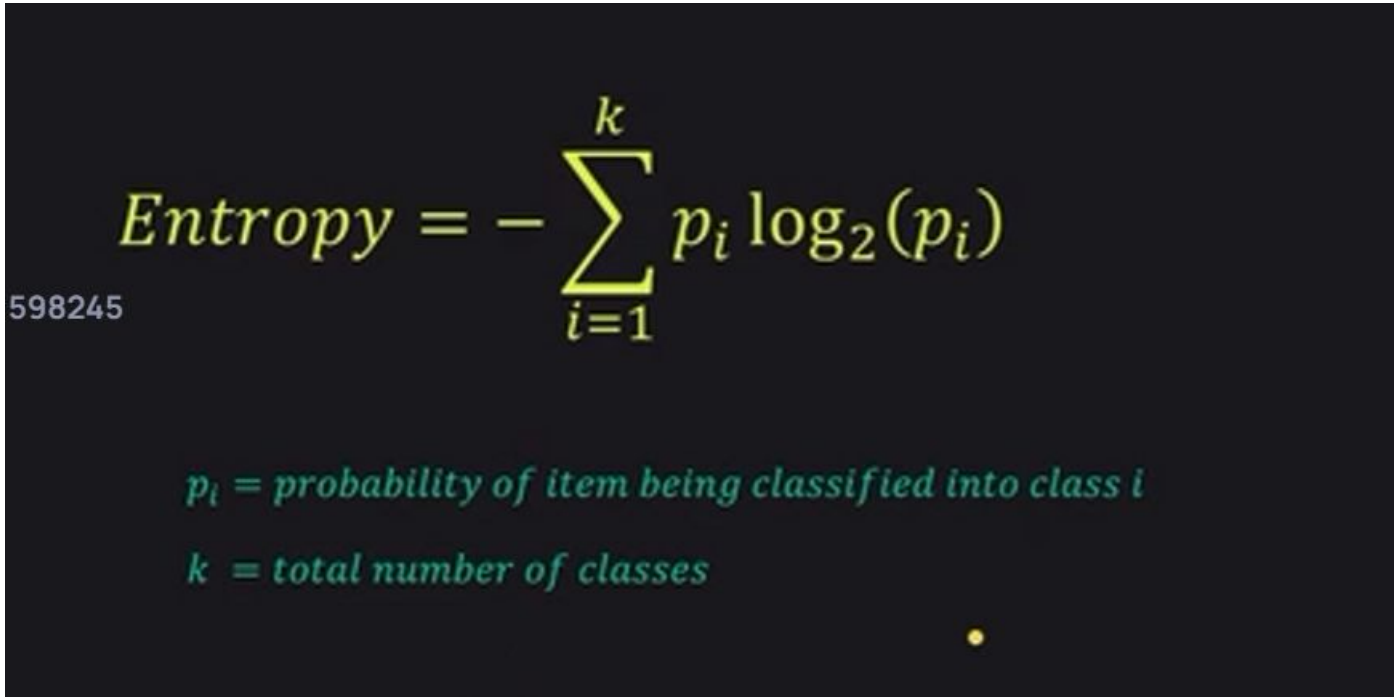
Gini impurity

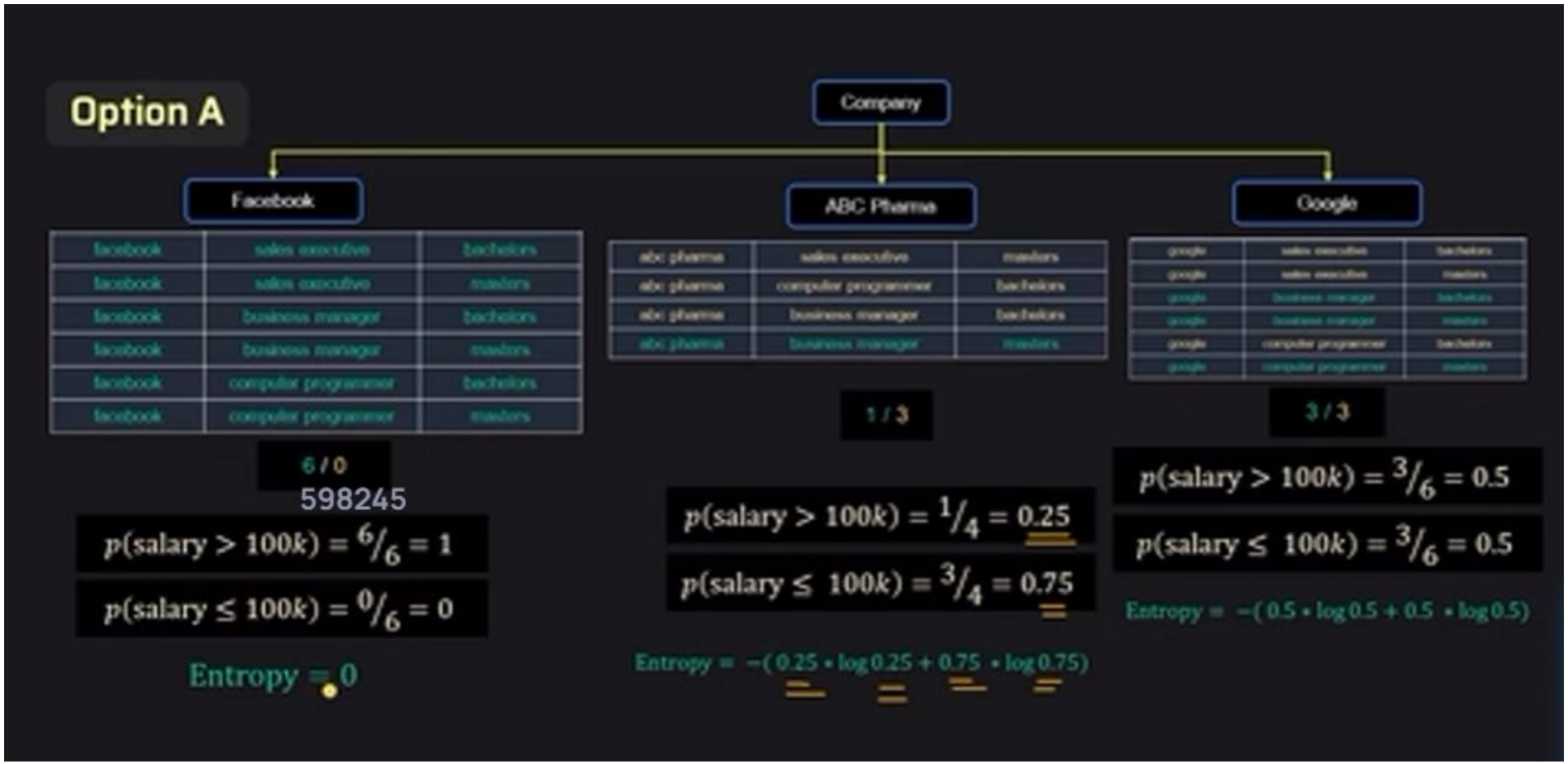


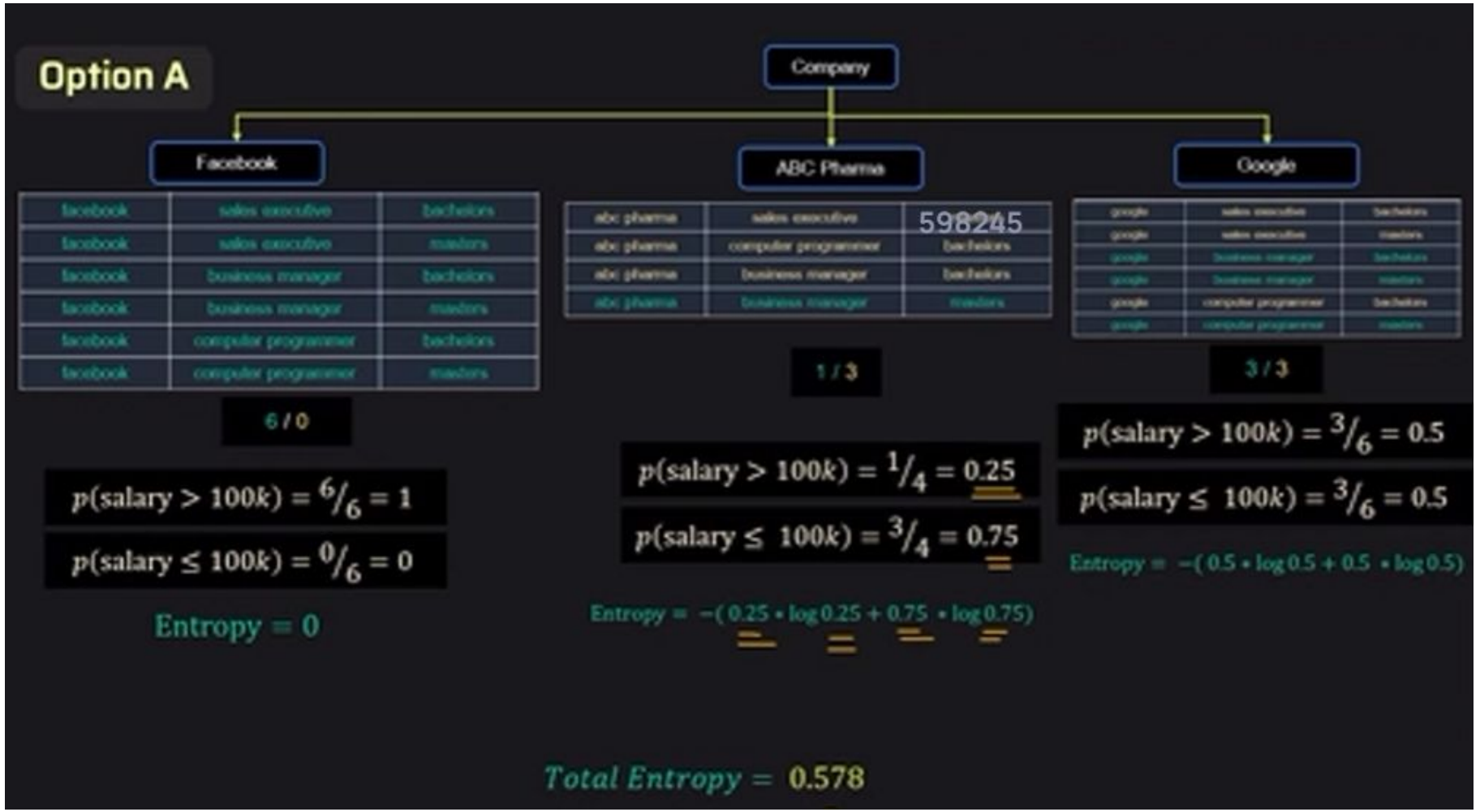
Option A is selected since it has got less impurity
Formula=1-individual probability

Entropy/Information Gain

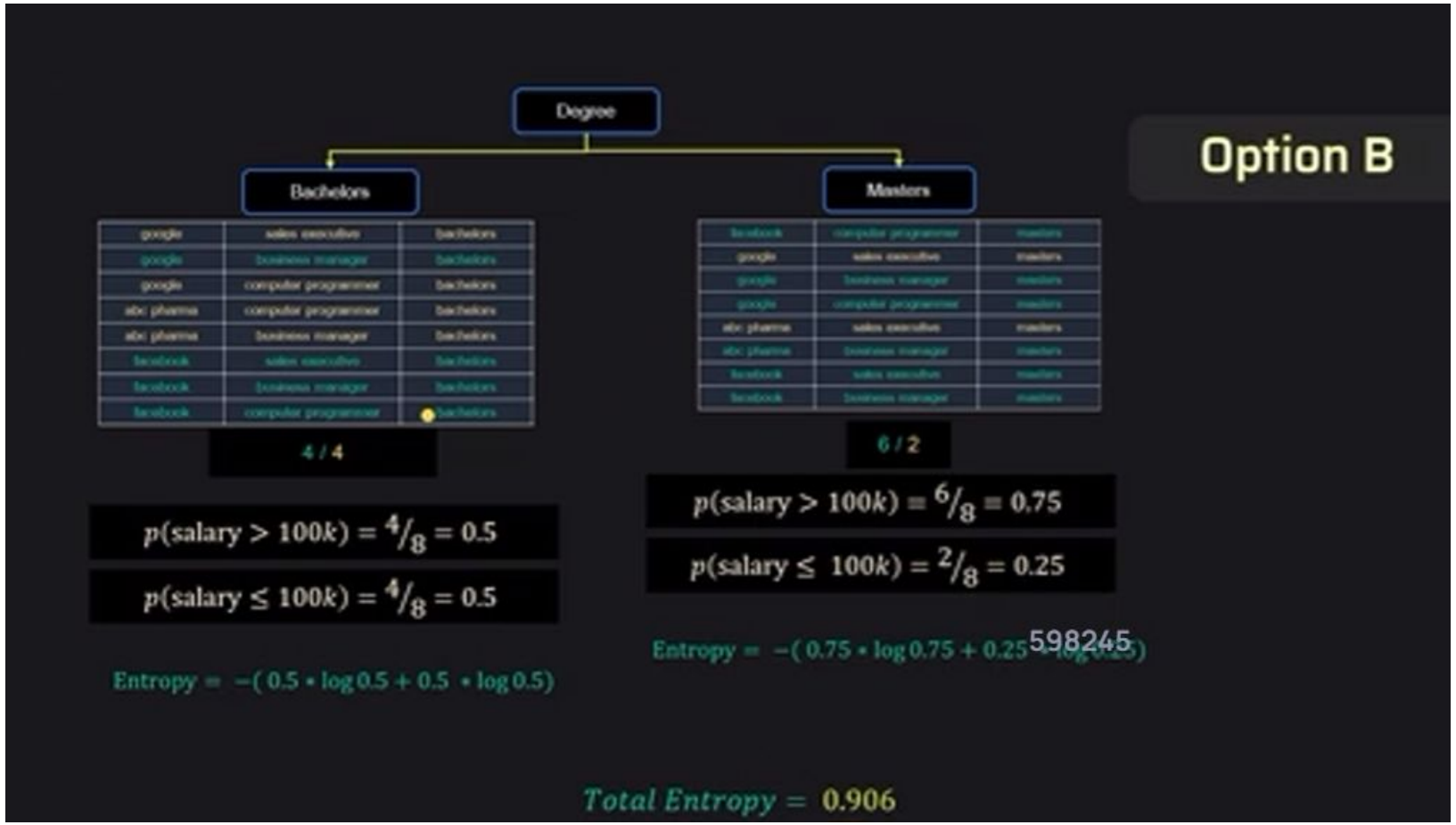
- Similar to Gini impurity
- Find the split where you get more pure dataset




Option A
Company
Facebook
ABC Pharma
Google
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
google sales executive bachelors
google sales executive masters
google business manager bachelors
google business manager masters
google computer programmer bachelors
google computer programmer masters
6 / 0
1 / 3
3 / 3
598245
$p(\text{salary} > 100k) = {}^{6}/_{6} = 1$
$p(\text{salary} \leq 100k) = {}^{0}/_{6} = 0$
Entropy = 0
$p(\text{salary} > 100k) = {}^{1}/_{4} = 0.25$
$p(\text{salary} \leq 100k) = {}^{3}/_{4} = 0.75$
Entropy = $-(0.25 * \log 0.25 + 0.75 * \log 0.75)$
$p(\text{salary} > 100k) = {}^{3}/_{6} = 0.5$
$p(\text{salary} \leq 100k) = {}^{3}/_{6} = 0.5$
Entropy = $-(0.5 * \log 0.5 + 0.5 * \log 0.5)$


Option A
Company
Facebook
ABC Pharma
Google
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
$p(\text{salary} > 100k) = {}^{6}/_{6} = 1$
$p(\text{salary} \leq 100k) = {}^{0}/_{6} = 0$
Entropy = 0
abc pharma | sales executive | bachelors
abc pharma | computer programmer | bachelors
abc pharma | business manager | bachelors
abc pharma | business manager | masters
598245
1 / 3
$p(\text{salary} > 100k) = {}^{1}/_{4} = 0.25$
$p(\text{salary} \leq 100k) = {}^{3}/_{4} = 0.75$
Entropy = $-(0.25 * \log 0.25 + 0.75 * \log 0.75)$
google | sales executive | bachelors
google | sales executive | masters
google | business manager | bachelors
google | business manager | masters
google | computer programmer | bachelors
google | computer programmer | masters
3 / 3
$p(\text{salary} > 100k) = {}^{3}/_{6} = 0.5$
$p(\text{salary} \leq 100k) = {}^{3}/_{6} = 0.5$
Entropy = $-(0.5 * \log 0.5 + 0.5 * \log 0.5)$
Total Entropy = 0.578

- Entropy is a measure of how much randomness we have
- Calculate the individual entropy and find the weighted sum
- More entropy means more randomness
- Less randomness means high information gain



- Option A is selected because option B is having high entropy

Construct a decision tree for the following dataset.

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attributes are

1. Outlook
2. Temperature
3. Humidity
4. Wind

Target

Play tennis or not?