CLASSIFICATION

CLASSIFICATION ALGORITHM

- In Classification, a program learns from the given dataset or observations and then classifies new observations into some classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called targets/labels or categories.
- It is a type of supervised learning where the output variable is categorical.
- In the classification algorithm, a discrete output function(y) is mapped to the input variable(x).

```
Y=f(x), where y = categorical output
```

LEARNERS IN CLASSIFICATION PROBLEMS

Eager learners are machine learning algorithms that first build a model from the training dataset before making any prediction on future datasets. They spend more time during the training process because of their eagerness to have a better generalization during the training from learning the weights, but they require less time to make predictions.

Most machine learning algorithms are eager learners, and below are some examples:

- Logistic Regression.
- Support Vector Machine.
- Decision Trees.
- Artificial Neural Networks.

Lazy learners or instance-based learners, on the other hand, do not create any model immediately from the training data, and this is where the lazy aspect comes from.

- K-Nearest Neighbour.

USE CASES OF CLASSIFICATION PROBLEMS

Classification algorithms can be used in different places. Below are some popular use cases of Classification Algorithms:

- ✓ Email Spam Detection
- ✓ Identifications of Cancer tumor cells

EVALUATING A CLASSIFICATION MODEL

1. Log Loss or Cross-Entropy Loss:

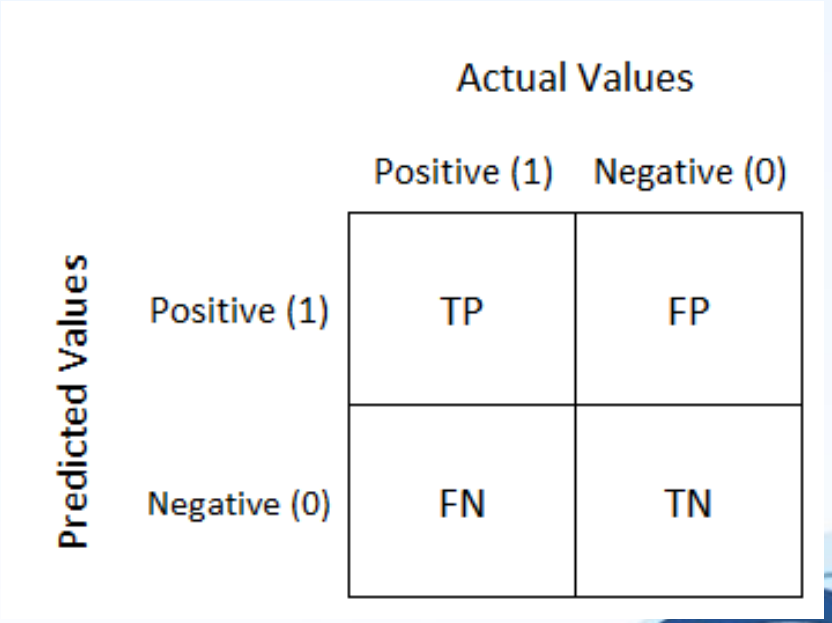
Here the model output is a probability value between 0 and 1.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\log(pic)$$

- N is the number of samples (instances) in the dataset.
- M is the number of classes in the classification problem.
- yic is an indicator (0 or 1) if class c is the correct classification for instance i.
- pic is the predicted probability that instance i belongs to class c according to the model.

EVALUATING A CLASSIFICATION MODEL

- A **confusion matrix** is a matrix that summarizes the performance of a machine learning model on a set of test data.
- It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions



K-NEAREST NEIGHBOUR

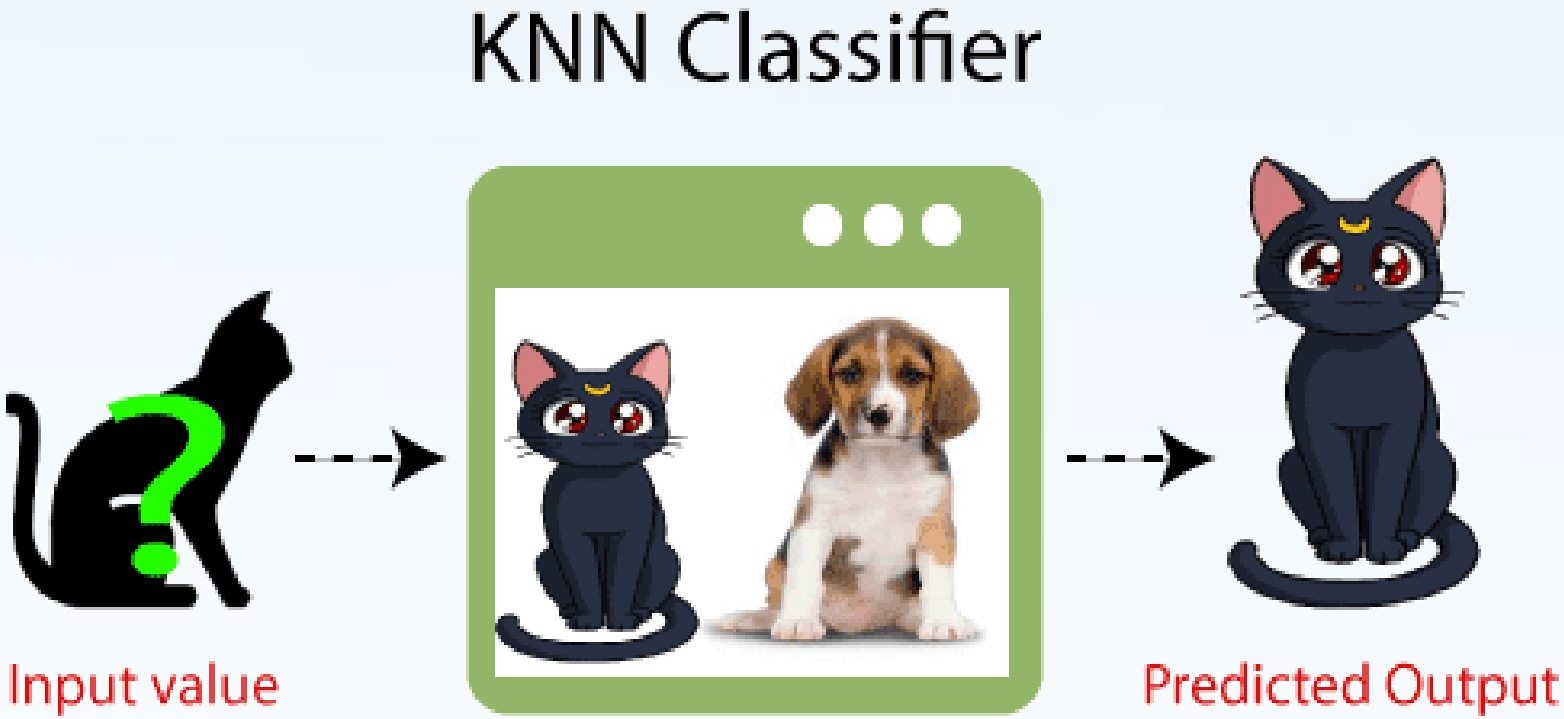
EXAMPLE

- Imagine entering a restaurant where the atmosphere is enveloped in total darkness.
- You are seated at your table and begin your culinary adventure, relying solely on your senses of touch, taste, and smell.
- This dining experience can be a metaphor for a type of machine learning technique known as nearest neighbor classification.
- Nearest neighbor classifiers excel in situations where the relationships among features and target classes are complex and not easily defined, yet items within the same class are quite similar.

OVERVIEW OF K-NEAREST NEIGHBOUR

- The K-Nearest Neighbors (K-NN) algorithm falls under the category of supervised learning.
- It operates on the principle that similar data points are likely to belong to the same category. When a new data point appears, K-NN searches for the 'K' training examples that are closest to this new point.
- K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.
- K-Nearest Neighbors (K-NN) is considered a **non-parametric algorithm** because it does not make any assumptions about the underlying distribution of the data.

OVERVIEW



WHY IS THE K-NN ALGORITHM LAZY?

- Classification algorithms based on nearest-neighbor methods are considered lazy learning algorithms because they do not involve abstraction or generalization during the training phase.

Key Characteristics of Lazy Learning Algorithms:

✓ No Abstraction or Generalization

✓ Rapid Training Phase

✓ Slow Prediction Phase

✓ Instance-Based Learning

✓ Non-Parametric Nature

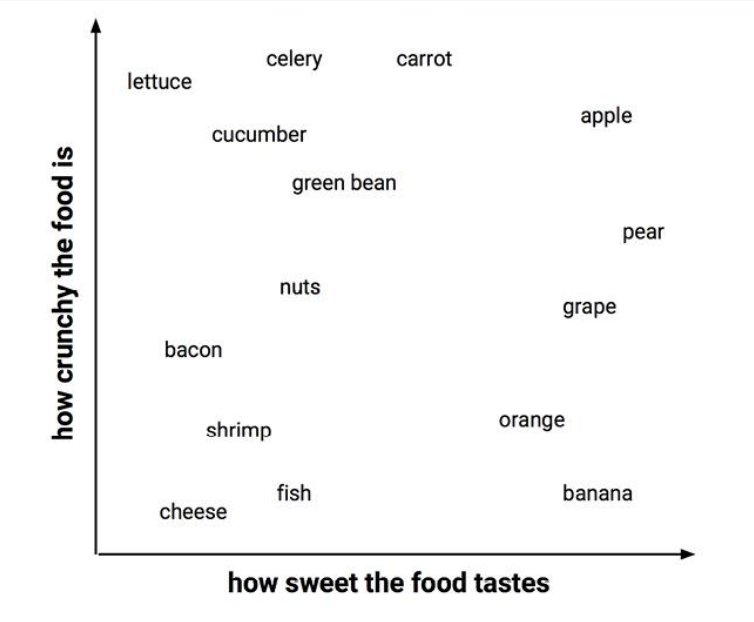
MEASURING SIMILARITY WITH THE DISTANCE

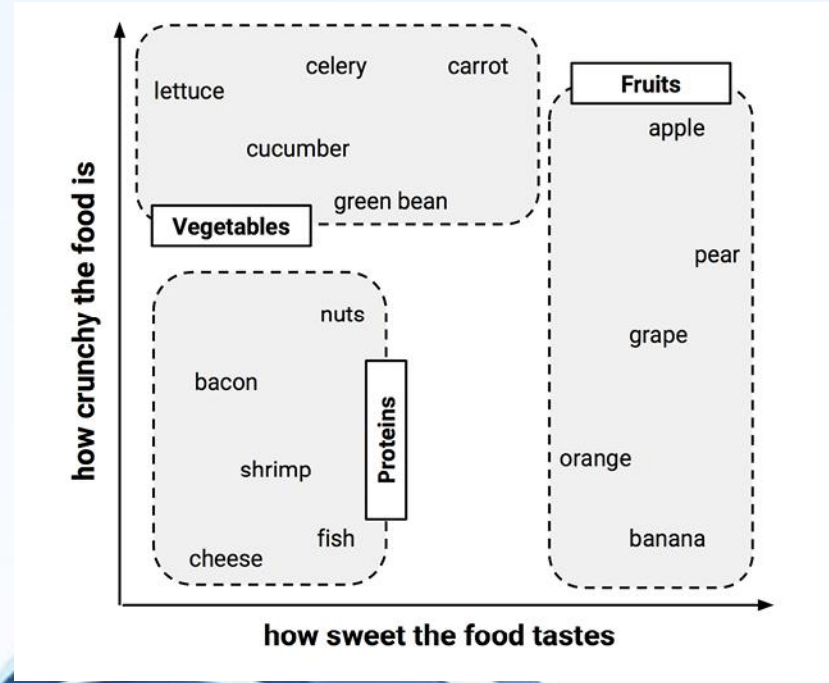
The first is a measure from 1 to 10 of how crunchy the ingredient is, and the second is a score from 1 to 10 measuring how sweet the ingredient tastes.

Ingredient	Sweetness	Crunchiness	Food type
Apple	10	9	Fruit
Bacon	1	4	Protein
Banana	10	1	Fruit
Carrot	7	10	Vegetable
Celery	3	10	Vegetable

MEASURING SIMILARITY WITH THE DISTANCE

Similar types of food tend to be grouped closely together. As illustrated in *the below figure*, vegetables tend to be crunchy but not sweet; fruits tend to be sweet and either crunchy or not crunchy; and proteins tend to be neither crunchy nor sweet:





MEASURING SIMILARITY WITH THE DISTANCE

- Locating the tomato's nearest neighbors requires a **distance function**, which is a formula that measures the similarity between two instances.
- Euclidean distance is specified by the following formula, where p and q are the examples to be compared, each having n features

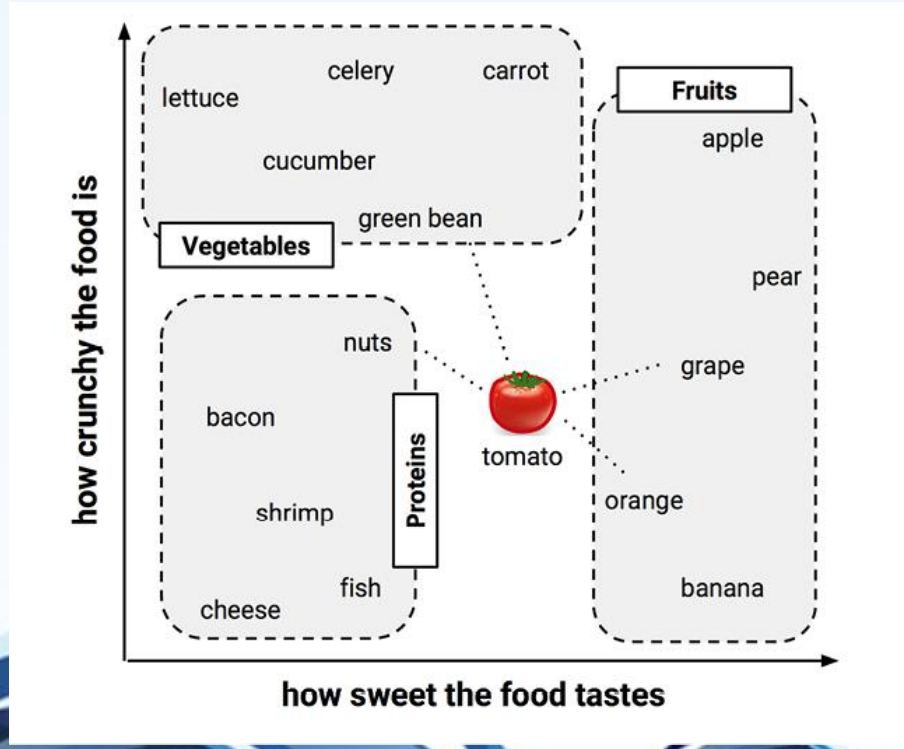
$$\text{dist}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

- For example, to calculate the distance between the tomato

(sweetness = 6, crunchiness = 4), and the green bean

(sweetness = 3, crunchiness = 7),

$$\text{dist}(\text{tomato}, \text{green bean}) = \sqrt{(6-3)^2 + (4-7)^2} = 4.2$$



MEASURING SIMILARITY WITH THE DISTANCE

To classify the tomato as a vegetable, protein, or fruit, we'll begin by assigning the tomato the food type of its single nearest neighbor. This is called 1-NN classification because *k = 1*.

The orange is the single nearest neighbor to the tomato, with a distance of 1.4. Because an orange is a fruit, the 1-NN algorithm would classify a tomato as a fruit.

Ingredient	Sweetness	Crunchiness	Food type	Distance to the tomato
Grape	8	5	Fruit	sqrt((6 - 8)^2 + (4 - 5)^2) = 2.2
Green bean	3	7	Vegetable	sqrt((6 - 3)^2 + (4 - 7)^2) = 4.2
Nuts	3	6	Protein	sqrt((6 - 3)^2 + (4 - 6)^2) = 3.6
Orange	7	3	Fruit	sqrt((6 - 7)^2 + (4 - 3)^2) = 1.4

CHOOSING AN APPROPRIATE K

Some common approaches and considerations:

Rule of Thumb:

- One common heuristic is to start with k equal to the square root of the number of training examples

2. Empirical Testing:

- Another approach involves testing multiple values of k on various test datasets and evaluating their classification performance.

ADVANTAGES AND DISADVANTAGES OF K-NEAREST NEIGHBORS (KNN)

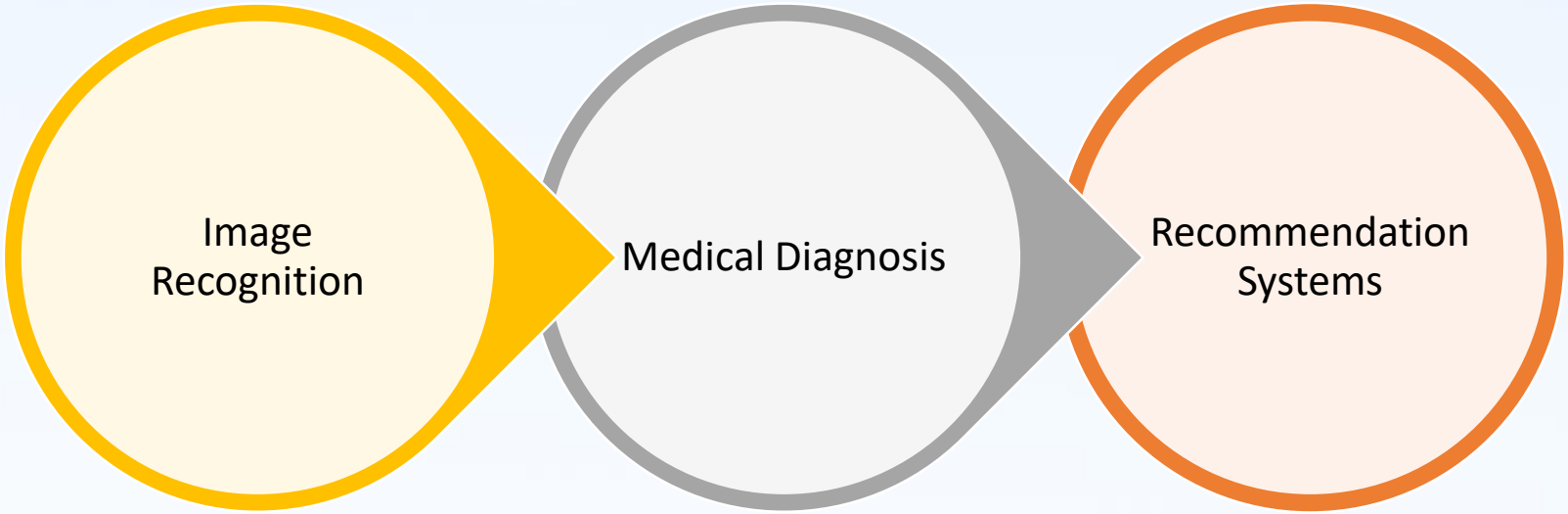
Advantages:

- Simplicity and Intuitiveness
- No Training Phase
- Versatility

Disadvantages

- Computational Cost
- Sensitive to Irrelevant Features
- Careful Parameter Selection

REAL LIFE USE CASES OF KNN



Activity: Predicting Product Purchase with K-Nearest Neighbors (KNN) Classifier

PROBABILISTIC LEARNING – CLASSIFICATION USING NAIVE BAYES

PROBABILISTIC LEARNING – CLASSIFICATION USING NAIVE BAYES

- When a meteorologist provides a weather forecast, precipitation is typically described with phrases like “70 percent chance of rain.”
- Weather estimates are based on probabilistic methods, which are those concerned with describing uncertainty. They use data on past events to extrapolate future events.
- In the case of the weather, the chance of rain describes the proportion of prior days with similar atmospheric conditions on which precipitation occurred.

UNDERSTANDING NAIVE BAYE'S

- A probability is a number between zero and one that captures the chance that an event will occur in light of the available evidence.
- Classifiers based on Bayesian methods utilize training data to calculate the probability of each outcome based on the evidence provided by feature values.

.

UNDERSTANDING NAIVE BAYE'S

- Bayesian probability theory is rooted in the idea that the estimated likelihood of an **event**, or potential outcome, should be based on the evidence at hand across multiple **trials**, or opportunities for the event to occur.

  ✓ The probability of an event is estimated from observed data by dividing the number of trials in which the event occurred by the total number of trials.

  ✓ For instance, if it rained 3 out of 10 days with similar conditions as today, the probability of rain today can be estimated as *3 / 10 = 0.30* or 30 percent.

Event	Trial
Heads result	A coin flip
Rainy weather	A single day (or another time period)
Message is spam	An incoming email message
Candidate becomes president	A presidential election
Mortality	A hospital patient
Winning the lottery	A lottery ticket

UNDERSTANDING NAIVE BAYE'S

- To denote these probabilities, we use notation in the form *P(A)*
- For example, *P(rain) = 0.30* to indicate a 30 percent chance of rain
- Because a trial always results in some outcome happening, the probability of all possible outcomes of a trial must always sum to one.
- Knowing the P(spam) = 0.20 allows us to calculate P(ham) = 1 – 0.20 = 0.80. This only works because spam and ham are mutually exclusive and exhaustive events, which implies that they cannot occur at the same time and are the only possible outcomes.

UNDERSTANDING NAIVE BAYE'S

- Events are exhaustive if they represent all possible outcomes of a trial. That is, at least one of the events must occur.

A complementary event is an event that represents all outcomes not included in the event of interest. If *A* is an event, then its complement, Ac, consists of all outcomes where *A* does not occur. For example:

- If A is the event "it rains today," then Ac(complement of *A*) is the event "it does not rain today."
- If *A* is the event "an email is spam," then *Ac* is the event "an email is ham."

UNDERSTANDING JOINT PROBABILITY

Scenario: You roll two six-sided dice. Each die has faces numbered from 1 to 6. We want to find the joint probability of two events:

Event A: The first die shows a 3.

Event B: The second die shows a 5.

Steps to Calculate Joint Probability

1.Determine Total Possible Outcomes: When rolling two dice, each die has 6 possible outcomes. Therefore, the total number of possible outcomes is:

$6 \times 6 = 36$

2. Count Favorable Outcomes: We want the first die to show a 3 and the second die to show a 5. There is exactly one such outcome: (3, 5).

UNDERSTANDING JOINT PROBABILITY

Calculate Joint Probability:

The joint probability $P(A \cap B)$ is the ratio of the number of favourable outcomes to the total number of possible outcomes.

$$P(\text{First die is 3 and second die is 5}) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

$$P(\text{First die is 3 and second die is 5}) = \frac{1}{36}$$

UNDERSTANDING JOINT PROBABILITY

Independent Events

- If the two events are unrelated, they are called independent events. This is not to say that independent events cannot occur at the same time.
- If all events were independent, it would be impossible to predict one event by observing another. In other words, dependent events are the basis of predictive modeling. Just as the presence of clouds is predictive of a rainy day

COMPUTING CONDITIONAL PROBABILITY WITH BAYES' THEOREM

- The relationships between dependent events can be described using Bayes' theorem, which provides a way of thinking about how to revise an estimate of the probability of one event in light of the evidence provided by another. One formulation is as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)\, P(A)}{P(B)}$$

COMPUTING CONDITIONAL PROBABILITY WITH BAYES' THEOREM

- Suppose a patient exhibits symptoms that could be indicative of a rare disease, such as Disease X. The symptoms are often associated with Disease X, but they can also occur due to other factors. We want to estimate the probability that the patient has Disease X given the symptoms they exhibit.

- **Prior Probability (Initial Belief)**: Let's say that based on historical data and medical knowledge, the prevalence of Disease X in the general population is very low, about 0.1%.

  $P(\text{Disease X})=0.001$

- **Likelihood of Symptoms Given Disease (Conditional Probability)**: Medical studies show that among patients who have Disease X, 95% of them exhibit the specific symptoms observed in the patient. So, the conditional probability of observing the symptoms given that the patient has Disease X is:

  $P(\text{Symptoms}|\text{Disease X})=0.95$

COMPUTING CONDITIONAL PROBABILITY WITH BAYES' THEOREM

Overall Likelihood of Symptoms (Marginal Probability): Among patients without Disease X, 10% of them exhibit the same symptoms. P(Symptoms)= 0.10

$$P(\text{Disease X}|\text{Symptoms}) = \frac{P(\text{Symptoms}|\text{Disease X}) \times P(\text{Disease X})}{P(\text{Symptoms})}$$

P(Disease X|Symptoms)=0.95×0.001/0.10

P(Disease X|Symptoms)=0.0095

So, given the observed symptoms, the posterior probability of the patient having Disease X is approximately 0.95%.

COMPUTING CONDITIONAL PROBABILITY WITH BAYES' THEOREM

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
- The Naïve Bayes algorithm is 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.
- Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

ADVANTAGES OF BAYE'S THEOREM

- Naïve Bayes is one of the fastest and easiest ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.

Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

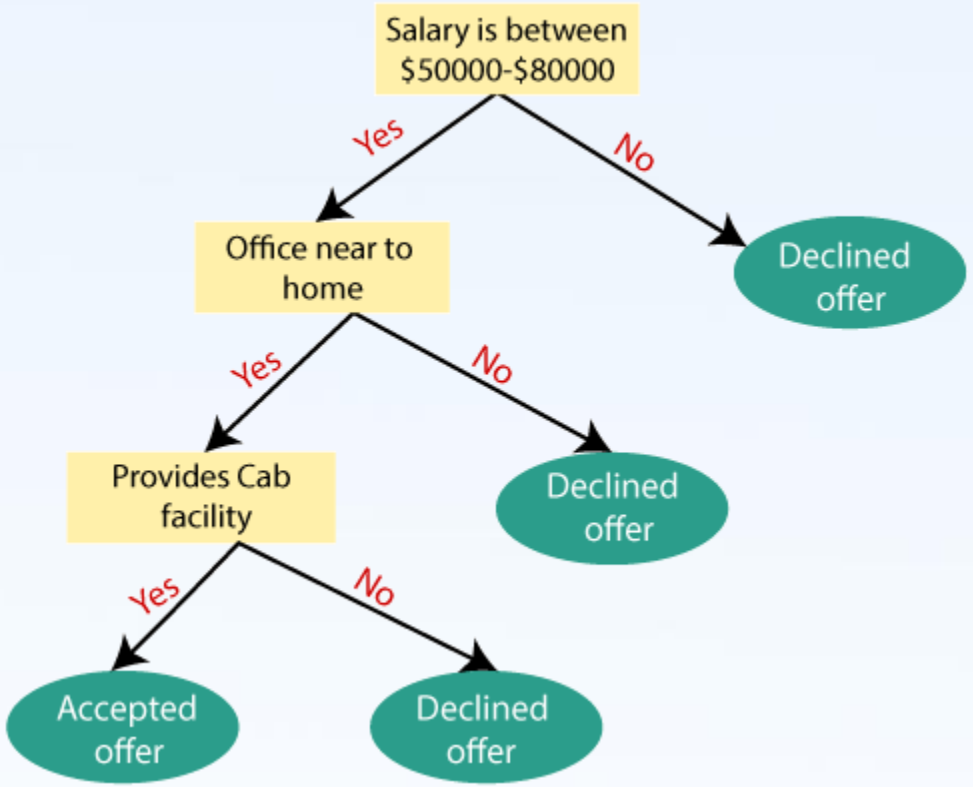
APPLICATIONS

- It is used for **Credit Scoring**.
- It is used in **medical data classification**.
- It is used in Text classification such as **Spam filtering** and **Sentiment analysis**.

DIVIDE AND CONQUER – CLASSIFICATION USING DECISION TREES AND RULES

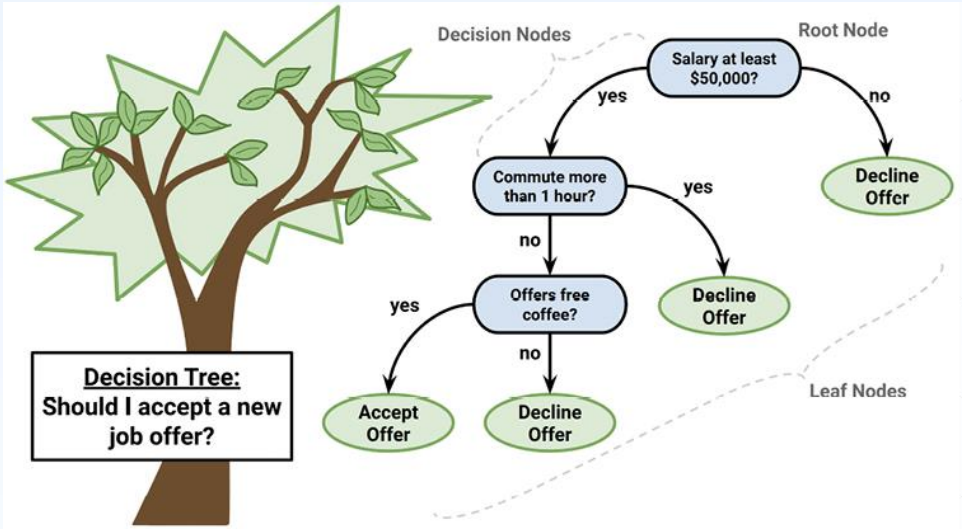
DIVIDE AND CONQUER – CLASSIFICATION USING DECISION TREES AND RULES

- When deciding between job offers, many people begin by making lists of pros and cons, then eliminate options using simple rules.
- For instance, they may decide, “If I have to commute for more than an hour, I will be unhappy,” or “If I make less than $50K, I can’t support my family.”

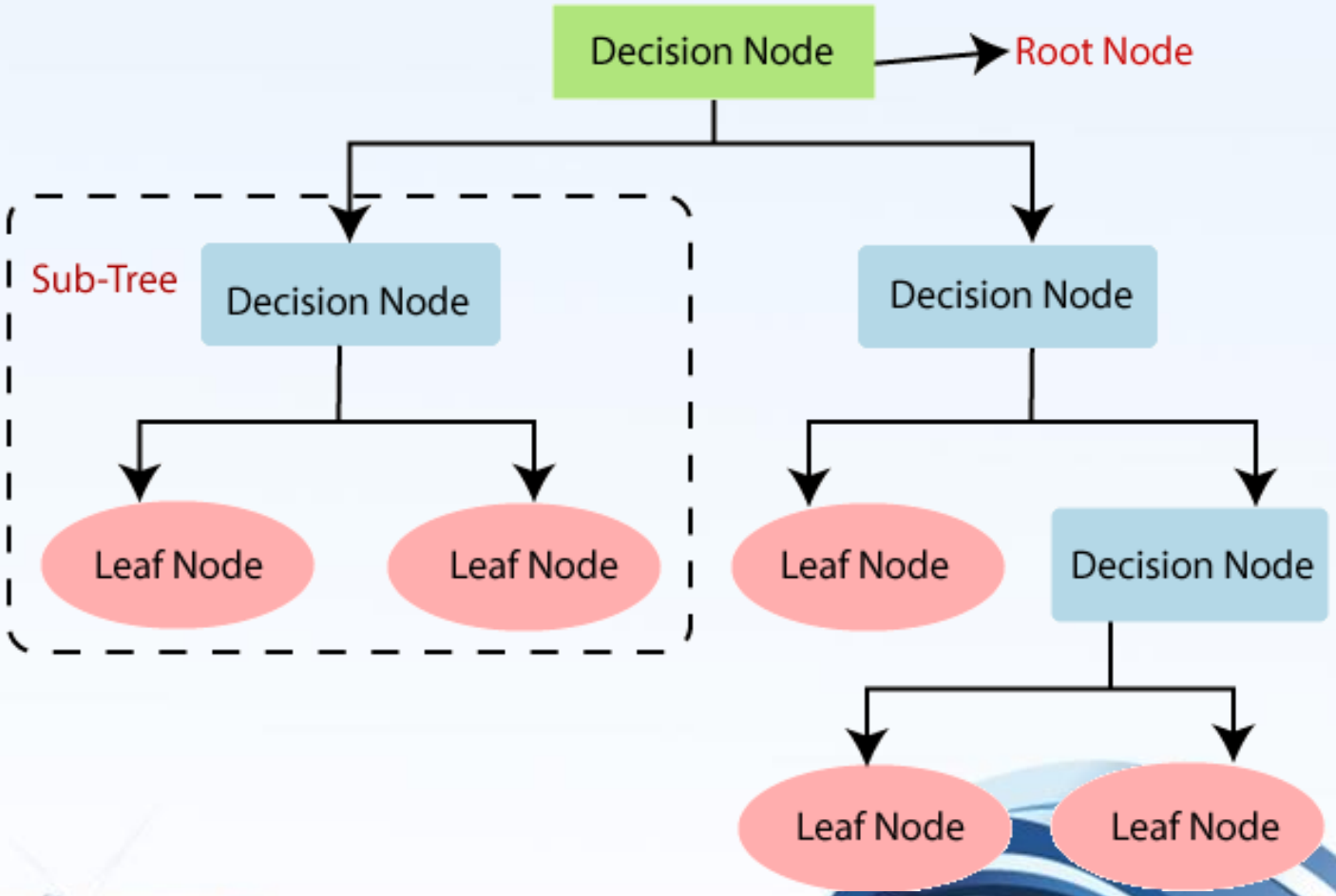


UNDERSTANDING DECISION TREES

- Decision tree learners are powerful classifiers that utilize a **tree structure** to model the relationships among the features and the potential outcomes.
- Let's consider the following tree, which predicts whether a job offer should be accepted.
- If a final decision can be made, the tree terminates in **leaf nodes** that denote the action to be taken as the result of the series of decisions.



UNDERSTANDING DECISION TREES



WHY USE DECISION TREES?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model.

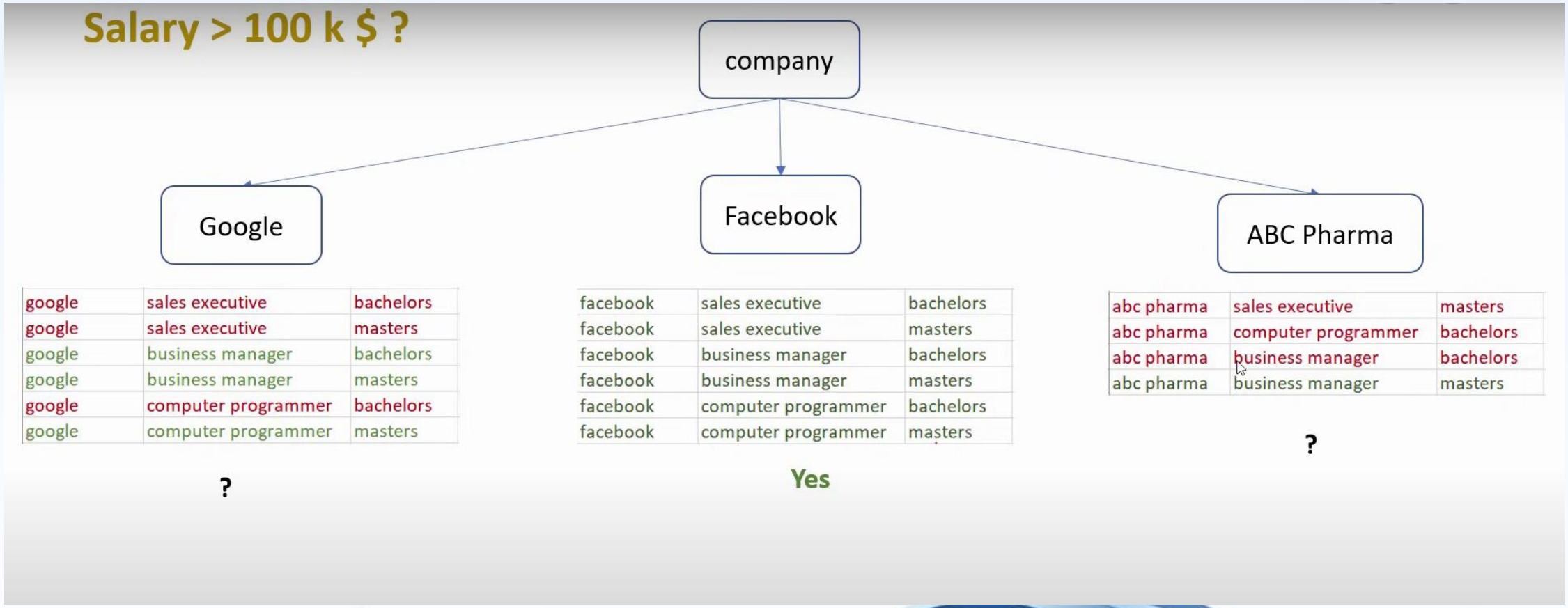
Below are the two reasons for using the Decision tree:

- It is easy to understand.
- It forms a tree-like structure.
- $\sum_{i=1}^{n} P_i log_2(pi)$

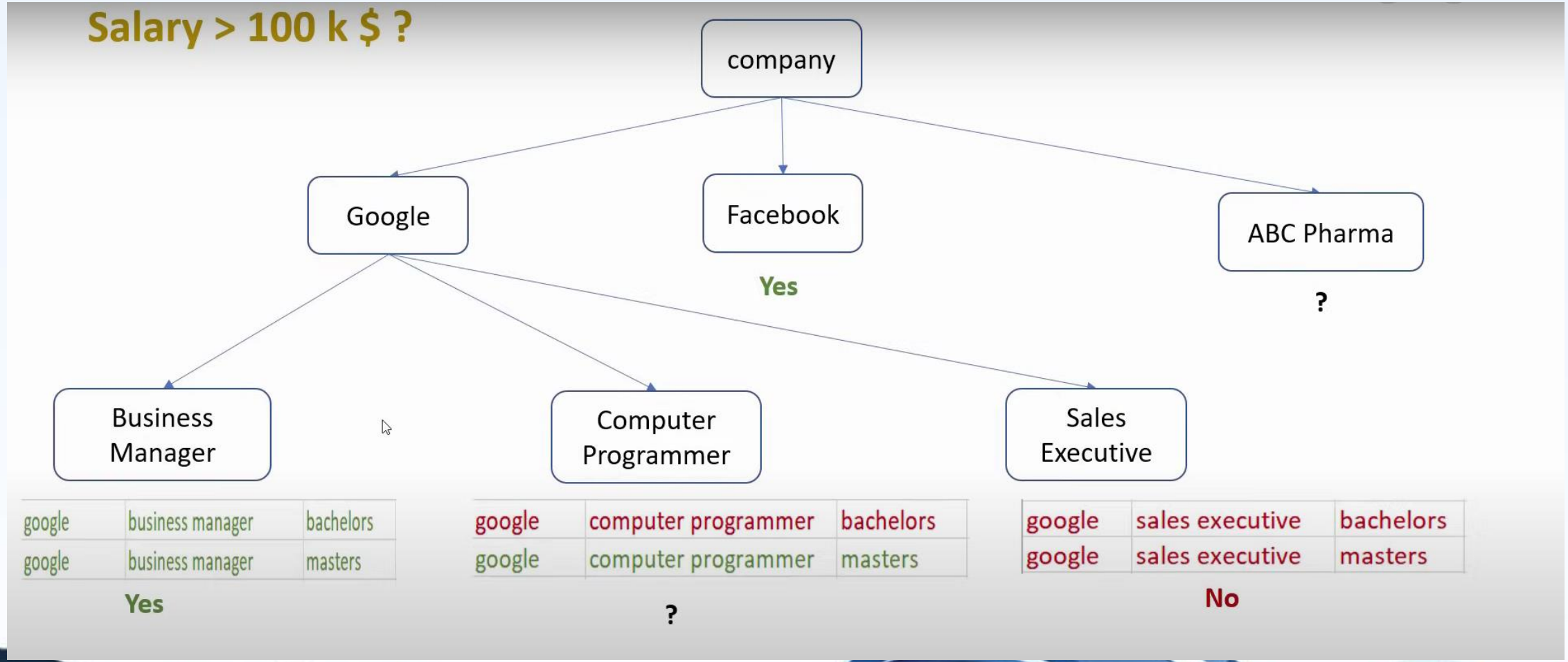
EXAMPE

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

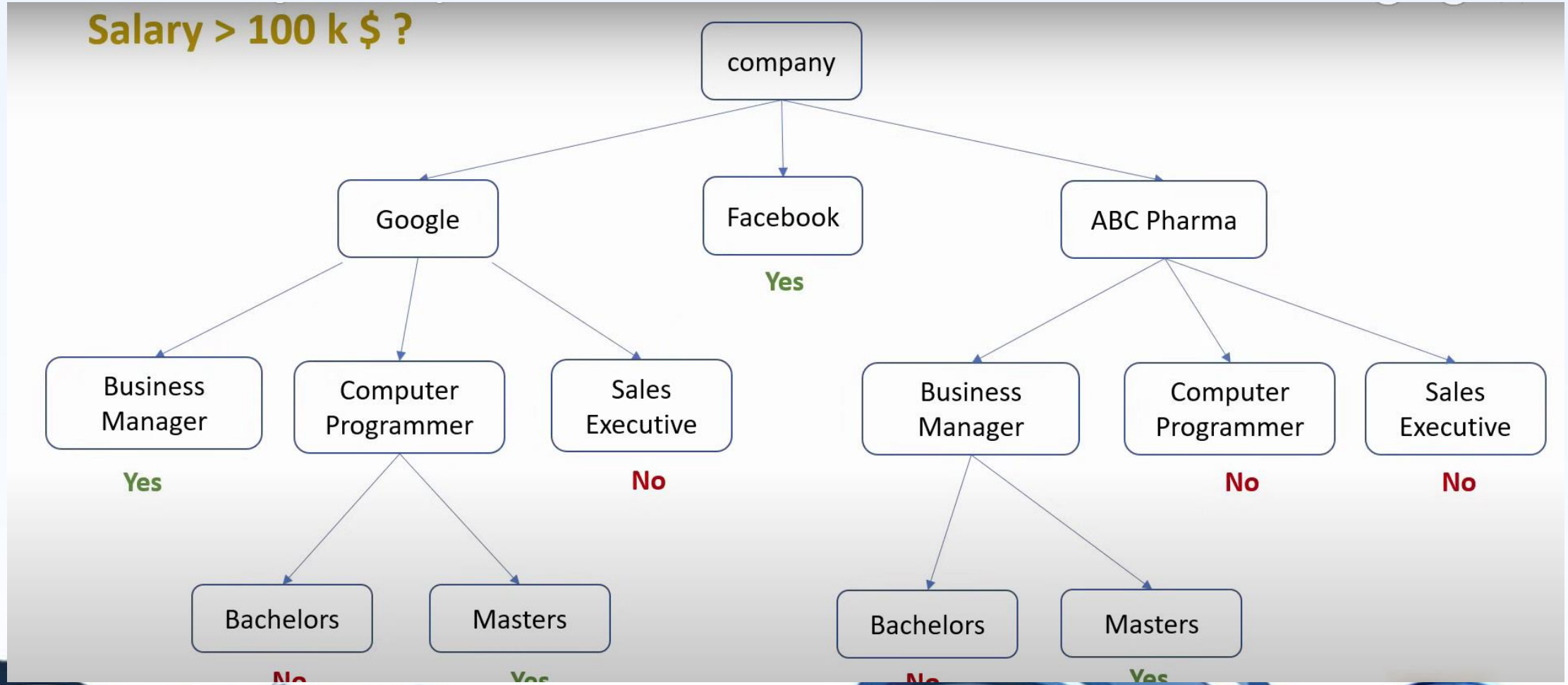
EXAMPE


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

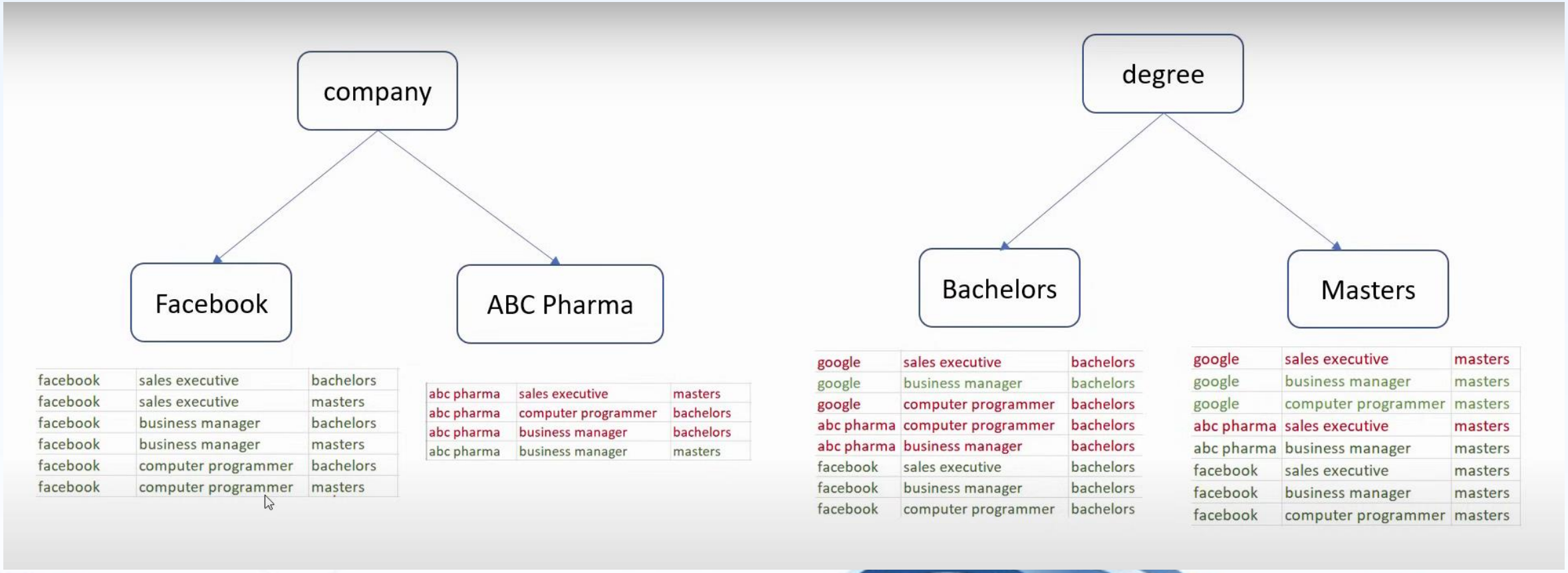
EXAMPE


Salary > 100 k $?
company
Google
Facebook
ABC Pharma
Yes
?
Business Manager
Computer Programmer
Sales Executive
google business manager bachelors
google business manager masters
Yes
google computer programmer bachelors
google computer programmer masters
?
google sales executive bachelors
google sales executive masters
No

EXAMPLE


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
Bachelors
Masters
Bachelors
Masters

EXAMPLE



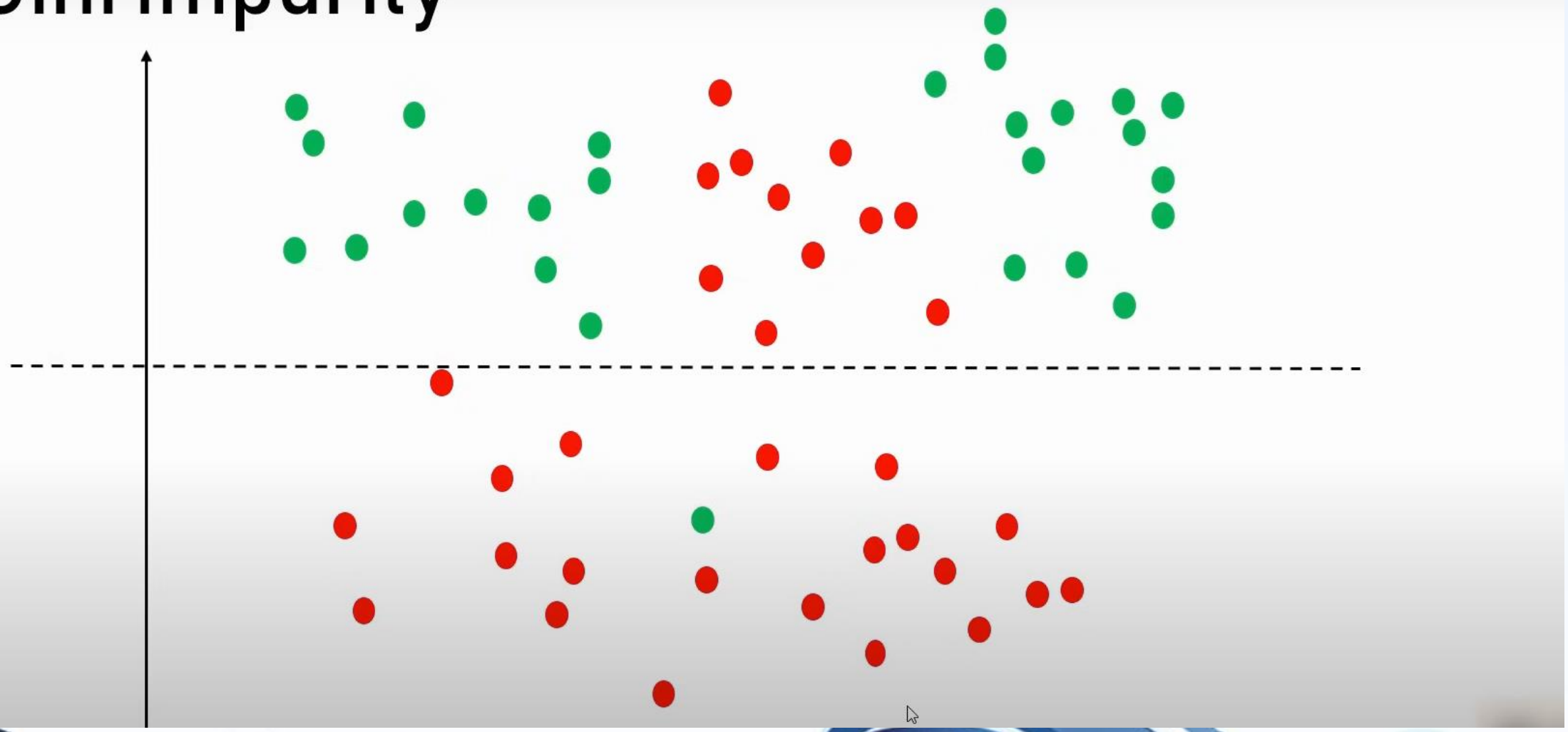
facebook	sales executive	bachelors
facebook	sales executive	masters
facebook	business manager	bachelors
facebook	business manager	masters
facebook	computer programmer	bachelors
facebook	computer programmer	masters

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	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

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

Gini Impurity



ADVANTAGES OF THE DECISION TREE

- It is simple to understand
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm.**
- For more class labels, the computational complexity of the decision tree may increase.

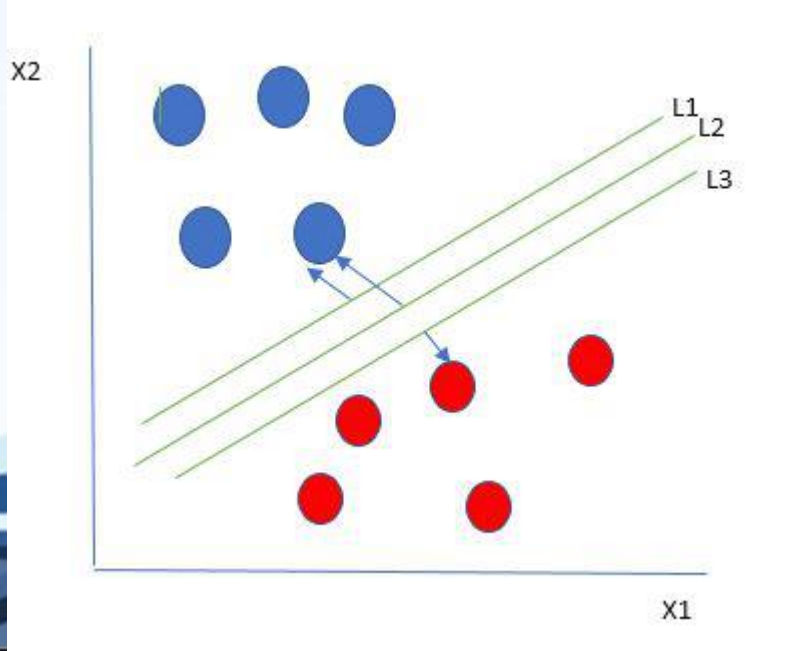
Activity: Decision Tree algorithm for predicting customer Purchases

SUPPORT-VECTOR MACHINE

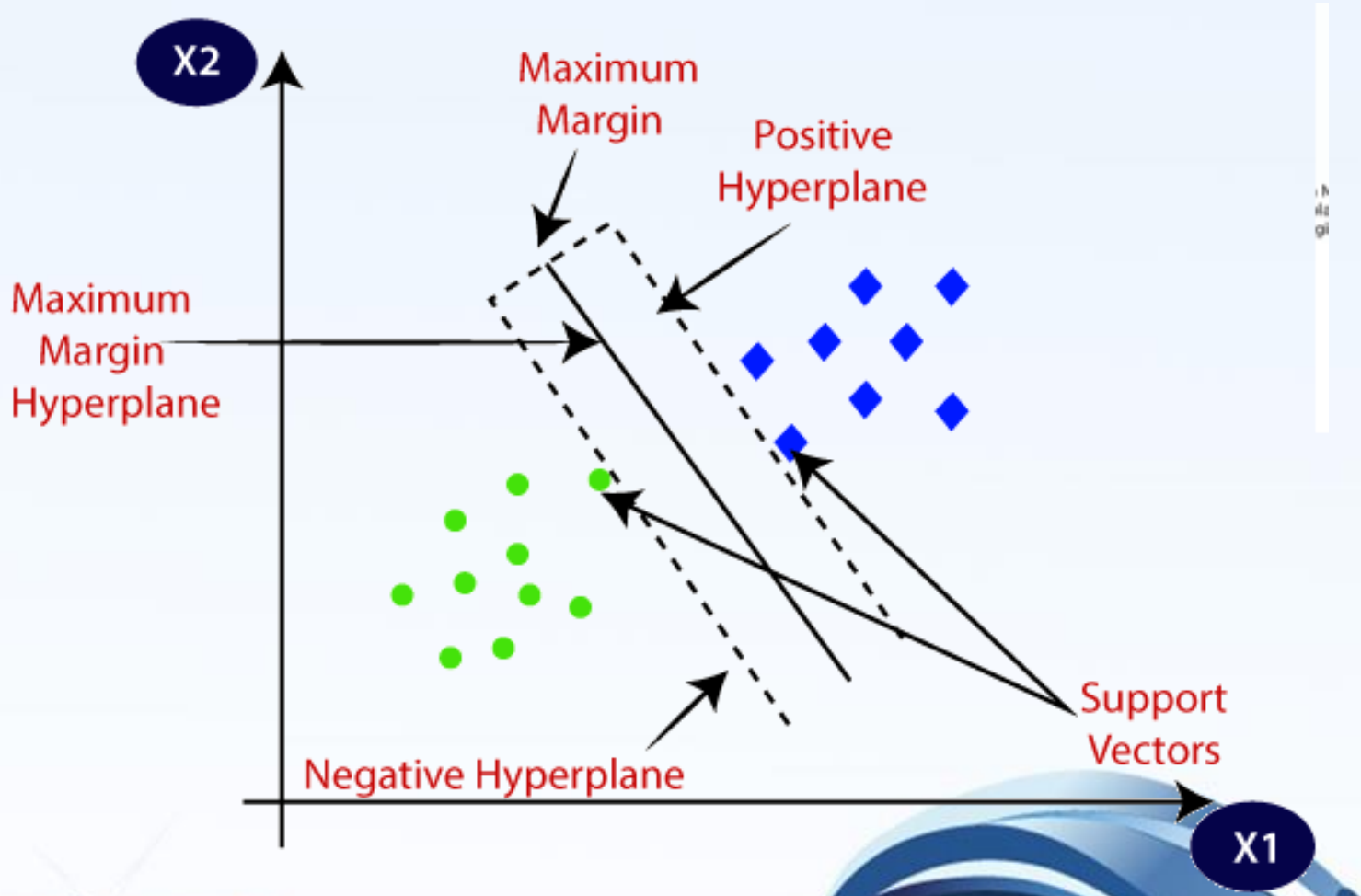
What is SVM?

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

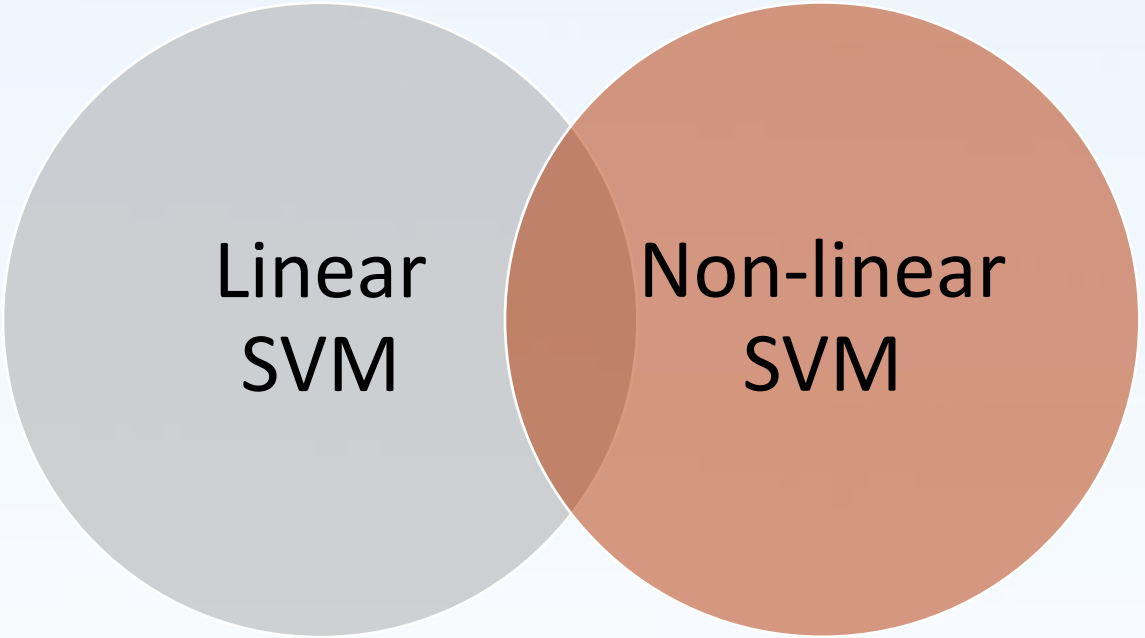
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



What is SVM?

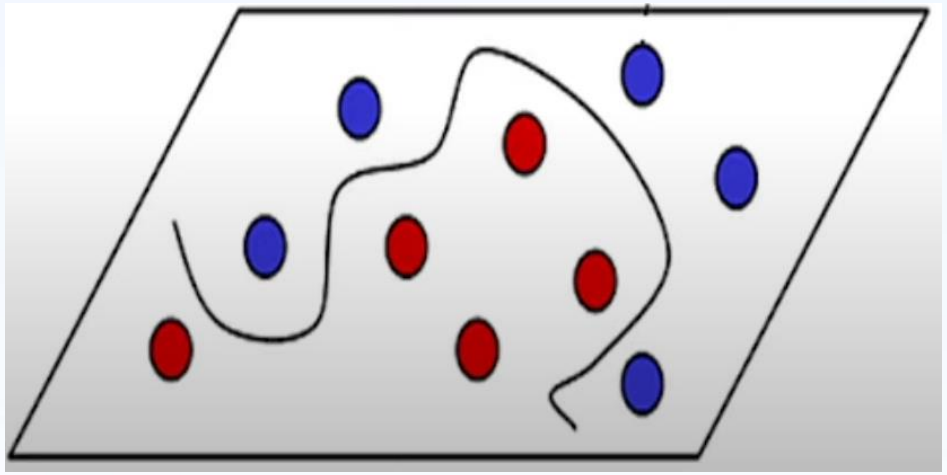

X2
Maximum
Margin
Positive
Hyperplane
Maximum
Margin
Hyperplane
Support
Vectors
Negative Hyperplane
X1

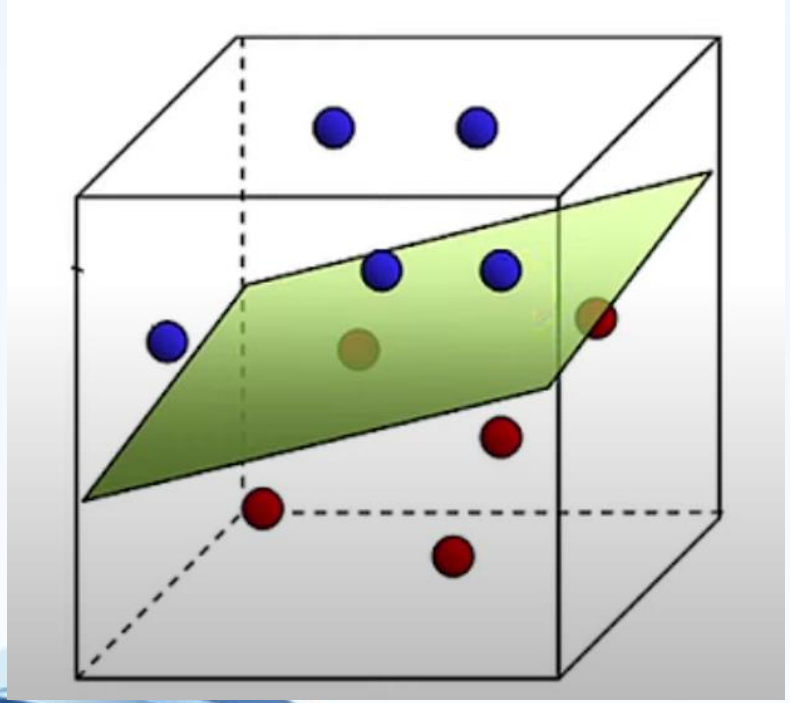
Types of SVM



Kernel Trick

- The **kernel trick** is a fundamental technique used in Support Vector Machines (SVM) and other machine learning algorithms to handle non-linearly separable data.





Major Kernel Function in Support Vector Machine

In Support Vector Machines (SVMs), several types of kernel functions can be used to map the input data into a higher-dimensional feature space. The choice of kernel function depends on the specific problem and the characteristics of the data.

Linear Kernel: It defines the dot product between the input vectors in the original feature space.

$$K(x, y) = x \cdot y$$

Where x and y are the input feature vectors. The dot product of the input vectors is a measure of their similarity or distance in the original feature space.

Major Kernel Function in Support Vector Machine

Polynomial Kernel: It is a nonlinear kernel function that employs polynomial functions to transfer the input data into a higher-dimensional feature space.

```
K(x, y) = (x. y + c)d
```

The polynomial kernel has the benefit of being able to detect both linear and nonlinear correlations in the data. It can be difficult to select the proper degree of the polynomial, though, as a larger degree can result in overfitting while a lower degree cannot adequately represent the underlying relationships in the data.

Major Kernel Function in Support Vector Machine

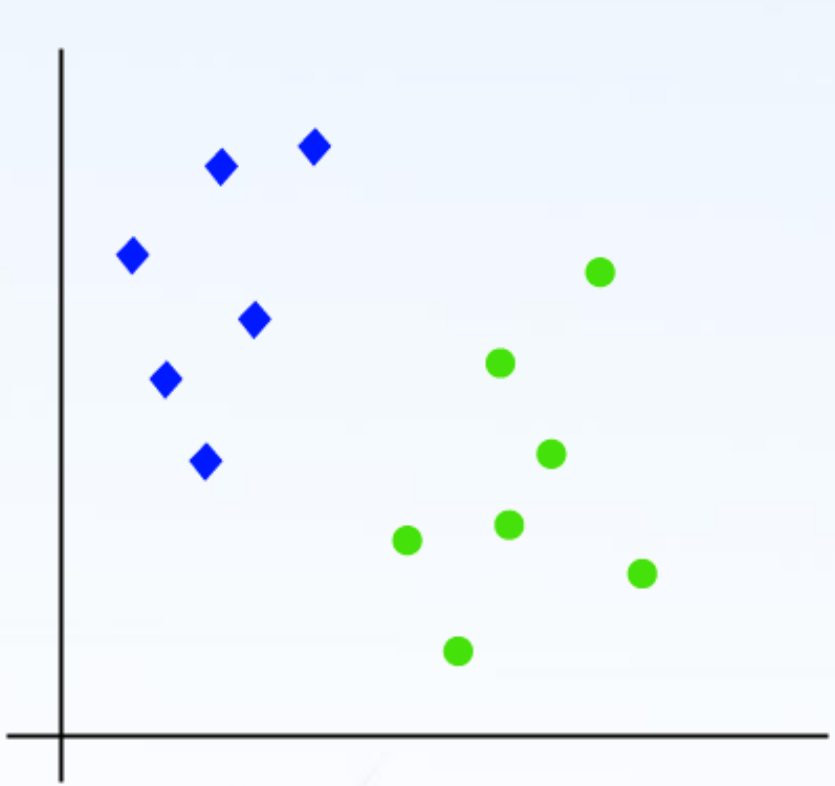
Gaussian Radial function:

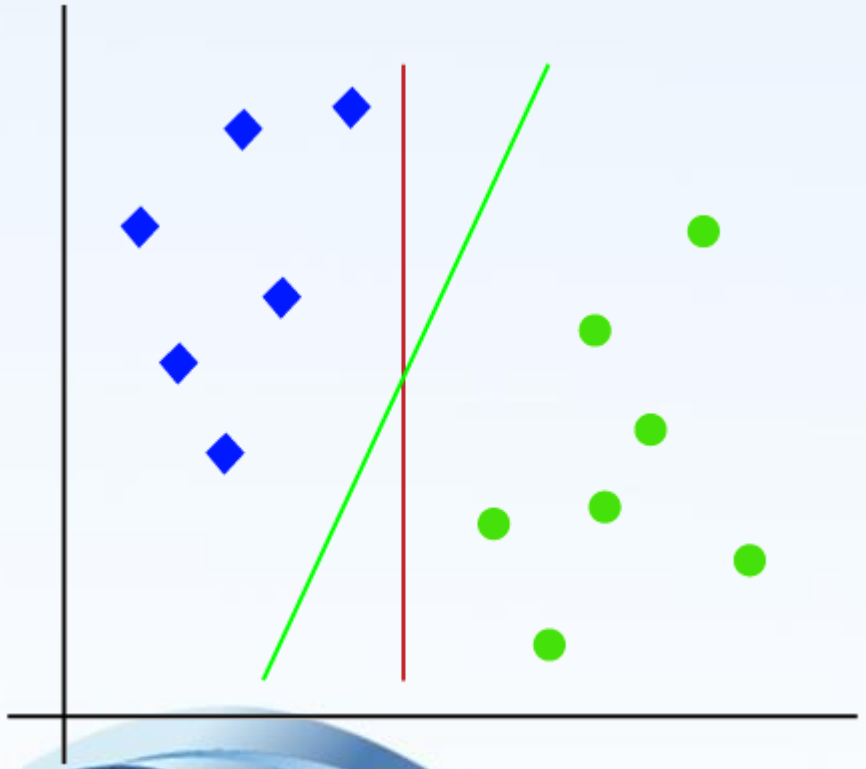
The Gaussian kernel, also known as the radial basis function (RBF) kernel, is a popular kernel function used in machine learning, particularly in SVMs (Support Vector Machines). It is a nonlinear kernel function that maps the input data into a higher-dimensional feature space using a Gaussian function.

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

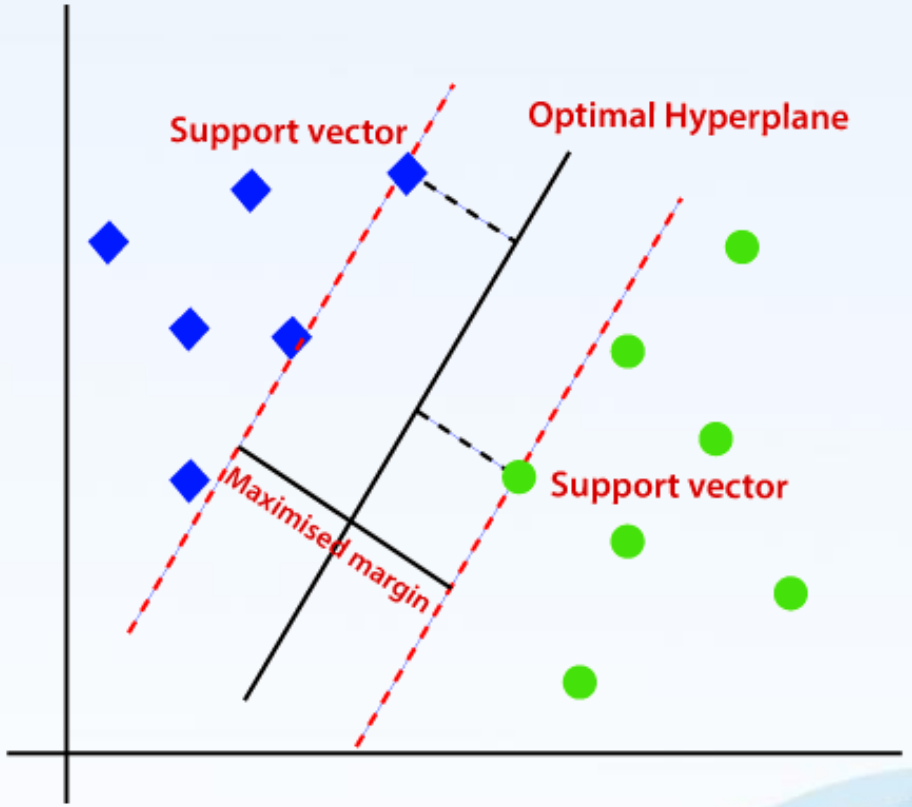
How does SVM works?

We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue





How does SVM works?



Activity: Analyze and segment users based on their age and estimated salary using SVM

End