Unit 6:
Machine Learning Algorithms

MACHINE LEARNING IN A NUTSHELL:

Machine Learning (ML) is a part of artificial intelligence (AI) that focuses on teaching computers to learn from data and make decisions without being explicitly programmed.

- ML algorithms learn from various types of data, including images, text, sensor readings, and historical records.
- Some common ML algorithms include decision trees, neural networks, and support vector machines.
- It powers recommendation systems like those used by Netflix, speech recognition, medical diagnosis, and autonomous vehicles. ML is also behind chatbots, personalized ads, and fraud detection systems.
- ML also presents challenges. Overfitting, where models become too specialized on training data, can lead to poor performance on new data. Bias in training data can result in biased predictions, and some models are difficult to interpret, acting as black boxes. Despite these challenges, ML transforms data into knowledge, enabling computers to learn, adapt, and make decisions autonomously.

- Artificial intelligence (AI) and machine learning (ML) have significantly impacted various aspects of our lives. From transportation and finance to healthcare and entertainment, AI algorithms are pervasive. They power self-driving cars, fraud detection systems, personalized shopping experiences, and virtual assistants like Siri and Alexa. As technology continues to evolve, the influence of AI and ML is only expected to grow, shaping the future of our society and culture.

Activity 1: Autodraw - Experience the power of machine learning with Autodraw! Autodraw combines machine learning with the creativity of talented artists, allowing you to draw things quickly and effortlessly. Visit the following link to play the game:

Types of Machine Learning

Types of Machine Learning

Machine Learning

Supervised Learning

Task Driven
(Classification/Regression)

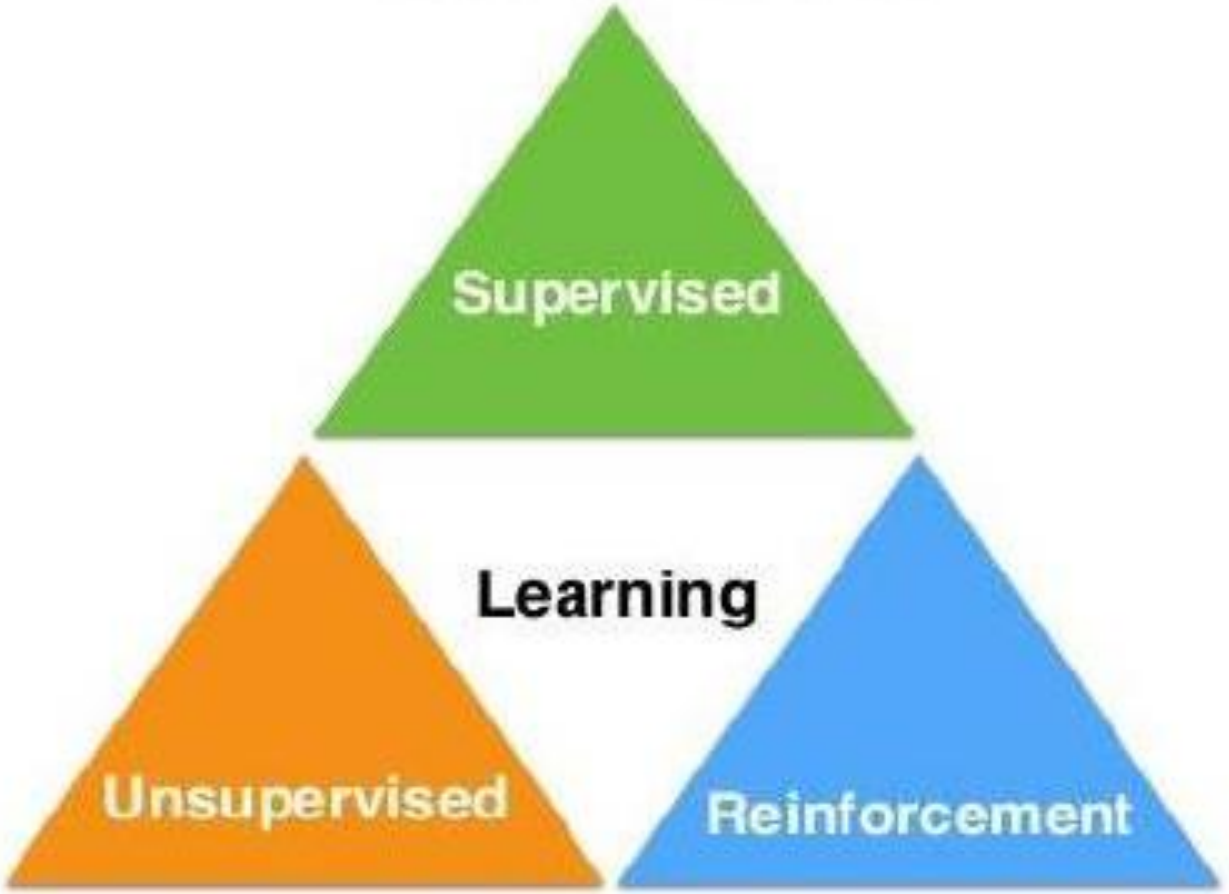
Unsupervised Learning

Data Driven
(Clustering)

Reinforcement Learning

Learning from
mistakes
(Playing Games)

- Labeled data
- Direct feedback
- Predict outcome/future



- No labels
- No feedback
- "Find hidden structure"

- Decision process
- Reward system
- Learn series of actions

A. Supervised Learning

It is a powerful approach that allows machines to learn from labeled data, making predictions or decisions based on that learning. Within supervised learning, two primary types of algorithms emerge:

1. Regression – works with continuous data

2. Classification – works with discrete data

enjoy algorithms

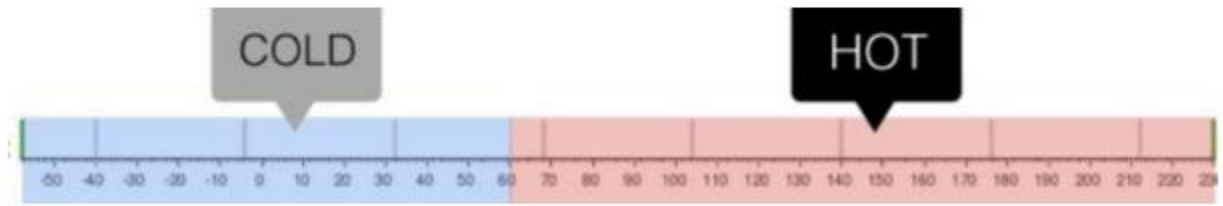
Regression

What will be the temperature tomorrow?

Fahrenheit

Classification

Will it be hot or cold tomorrow?



Fahrenheit

1. Regression

Understanding Correlation: The Foundation of Regression Analysis

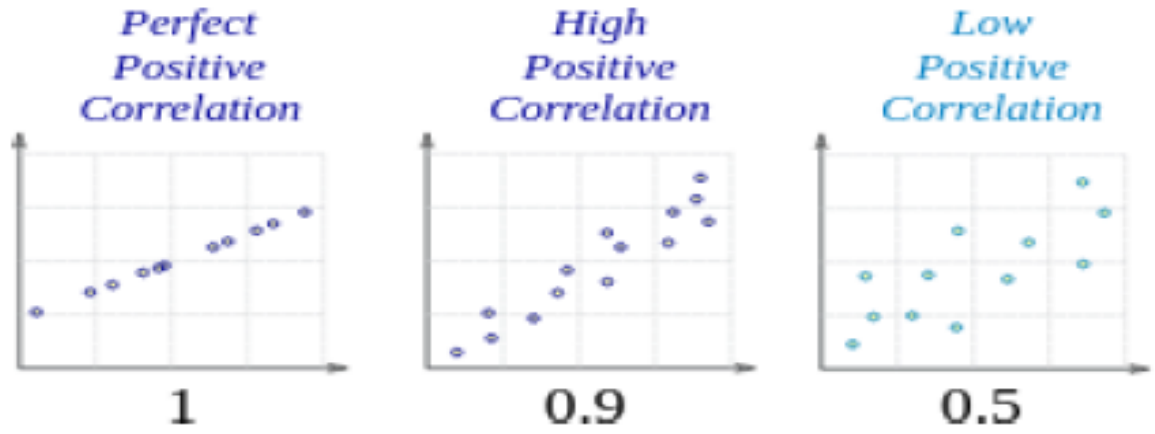
- In data analysis, correlation is a fundamental concept that helps us grasp the relationship between variables, laying the groundwork for predictive modeling and insightful analysis.

- Correlation is a measure of the strength of a linear relationship between two quantitative variables (e.g. price, sales).

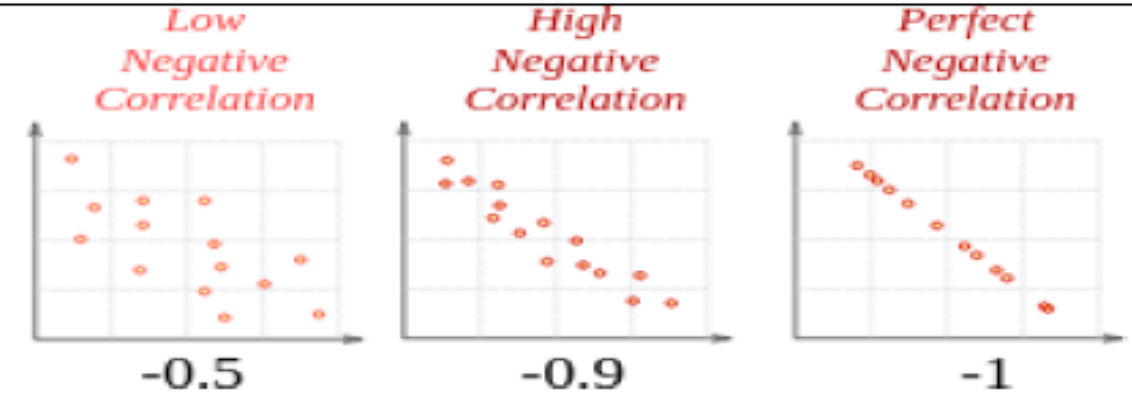
- If the change in one variable appears to be accompanied by a change in the other variable the two variables are said to be correlated and this inter dependence is called correlation.

Types of Correlation:

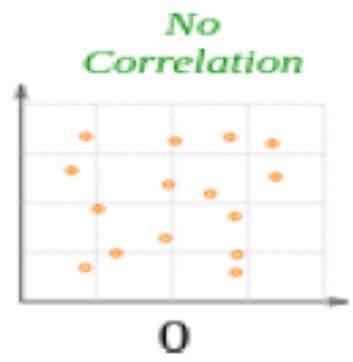
1. Positive Correlation: In a positive correlation, both variables move in the same direction. As one variable increases, the other also tends to increase, and vice versa.



2. Negative Correlation: Conversely, in a negative correlation, variables move in opposite directions. An increase in one variable is associated with a decrease in the other, and vice versa.

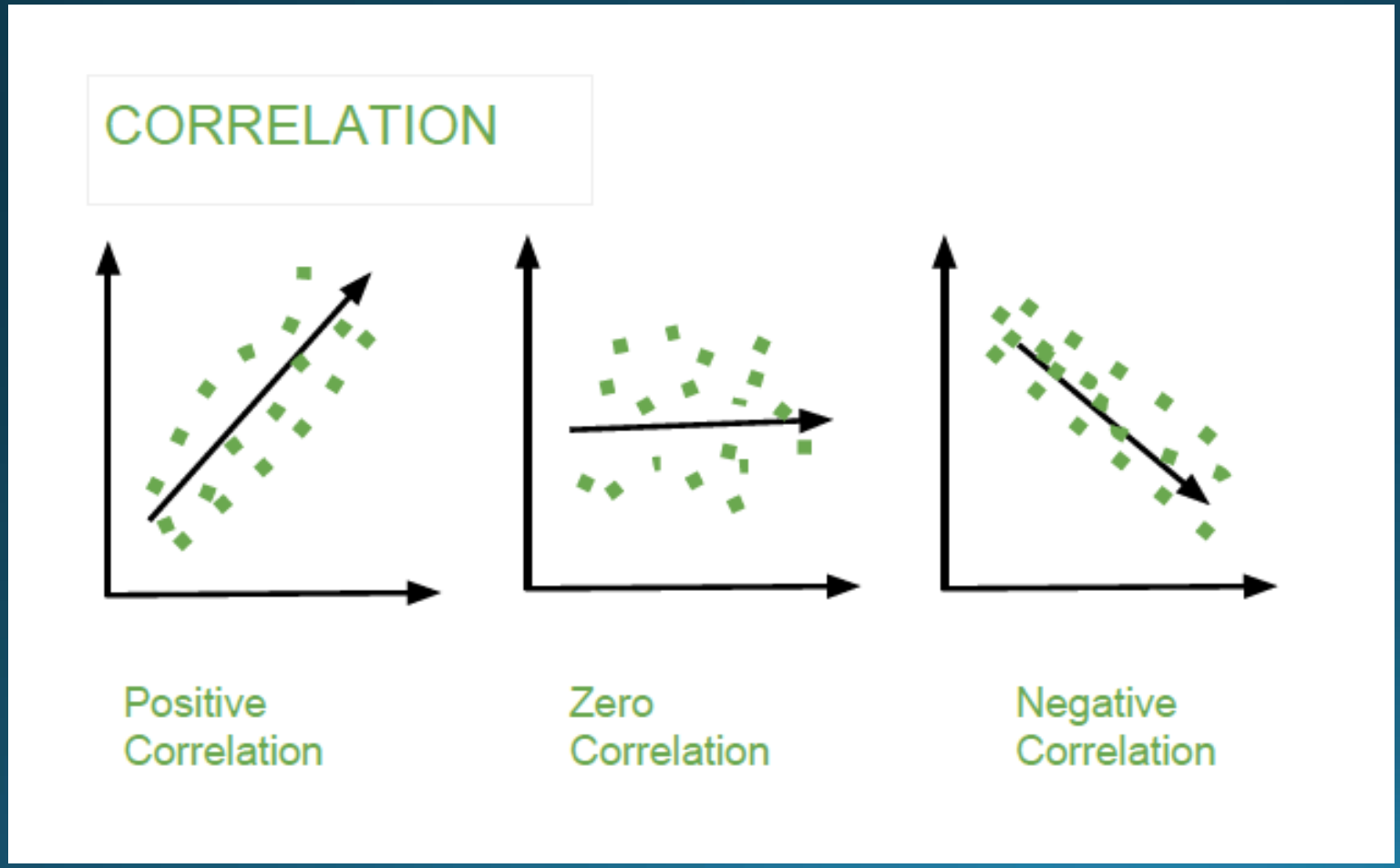


3. Zero Correlation: When there is no apparent relationship between two variables, they are said to have zero correlation. Changes in one variable do not predict changes in the other



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation



Causation: Causation indicates that one event is the result of the occurrence of the other event.

Example: Since there is hot weather, the person will use more sunscreen or eat more ice cream.

Sometimes, these two events may be correlated also. Example: smoking causes an increase in the risk of developing lung cancer or it can correlate with another like-smoking is correlated with alcoholism, but it does not cause alcoholism. Therefore, we can say causation is not always correlation.

PEARSON’S R

Pearson's correlation coefficient (often denoted as Pearson's r) is one of the crucial factors to consider when assessing the appropriateness of regression analysis.

Pearson's r measures the strength and direction of the linear relationship between two continuous variables.

The requirements when considering the use of Pearson's correlation coefficient are:

1. Scale of measurement should be interval or ratio.

2. Variables should be approximately normally distributed.

3. The association should be linear.

4. There should be no outliers in the data.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

r can take a range of values from +1 to -1

- A value of 0 indicates that there is no association between the two variables.

- A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

- A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Example 1

In the example below of 6 people with different ages and different weight, let us try calculating the value of the Pearson r.

Sr. No	Age (x)	Weight (y)
1	40	78
2	21	70
3	25	60
4	31	55
5	38	80
6	47	66

Solution:

For the Calculation of the Pearson Correlation Coefficient, we will first calculate the following values:

Sr. No	Age (x)	Weight (y)	xy	x^2	y^2
1	40	78	3120	1600	6084
2	21	70	1470	441	4900
3	25	60	1500	625	3600
4	31	55	1705	961	3025
5	38	80	3040	1444	6400
6	47	66	3102	2209	4356
Total (Σ)	202	409	13937	7280	28365

Here the total number of people is 6 so, **n=6**

Now the calculation of the Pearson R is as follows:

E12

=(6*D10-B10*C10)/SQRT((6*E10-B10^2)*(6*F10-C10^2))

Using Formula

	A		D	E	F	G
3	Sr. No	Age (x)	Weight (y)	xy	x^2	y^2
4	1	40	78	3120	1600	6084
5	2	21	70	1470	441	4900
6	3	25	60	1500	625	3600
7	4	31	55	1705	961	3025
8	5	38	80	3040	1444	6400
9	6	47	66	3102	2209	4356
10	Total (Σ)	202	409	13937	7280	28365
11						
12	Pearson Correlation Coefficient (r)			0.35		

r = (n (Σxy)- (Σx)(Σy))/(√ [n Σx2-(Σx)2][n Σy2– (Σy)2)

r = (6 * (13937)- (202)(409)) / (√ [6 *7280 -(202)2] * [6 * 28365- (409)2)

r = (6 * (13937)- (202) * (409))/(√ [6 *7280 -(202)2] * [6 * 28365- (409)2)

r = (83622- 82618)/(√ [43680 -40804] * [170190- 167281)

r = 1004/(√ [2876] * [2909)

r = 1004 / (√ 8366284)

r = 1004 / 2892.452938

r = 0.35

The value of the Pearson correlation coefficient is 0.35

It is important to note that, regression analysis may not be suitable in certain situations:

1. No Correlation: If there is no correlation between the variables, meaning they change independently of each other, regression analysis will not provide meaningful insights or predictions.

2. Non-linear Relationships: While regression can model linear relationships well, it may not capture more complex, non-linear relationships effectively. In such cases, alternative techniques like polynomial regression or non-linear regression may be more appropriate.

3. Outliers: Outliers, or extreme data points, can disproportionately influence the regression model and lead to inaccurate predictions. In the presence of outliers, it is essential to assess their impact and consider alternative modeling approaches.

4. Violation of Assumptions: Regression analysis relies on certain assumptions, such as the linearity of relationships and the absence of multicollinearity (high correlation between predictor variables). If these assumptions are violated, the results of the regression analysis may be unreliable.

REGRESSION

Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables.

Its primary objective is to understand and predict the value of the dependent variable based on the values of the independent variables.

Regression analysis is particularly useful when dealing with continuous data, where variables can take on any value within a certain range.

For example, variables such as height, temperature, salary, and time are all continuous, meaning they can be measured along a continuous scale.

By analyzing the relationship between the independent and dependent variables, regression allows us to make predictions and understand how changes in one variable may impact the other.

This makes regression a powerful tool for forecasting, prediction, and understanding complex relationships in various fields such as economics, social sciences, and healthcare.

When we make a distribution in which there is an involvement of more than one variable, then such an analysis is called Regression Analysis.

Let there be two variables *x and y*. If *y* depends on *x,* then the result comes in the form of a simple regression. Furthermore, we name the variables *x* and *y* as:

y – Regression / Dependent / Explained Variable.
It is the variable we want to predict or understand.

x – Independent /Predictor / Explanator Variable
It is used to predict or explain changes in the dependent variable.

Therefore, if we use a simple linear regression model where *y* depends on *x,* then the regression line of *y* on *x* is:

$$y = a + bx + e$$

In this equation,

- **a** represents the **intercept** of the regression line with the y-axis.
- **b** represents the **slope** of the regression line, indicating the rate of change in y for a unit change in x.
- **e** represents the **error or residual**, which accounts for the difference between the observed values of y and the values predicted by the regression equation.

FINDING THE LINE

- Regression analysis relies on the concept of the regression line or curve, which represents the best-fit relationship between the variables involved.

- This line or curve is determined by minimizing the differences between the observed values of the dependent variable and the values predicted by the regression model.

- The least squares method is commonly employed to find this best-fit line or curve. This method minimizes the squared differences between observed and predicted values, ensuring that the regression line captures the overall trend or pattern in the data as accurately as possible.

- Through the least squares method, regression analysis yields estimate of the regression coefficients that define the best-fit relationship between the variables.

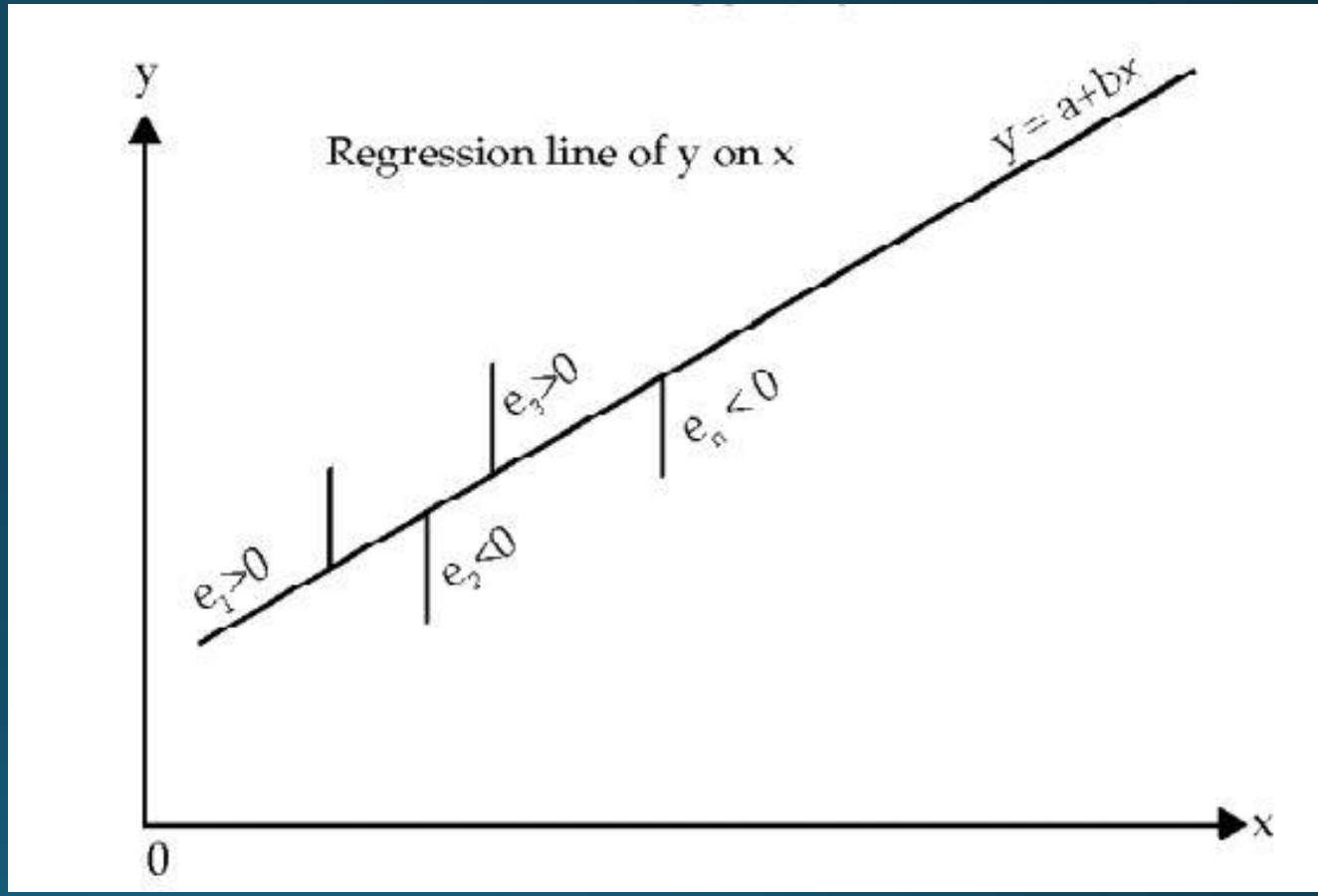
- These coefficients allow for making predictions about the dependent variable based on the values of the independent variable(s) with greater accuracy and reliability.

- As a result, this is widely used in regression analysis.

Properties of the Regression line:

- The line minimizes the sum of squared difference between the observed values (actual y-value) and the predicted value (ŷ value)

- The line passes through the mean of independent and dependent features.



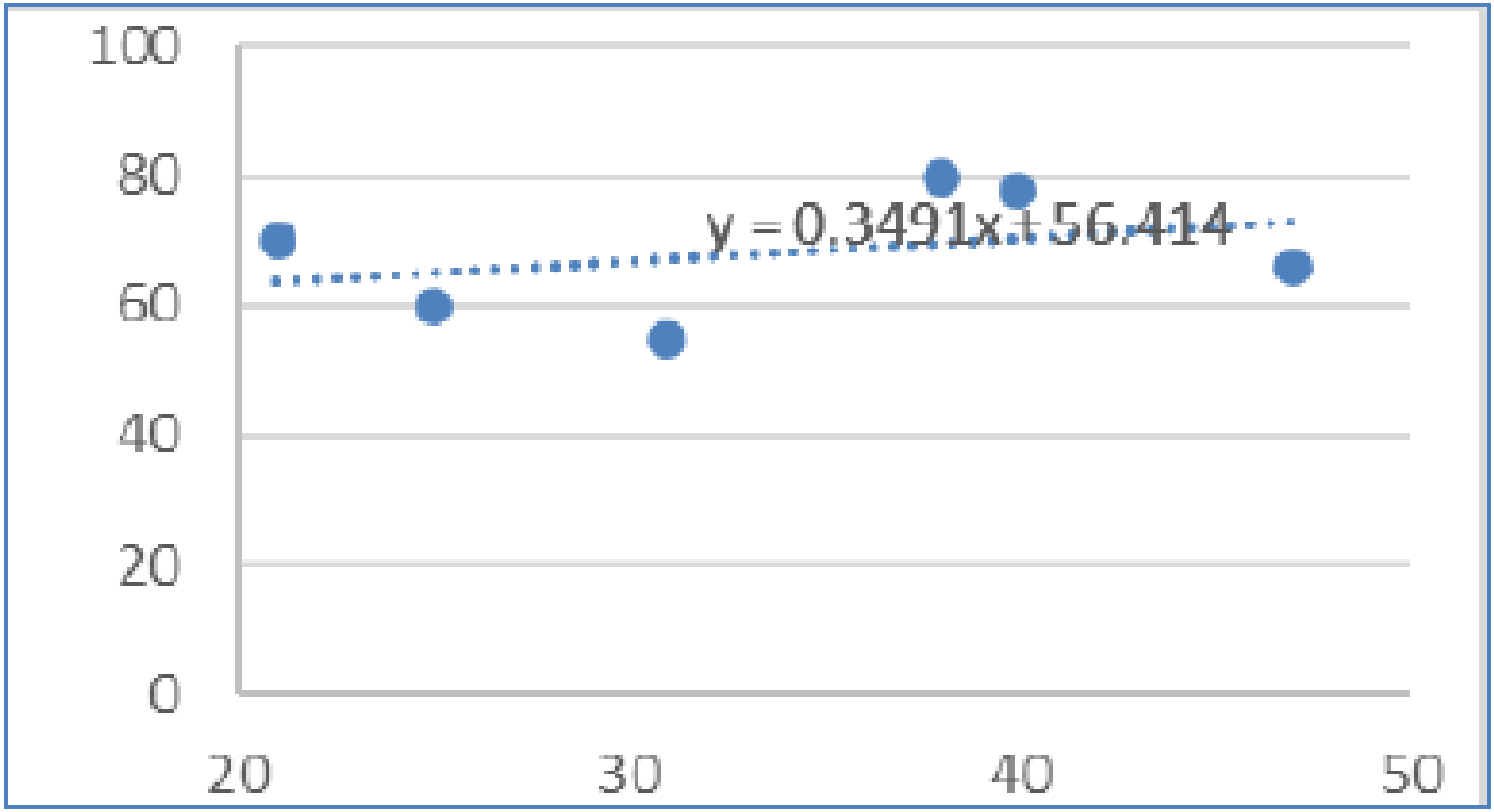
Example 1

In the example of 6 people with different ages and different weight, let us draw the line of best fit in Excel.

Sr. No	Age (x)	Weight (y)
1	40	78
2	21	70
3	25	60
4	31	55
5	38	80
6	47	66

Solution:

Step 1: Select the Age and Weight.
Step 2: Insert a scatter chart and make changes to the following:
Trendline Name: Linear, check Display Equation on Chart
X axis minimum: 20
Step 3: Let us verify the values of slope and intercept using slope() and intercept() function in excel.
Step 4: Click on any cell and type =slope (Now, select the values of Weight, and then type comma. Now, select the values of Age and press enter.
Step 5: Click on any cell and type =intercept (Now, select the values of Weight, and then type comma. Now, select the values of Age and press enter.


100
80
60
40
20
0
20
30
40
50
y = 0.3491x + 56.414

Some of the regression algorithms include

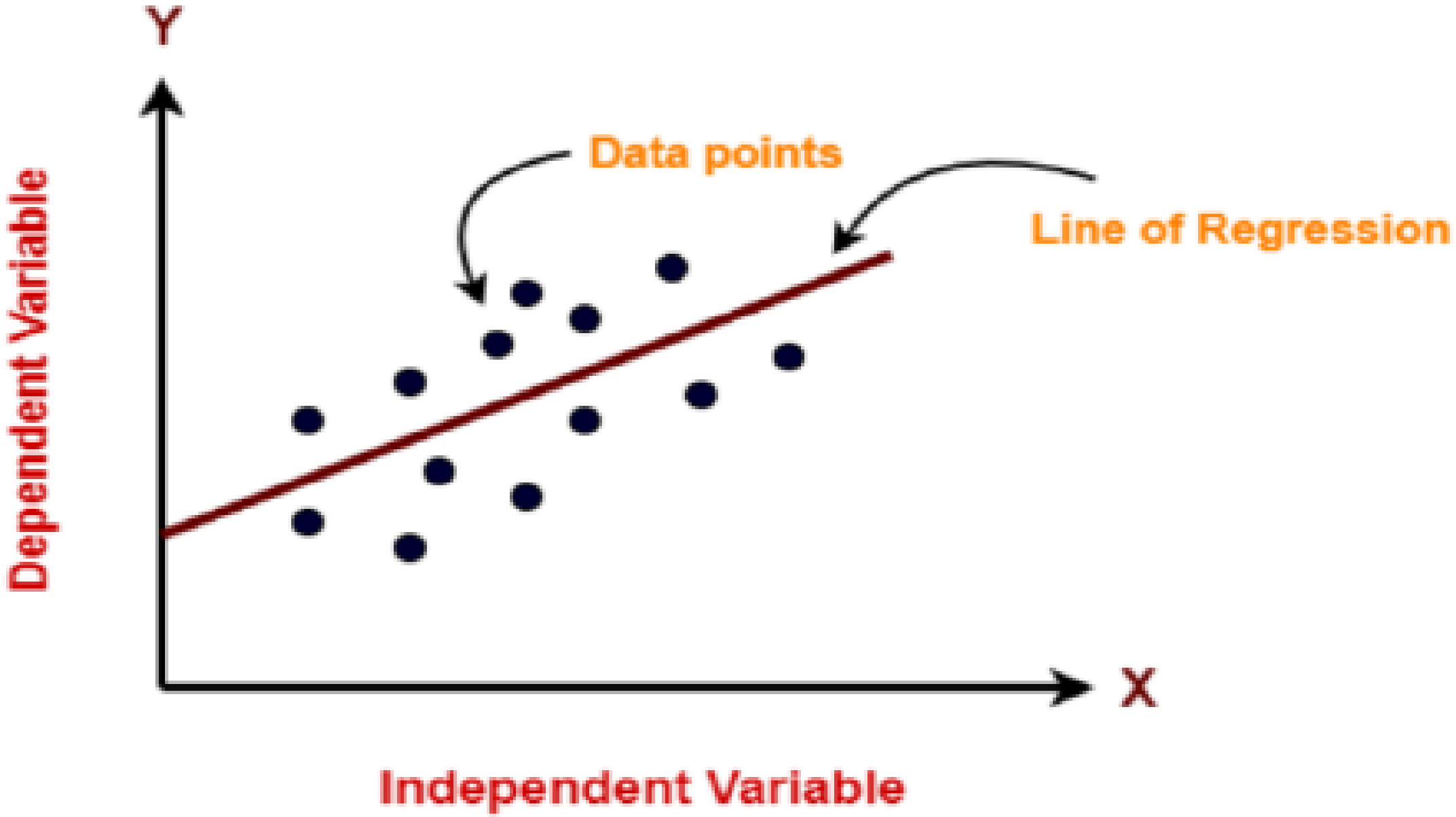
- Linear Regression,
- Logistic Regression,
- Decision Tree Regression,
- Random Forest Regression.

Linear Regression

The linear regression model consists of a predictor variable and a dependent variable related linearly to each other.
In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

a) Simple Linear Regression: The dependent variable's value is predicted using a single independent variable in simple linear regression.

b) Multiple Linear Regression: In multiple linear regression, more than one independent variable is used to predict the value of the dependent variable.


Y
Data points
Line of Regression
Dependent Variable
X
Independent Variable

Applications of Linear Regression:

- Market Analysis: Linear regression helps understand how different factors like pricing, sales quantity, advertising, and social media engagement relate to each other in the market.

- Sales Forecasting: It predicts future sales by analyzing past sales data along with factors like marketing spending, seasonal trends, and consumer behavior.

- Predicting Salary Based on Experience: Linear regression estimates a person's salary based on their years of experience, education, and job role, aiding in recruitment and compensation planning.

- Sports Analysis: Linear regression analyzes player and team performance by considering statistics, game conditions, and opponent strength, assisting coaches and team management in decision-making.

- Medical Research: Linear regression examines relationships between factors like age, weight, and health outcomes, helping researchers identify risk factors and evaluate interventions.

Advantages of Linear regression

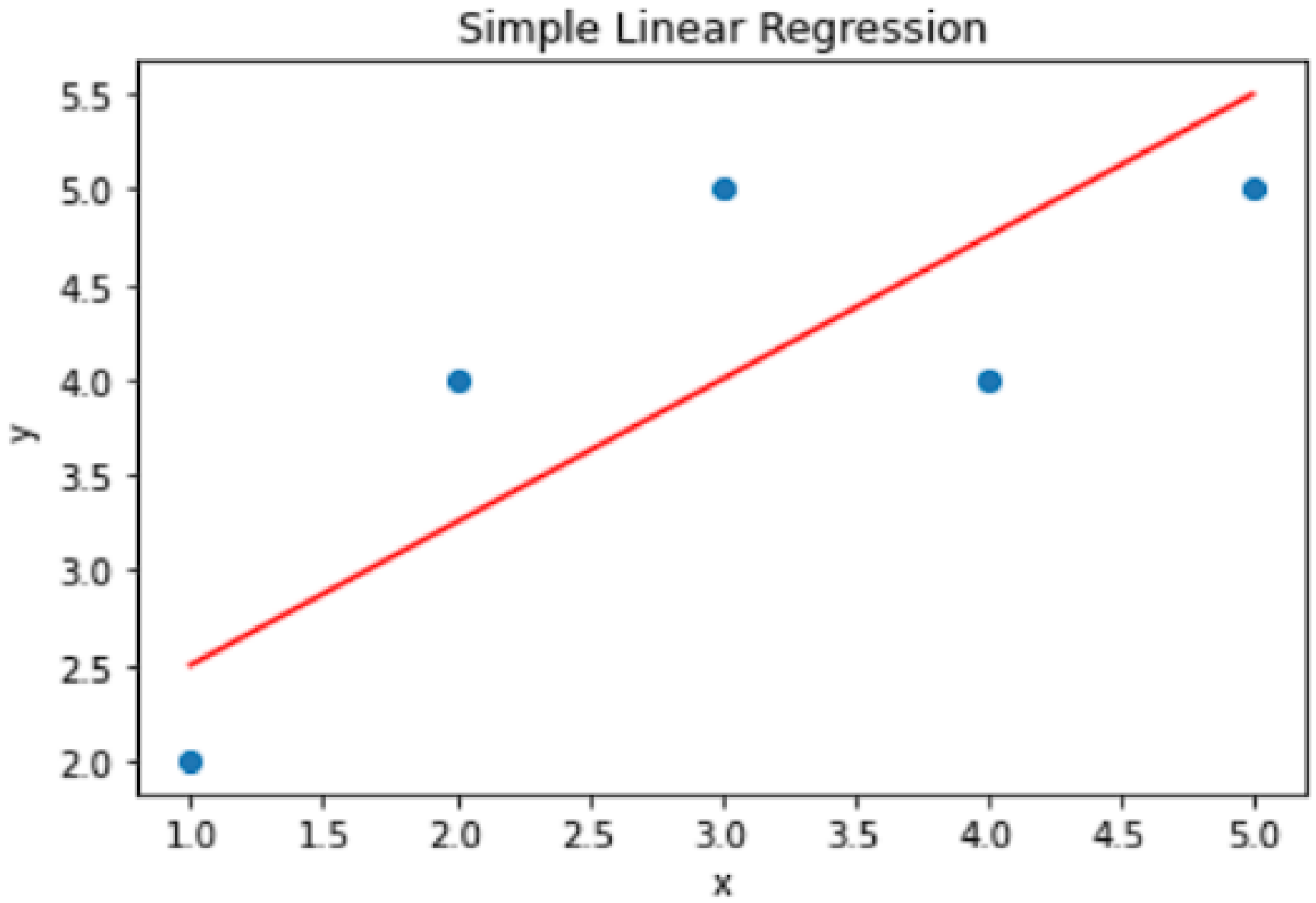
- Simple technique and easy to implement
- Efficient to train the machine on this model

Disadvantages of Linear regression

1. Sensitivity to outliers, which can significantly impact the analysis.
2. Limited to linear relationships between variables.

This program:

- Imports numpy for numerical calculations and matplotlib.pyplot for plotting.
- Defines sample data for x and y. You can replace this with your own data.
- Calculates mean, standard deviation, covariance, and slope.
- Calculates y-intercept based on slope and mean.
- Predicts y values for given x using the linear equation.
- Plots the data points and the regression line.
- Prints the estimated slope and intercept values.




```
Slope: 0.75
Intercept: 1.75
```

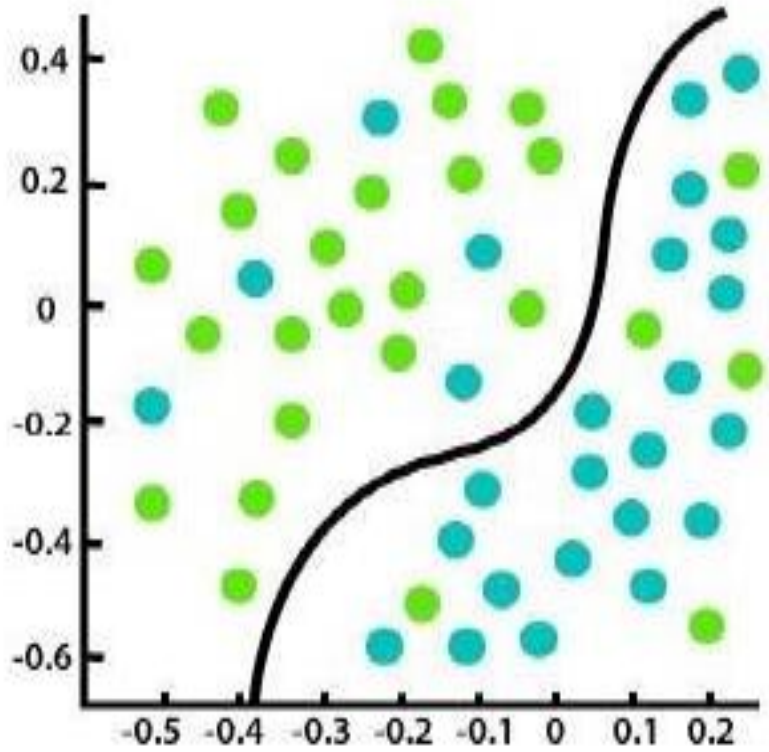
2. CLASSIFICATION

The main objective of classification is to assign labels to data instances based on their features or attributes.

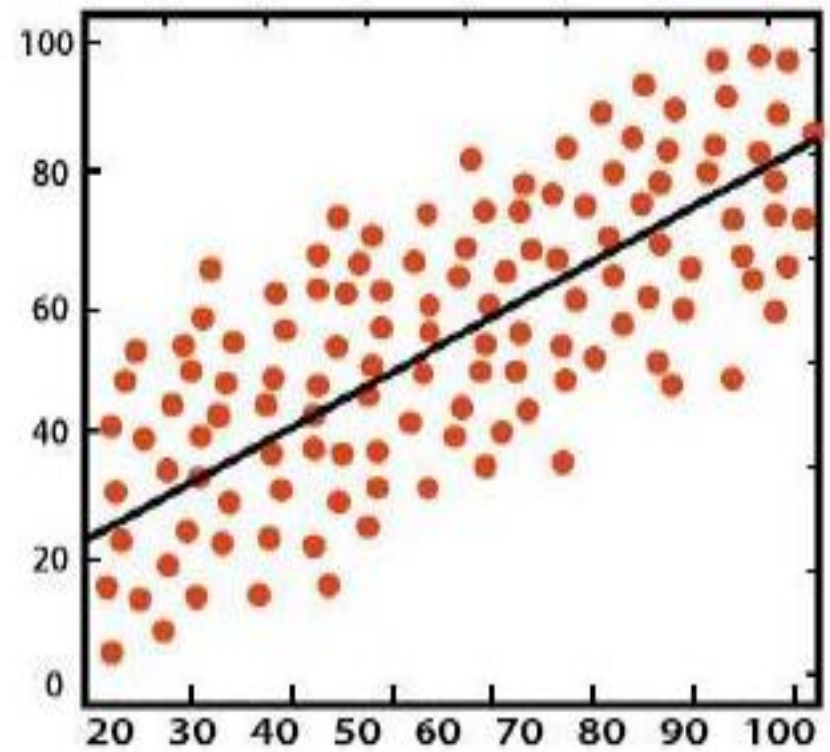
In classification, the data is typically labeled with class labels or categories, and the goal is to build a model that can accurately assign these labels to new, unseen data instances.

Look at the two graphs below and suggest which graph represents the classification problem.

Graph 1



Graph 2



_____ _____

How Classification Works

In classification tasks within machine learning, the process revolves around categorizing data into distinct groups or classes based on their features. Here is how it typically works:

- Classes or Categories: Data is divided into different classes or categories, each representing a specific outcome or group. For example, in a binary classification scenario, there are two classes: positive and negative.

- Features or Attributes: Each data instance is described by its features or attributes, which provide information about the instance. These features are crucial for the classification model to differentiate between different classes. For instance, in email classification, features might include words in the email text, sender information, and email subject.

- Training Data: The classification model is trained using a dataset known as training data. This dataset consists of labelled examples, where each data instance is associated with a class label. The model learns from this data to understand the relationship between the features and the corresponding class labels.
- Classification Model: An algorithm or technique is used to build the classification model. This model learns from the training data to predict the class labels of new, unseen data instances. It aims to generalize from the patterns and relationships in the training data to make accurate predictions.
- Prediction or Inference: Once trained, the classification model is used to predict the class labels of new data instances. This process, known as prediction or inference, relies on the learned patterns and relationships from the training data.

Types of classification

The four main types of classification are:

1) Binary Classification

2) Multi-Class Classification

3) Multi-Label Classification

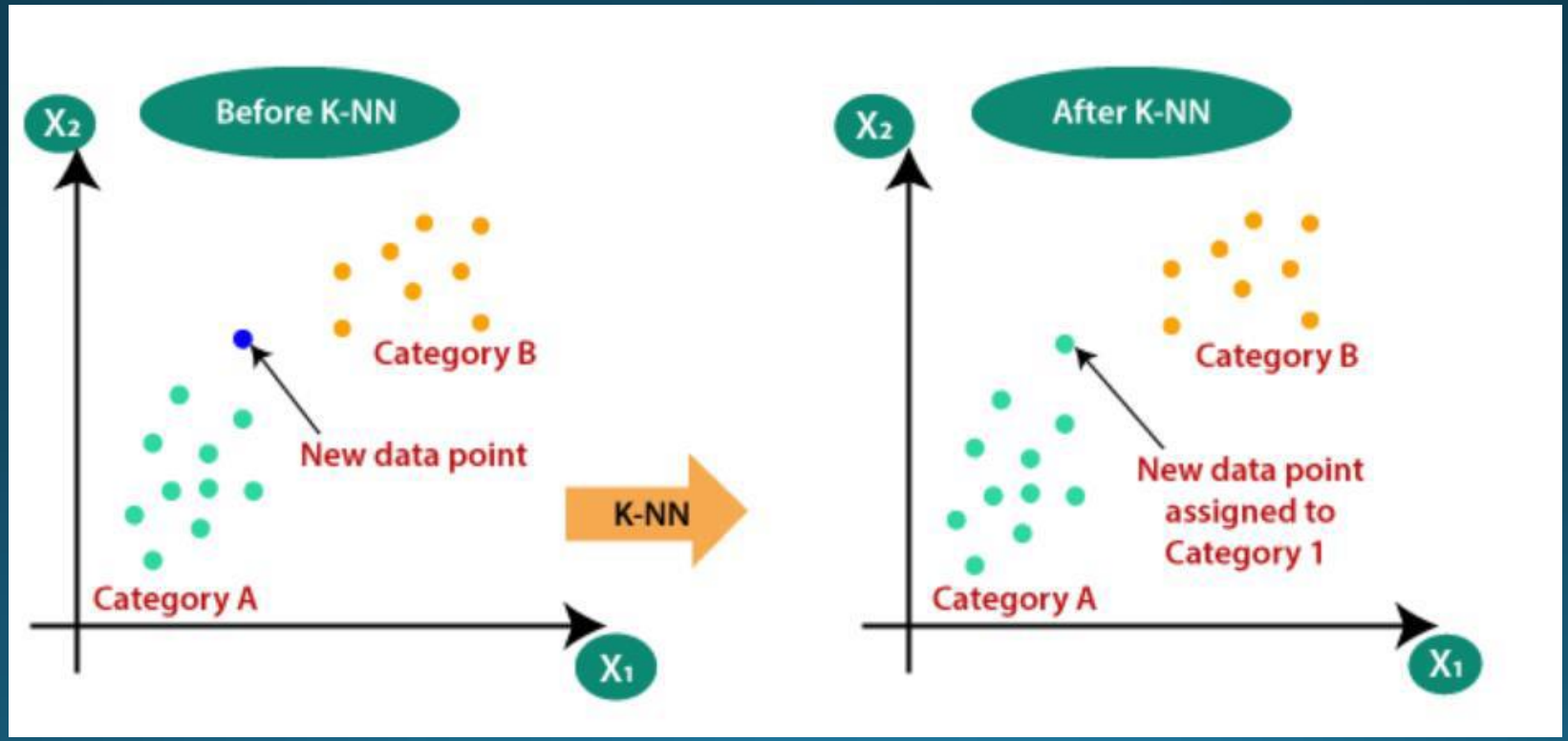
4) Imbalanced Classification

Classification Type	Binary Classification	Multi-Class Classification	Multi-Label Classification	Imbalanced Classification
Description	Classification tasks with two class labels.	Classification tasks with more than two class labels.	Classification tasks where each example may belong to multiple class labels.	Classification tasks with unequally distributed class labels, typically with a majority and minority class.
Example	• Email spam detection - spam or not • Conversion prediction - buy or not • Medical test - Cancer detected or not • Exam results - pass/fail	• Face classification • Plant species classification • Optical character recognition • Image classification into thousands of classes	• Photo classification - objects present in the photo (bicycle, apple, person, etc.)	• Fraud detection • Outlier detection • Medical diagnostic tests

K- Nearest Neighbour algorithm (KNN)

The K-Nearest Neighbors algorithm, commonly known as KNN or k-NN, is a versatile non-parametric supervised learning technique used for both classification and regression tasks. It operates based on the principle of proximity, making predictions or classifications by considering the similarity between data points.

Why KNN Algorithm is Needed: KNN is particularly useful when dealing with classification problems where the decision boundaries are not clearly defined or when the dataset does not have a well-defined structure. It provides a simple yet effective method for identifying the category or class of a new data point based on its similarity to existing data points.



Steps involved in k-NN

- Select the number K of the neighbors
- Calculate the Euclidean distance of K number of neighbors
- Take the K nearest neighbors as per the calculated Euclidean distance.
- Among these k neighbors, count the number of the data points in each category.
- Assign the new data points to that category for which the number of the neighbor is maximum.
- Our model is ready.

Applications of KNN:

- Image recognition and classification
- Recommendation systems
- Healthcare diagnostics
- Text mining and sentiment analysis
- Anomaly detection

Advantages of KNN:

- Easy to implement and understand.
- No explicit training phase; the model learns directly from the training data.
- Suitable for both classification and regression tasks.
- Robust to outliers and noisy data.

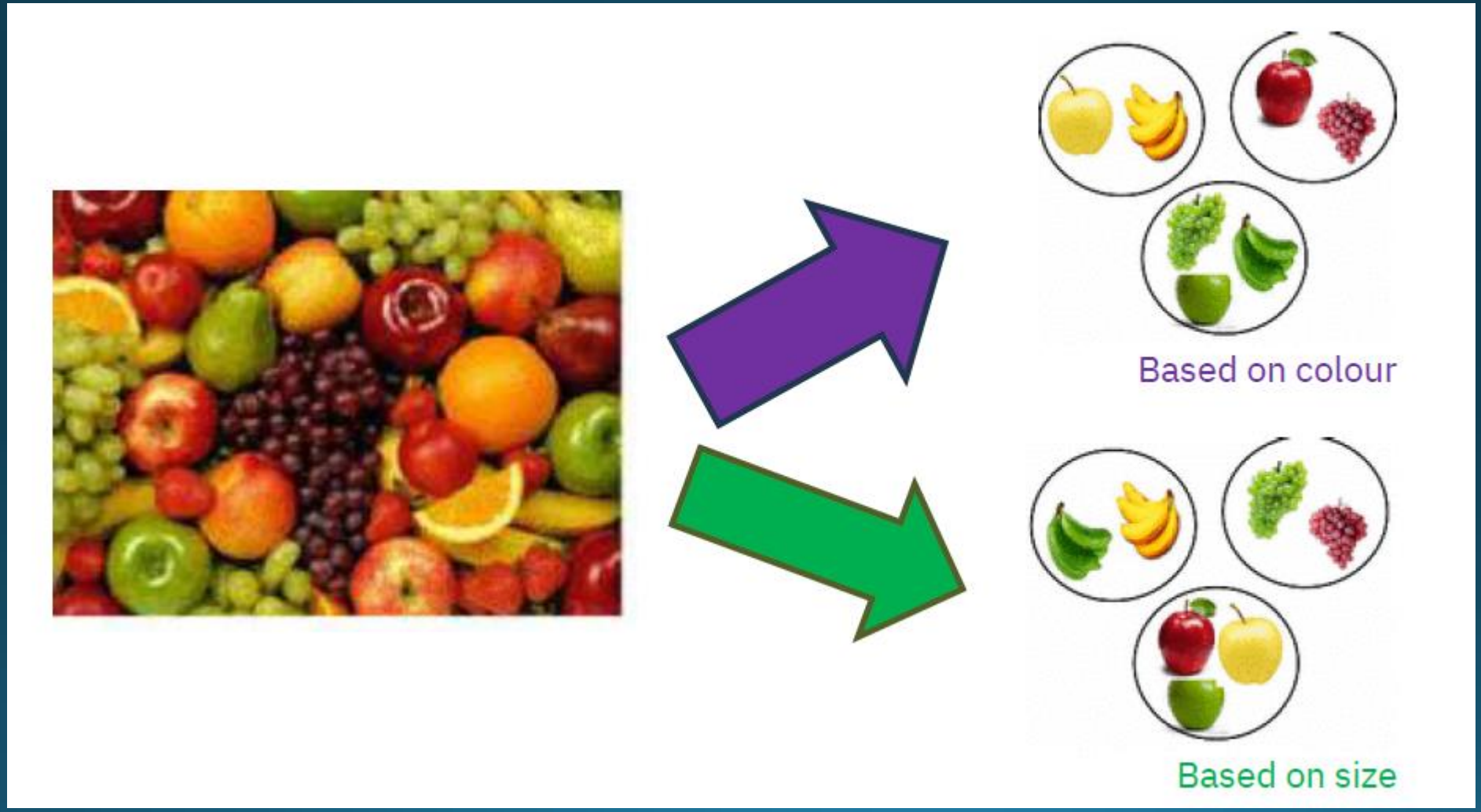
Limitations of KNN:

- Computationally expensive, especially for large datasets.
- Sensitivity to the choice of distance metric and the number of neighbors (K).
- Requires careful preprocessing and feature scaling.
- Not suitable for high-dimensional data due to the curse of dimensionality.

B. UNSUPERVISED LEARNING

3. CLUSTERING

Clustering, or cluster analysis, is a machine learning technique used to group unlabeled dataset into clusters or groups based on similarity.



- It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

- It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset. The clustering technique is commonly used for statistical data analysis.

How Clustering Works: To cluster data effectively, follow these key steps:

1) Prepare the Data: Select the right features for clustering and make sure the data is ready by scaling or transforming it as needed.

2) Create Similarity Metrics: Define how similar data points are by comparing their features. This similarity measure is crucial for clustering.

3) Run the Clustering Algorithm: Apply a clustering algorithm to group the data. Choose one that works well with your dataset size and characteristics.

4) Interpret the Results: Analyze the clusters to understand what they represent. Since clustering is unsupervised, interpretation is essential for assessing the quality of the clusters.

Types of Clustering Methods Some of the common clustering methods used in Machine learning are:

1) Partitioning Clustering

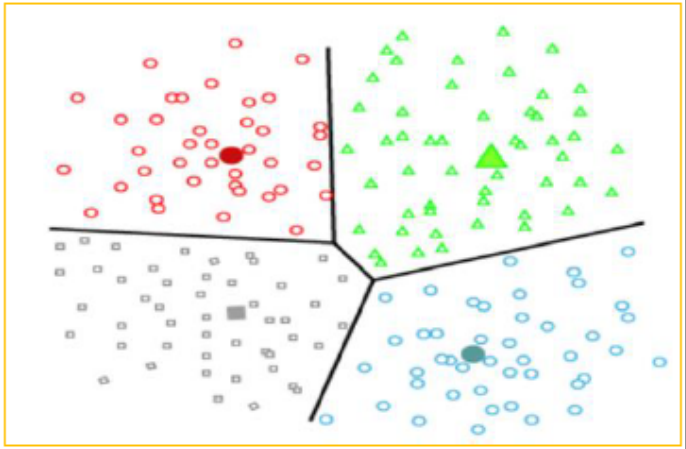
2) Density-Based Clustering

3) Distribution Model-Based Clustering

4) Hierarchical Clustering

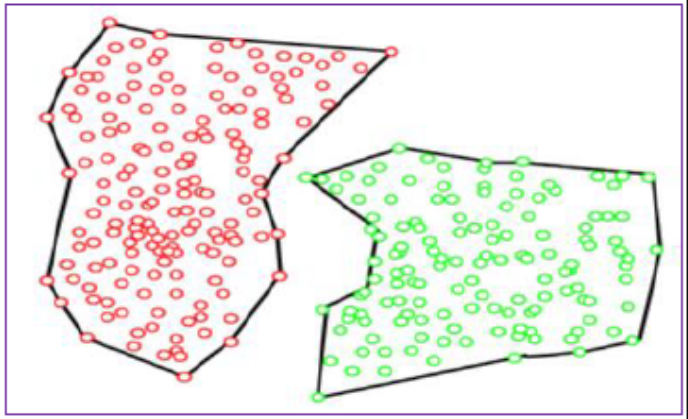
1. Partitioning Clustering

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the **centroid-based method**. The most common example of partitioning clustering is the K-Means Clustering algorithm. In this type, the dataset is divided into a set of k groups, where k is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



2. Density-Based Clustering

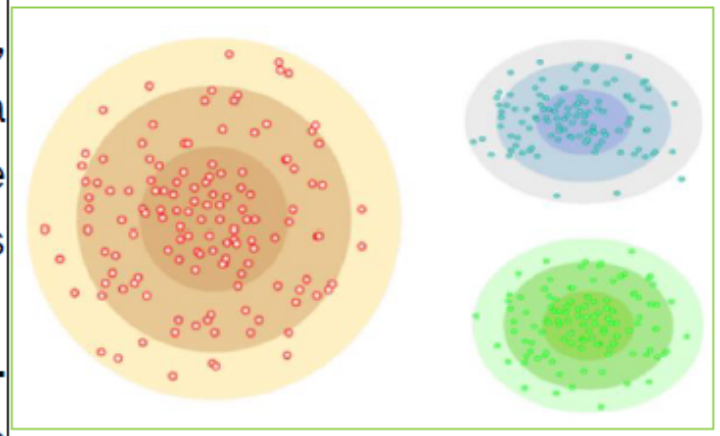
The density-based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas. These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.



3. Distribution Model-Based Clustering

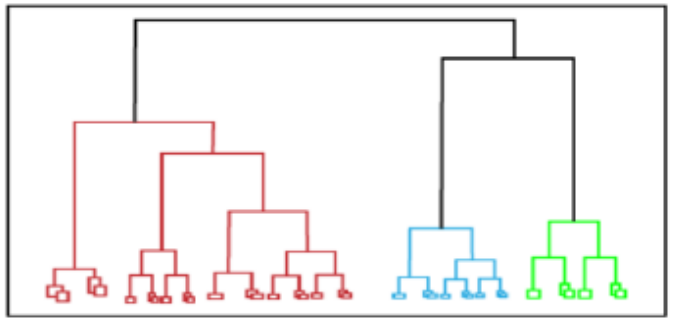
In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly **Gaussian Distribution**.

The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).



4. Hierarchical Clustering

Hierarchical clustering can be used as an alternative for the partitioned clustering as there is no requirement of pre-specifying the number of clusters to be created. In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**. The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm.**



K Means Clustering

It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm.

Steps involved K-Means Clustering: The working of the K-Means algorithm is explained in the below steps:

- Select the number K to decide the number of clusters.
- Select random K points or centroids. (It can be other from the input dataset).
- Assign each data point to their closest centroid, which will form the predefined K clusters.
- Calculate the variance and place a new centroid of each cluster.
- Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- If any reassignment occurs, then go to step-4 else go to FINISH.
- The model is ready.

Applications of K-Means Clustering:

- Market Segmentation: group customers based on similar purchasing behaviours or demographics for tailored marketing strategies.
- Image Segmentation: partition images into regions of similar colours to aid in tasks like object detection and compression.
- Document Clustering: categorize documents based on content similarity, aiding in organization and information retrieval.
- Anomaly Detection: identify outliers by clustering normal data points and detecting deviations.
- Customer Segmentation: segment customers for targeted marketing and personalized experiences.

Advantages of K-Means Clustering:

- Easy to implement, making it suitable for users of all levels.
- Handles large datasets with low computational resources.
- Works well with numerous features and data points.
- Are easy to understand, aiding in decision-making.
- Applicable across various domains and data types.

Limitations of K-Means Clustering:

- Results can vary based on initial centroid placement.
- Assumes clusters are spherical, which is not always true.
- Number of clusters must be known beforehand.
- Outliers can distort clusters due to their influence on centroids.
- May converge to suboptimal solutions instead of the global optimum.

This program does the following:

1. Generates synthetic data using make_blobs from sklearn.datasets.

2. Applies K-means clustering with n_clusters=4.

3. Plots the data points colored by their cluster assignments and shows the centroids as red circles.

You can adjust the parameters like the number of clusters, standard deviation, and number of samples in make_blobs to observe different clustering scenarios.

THANK YOU