Unit-IV
Supervised Machine Learning Models

Computer Department
AVPTI Rajkot

Unit Outcomes

- Define Supervised Learning
- List types of Supervised Learning, Describe K-Nearest Neighbour and Simple linear regression
- Advantage and disadvantage of supervised machine learning

Topics

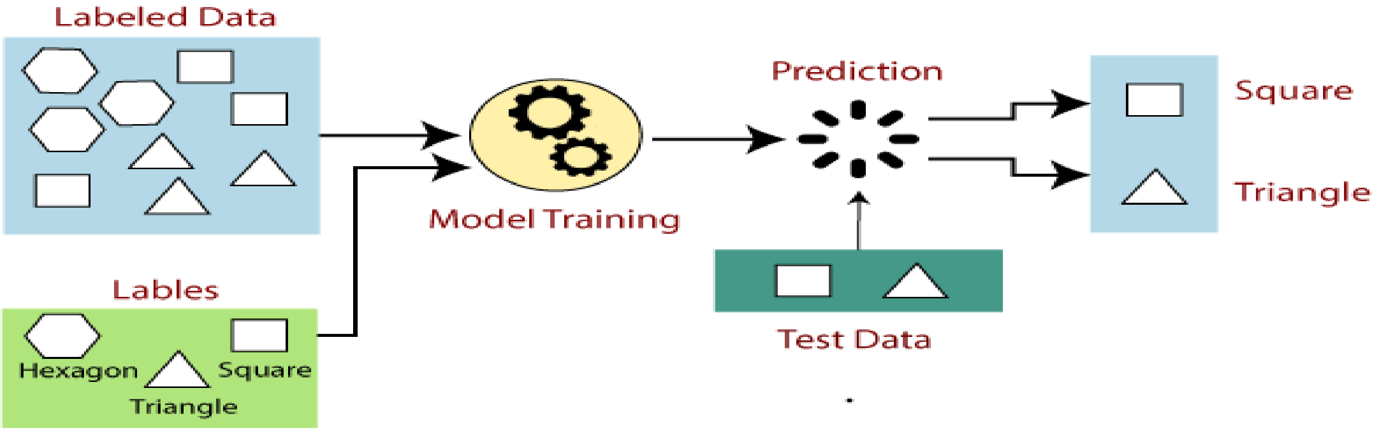
- Introduction of Supervised Learning
- Types of Supervised Learning :Classification and Regression (Algorithm KNN and Linear regression)
- Advantage and disadvantage of supervised machine learning

Supervised Machine Learning

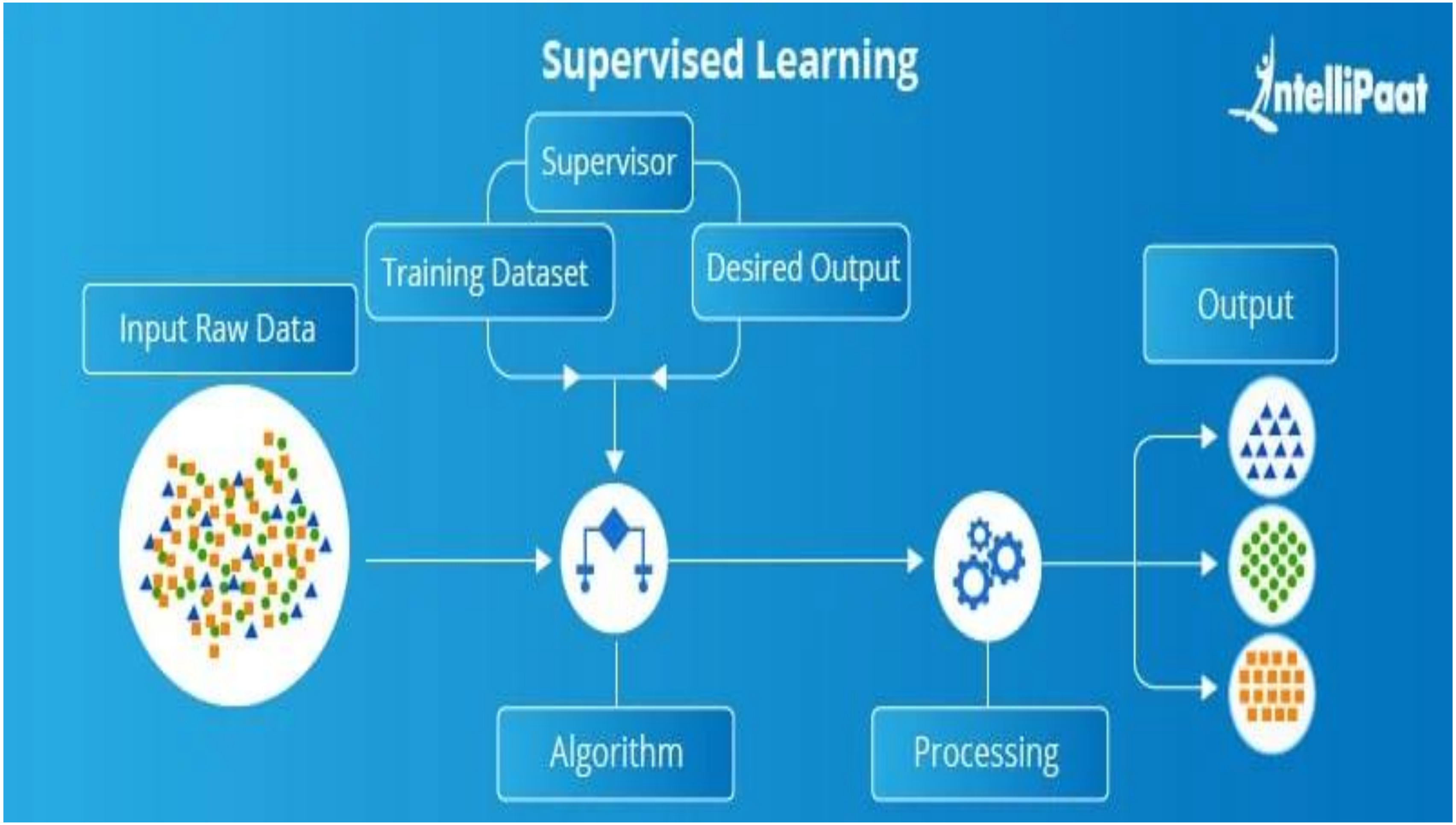
- Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.
- In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
- Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

How Supervised Learning Works?

- In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.



Continue…


Supervised Learning
IntelliPaat
Supervisor
Training Dataset
Desired Output
Input Raw Data
Output
Algorithm
Processing

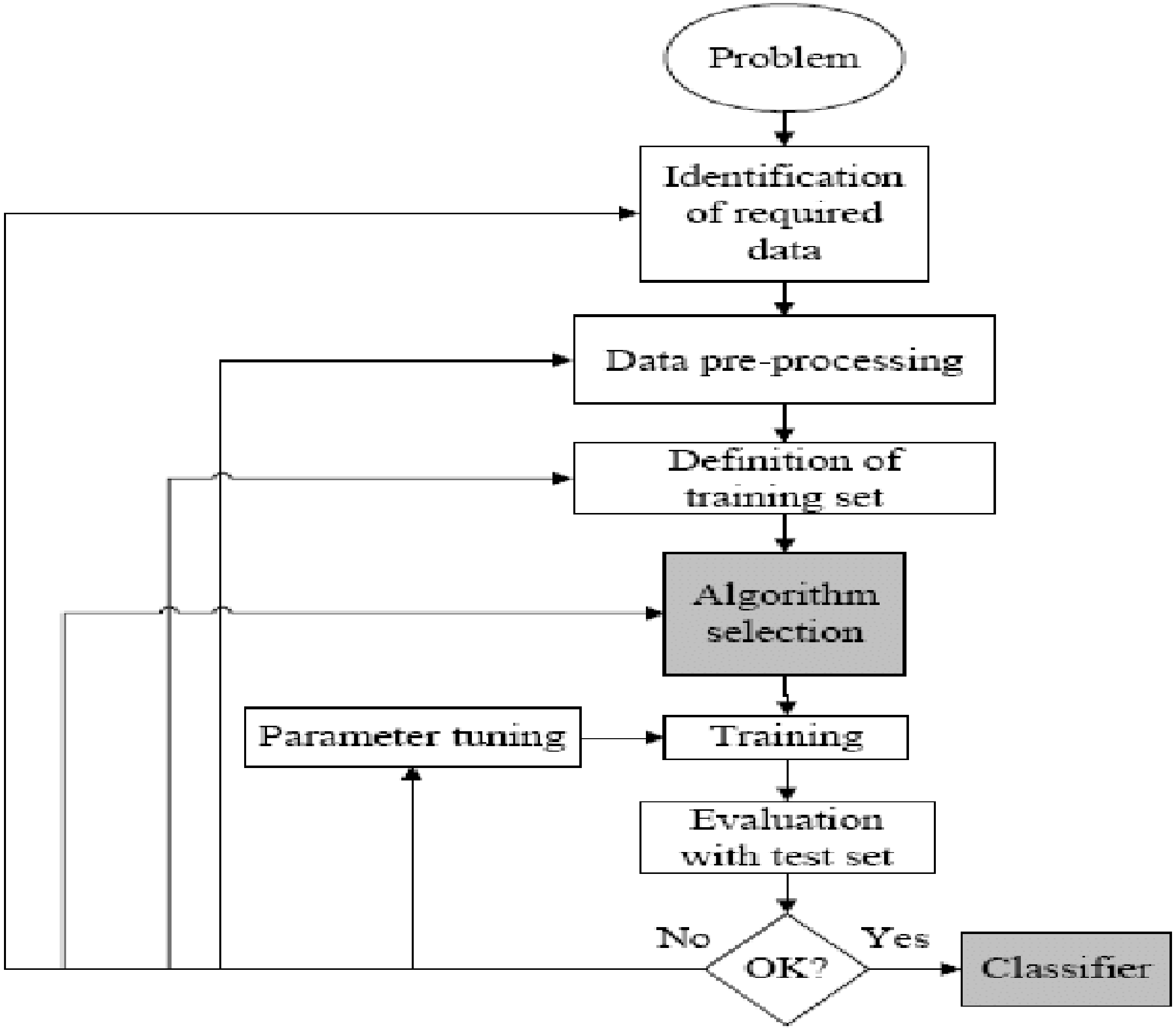
Applications of Supervised Learning

- **Image Segmentation**: Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
- **Medical Diagnosis:** Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labeled data with labels for disease conditions. With such a process, the machine can identify a disease for the new pat
- **Fraud Detection**: Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
- **Spam detection**: In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
- **Speech Recognition**: Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

Steps in Supervised Machine learning

- First Determine the type of training dataset
- Collect/Gather the labelled training data.
- Split the training dataset into training **dataset, test dataset, and validation dataset**.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
- Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

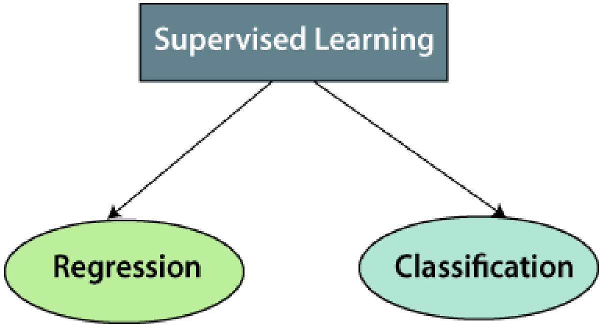
Continue…


Problem
Identification of required data
Data pre-processing
Definition of training set
Algorithm selection
Parameter tuning
Training
Evaluation with test set
No
OK?
Yes
Classifier

Advantages and Disadvantages of Supervised Learning

- **Advantages:**
  - Since supervised learning work with the labeled dataset so we can have an exact idea about the classes of objects.
  - These algorithms are helpful in predicting the output on the basis of prior experience.
- **Disadvantages:**
  - These algorithms are not able to solve complex tasks.
  - It may predict the wrong output if the test data is different from the training data.
  - It requires lots of computational time to train the algorithm.

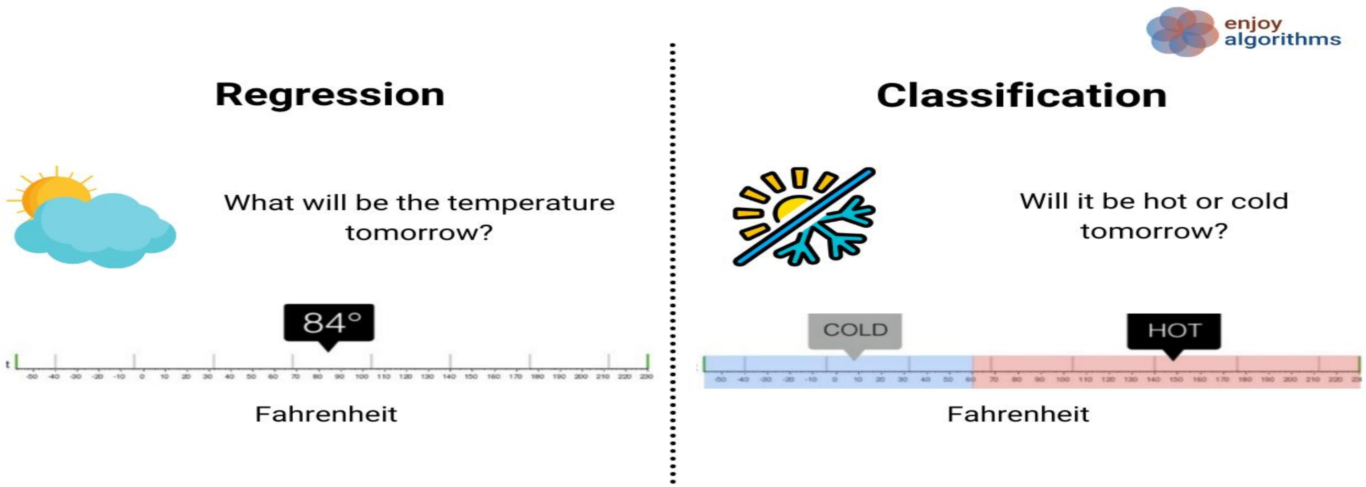
Types of supervised Machine learning Algorithms

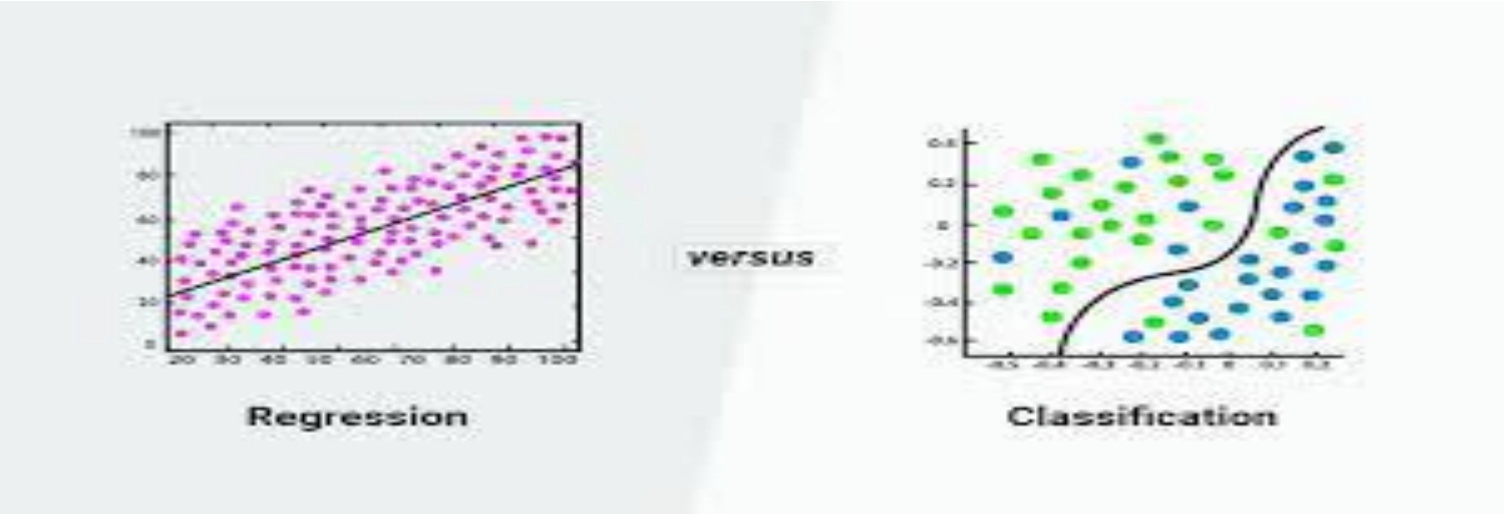


- Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

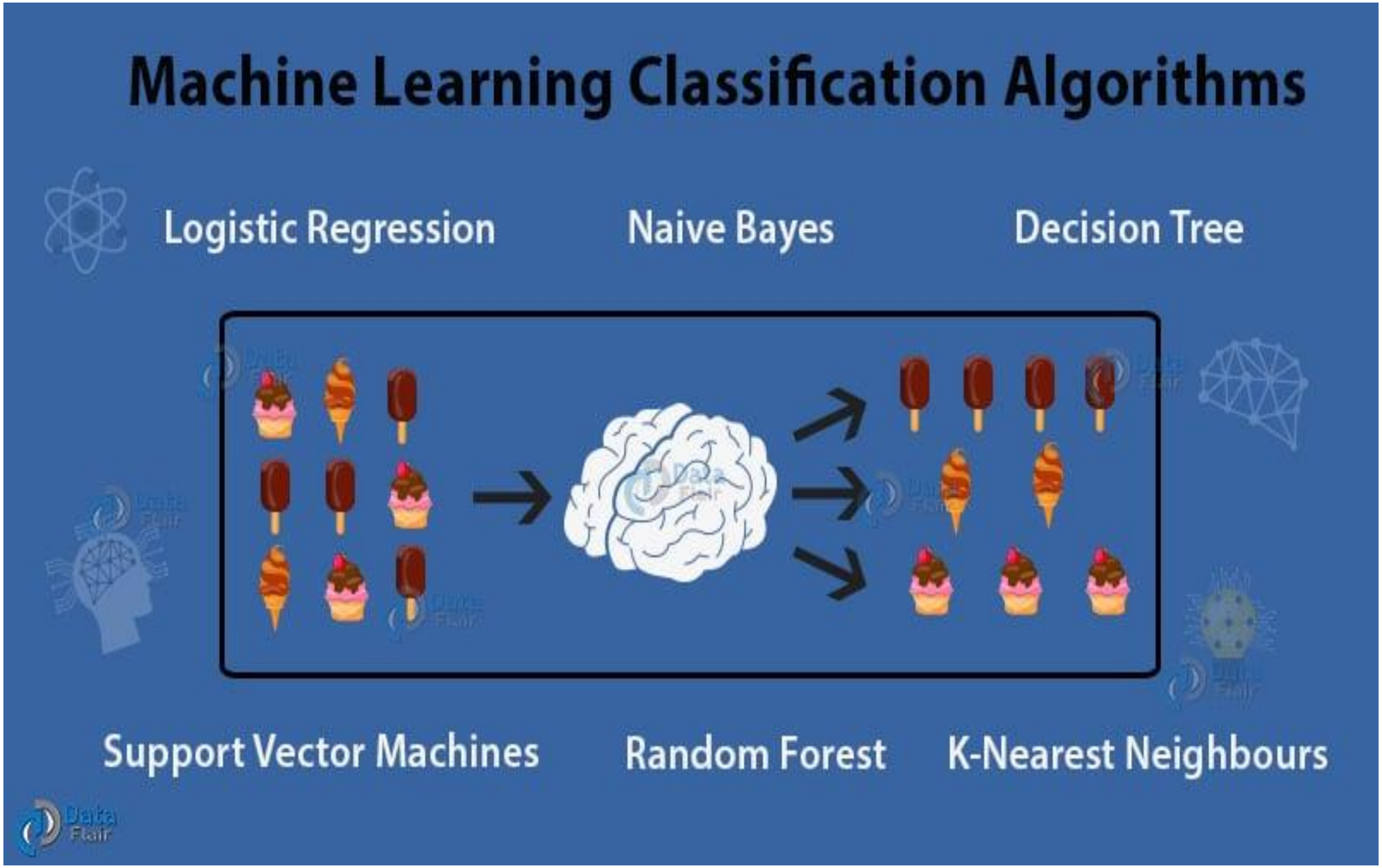
- Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Continue…

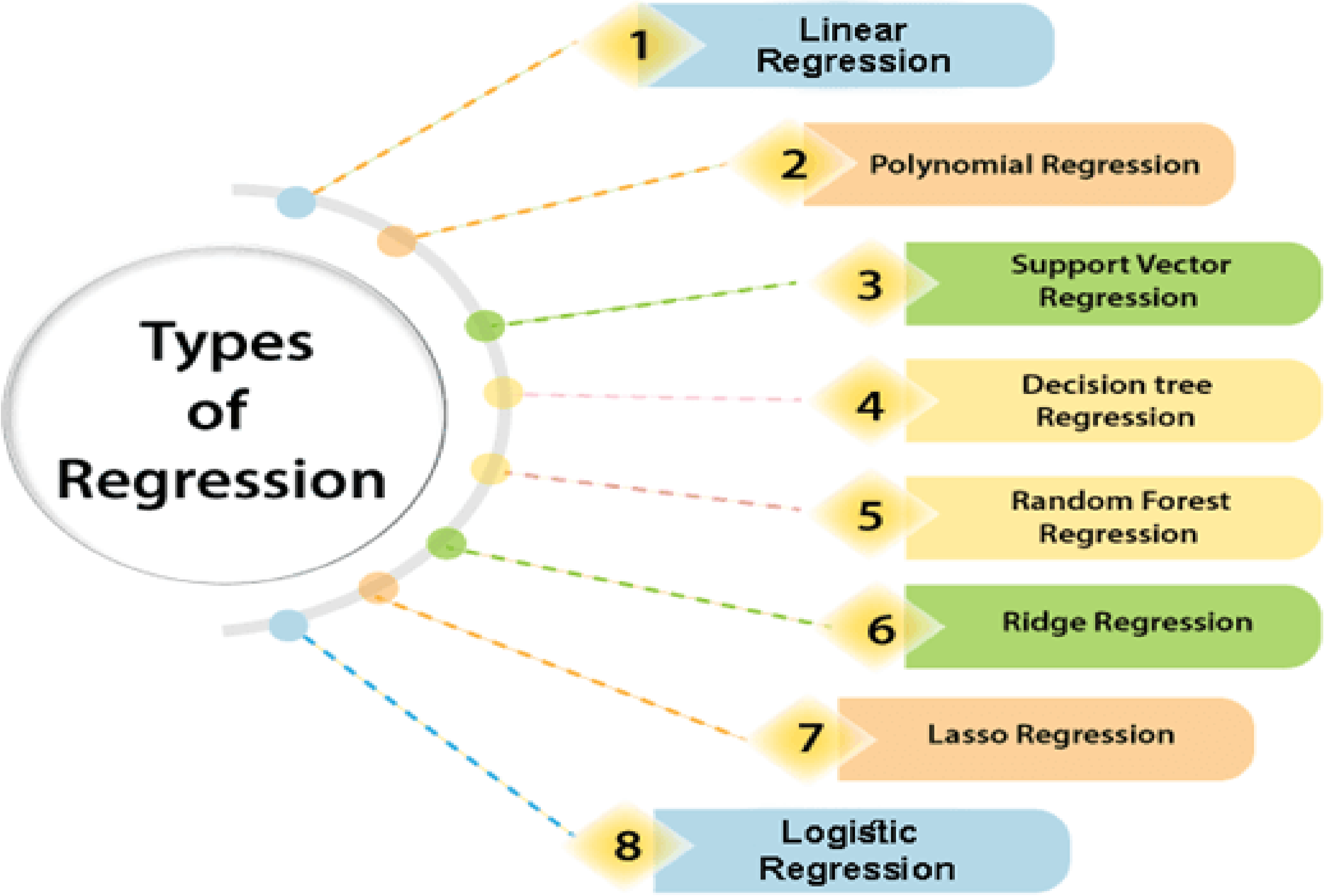

enjoy algorithms
Regression
What will be the temperature tomorrow?
84°
Fahrenheit
Classification
Will it be hot or cold tomorrow?
COLD
HOT
Fahrenheit


Regression
versus
Classification

Continue…


Machine Learning Classification Algorithms
Logistic Regression
Naive Bayes
Decision Tree
Support Vector Machines
Random Forest
K-Nearest Neighbours
Data Flair

Continue…


Types of Regression
1
Linear Regression
2
Polynomial Regression
3
Support Vector Regression
4
Decision tree Regression
5
Random Forest Regression
6
Ridge Regression
7
Lasso Regression
8
Logistic Regression

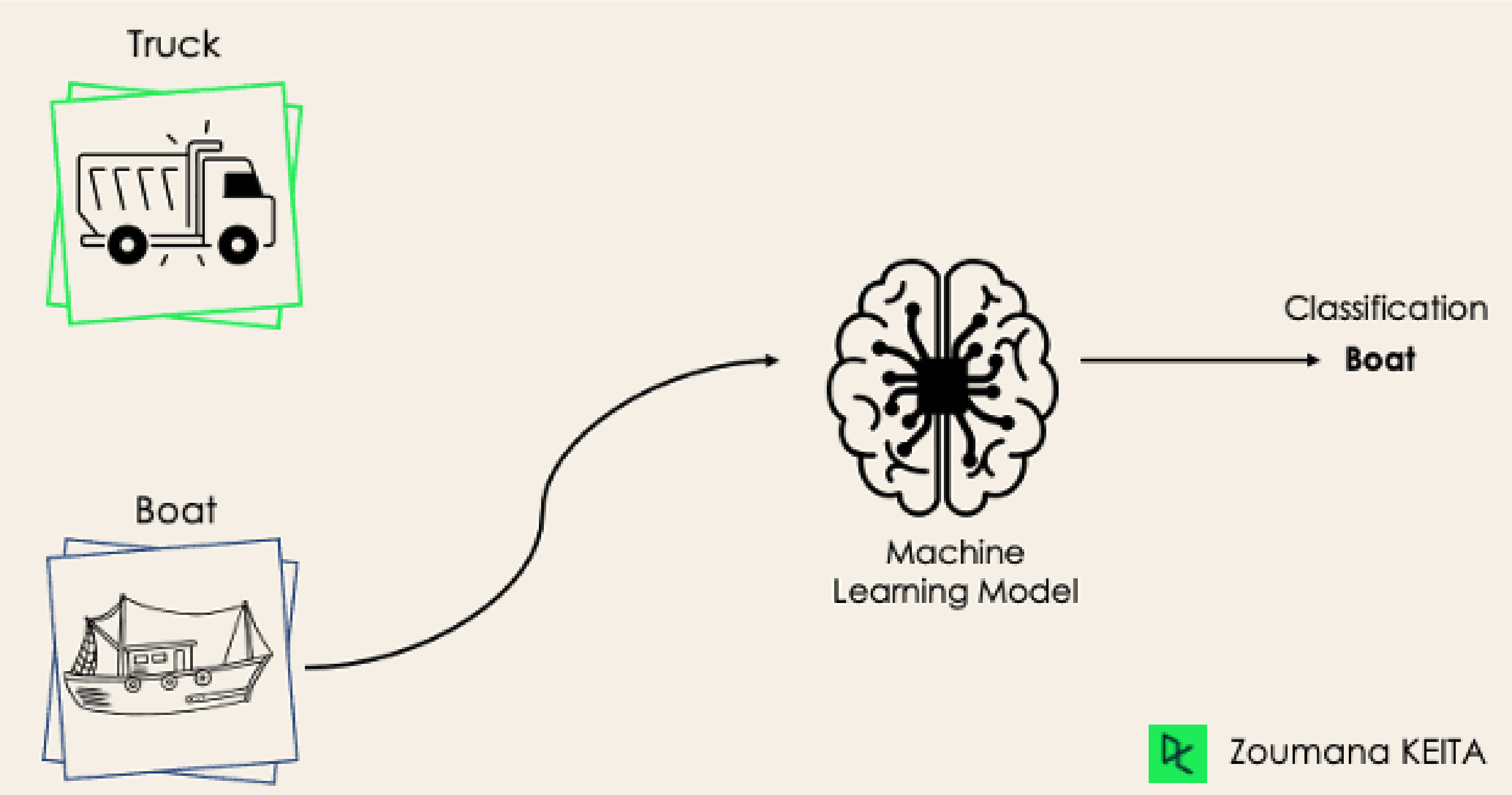
Classification

- Classification is defined as the process of recognition, understanding, and grouping of objects into preset categories.
- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

Types of classification

- **Binary Classification**
- In a binary classification task, the goal is to classify the input data into two mutually exclusive categories.
- The training data in such a situation is labeled in a binary format: true and false; positive and negative; O and 1; spam and not spam, etc. depending on the problem being tackled.
- For instance, we might want to detect whether a given image is a truck or a boat.

Continue…

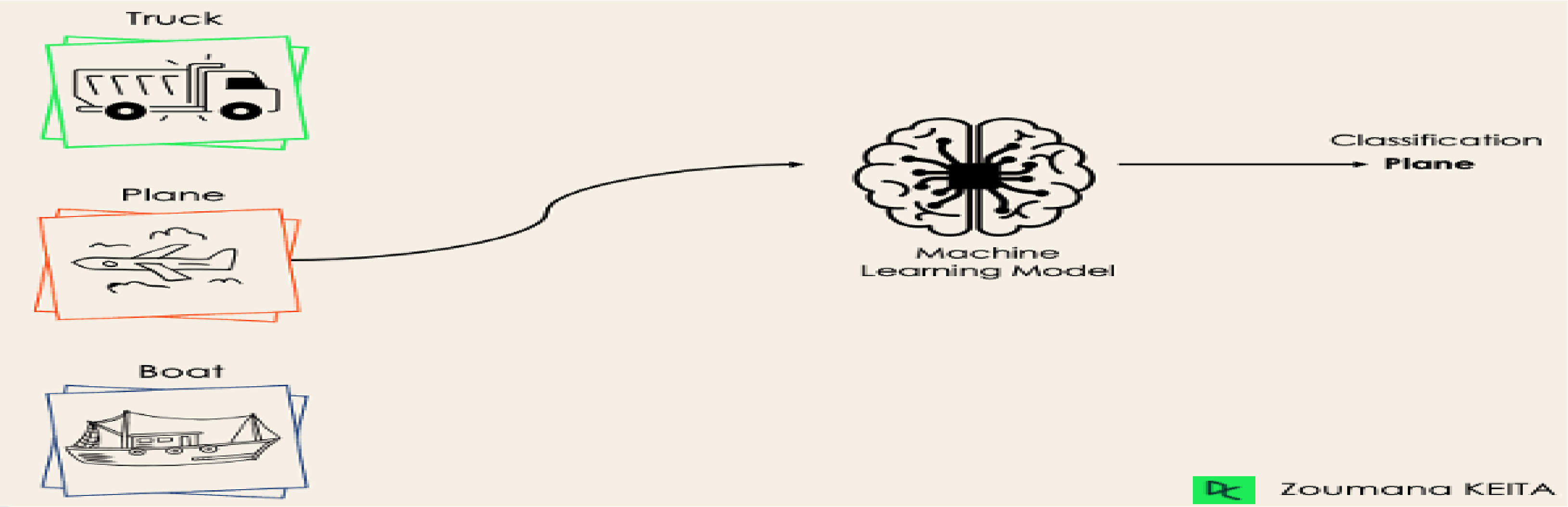

Truck
Boat
Machine
Learning Model
Classification
Boat
Zoumana KEITA

Continue…

- Popular algorithms that can be used for binary classification include:
- Logistic Regression
- **k-Nearest Neighbors**
- Decision Trees
- Support Vector Machine
- Naive Bayes

Multi-Class Classification

- It refers to those classification tasks that have more than two class labels where the goal is to predict to which class a given input example belongs to.
- In the following case, the model correctly classified the image to be a plane.



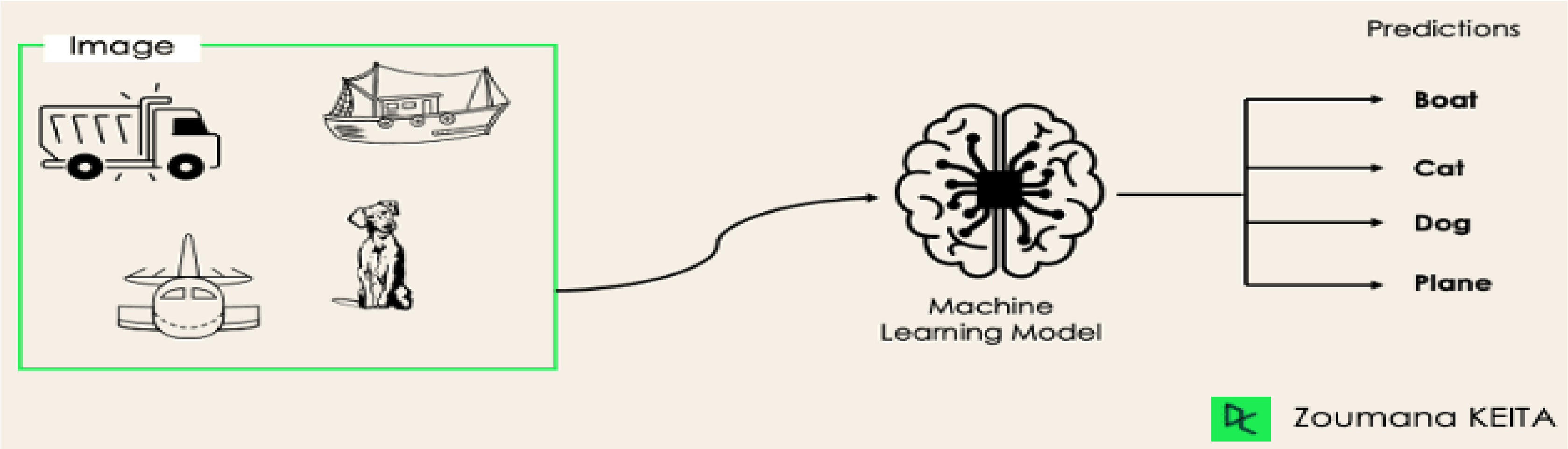
Continue…

- Most of the binary classification algorithms can be also used for multi-class classification. These algorithms include but are not limited to:
- Random Forest
- Naive Bayes
- K-Nearest Neighbors
- Gradient Boosting
- SVM
- Logistic Regression.

Multi-Label Classification

- It refers to those classification tasks that have two or more class labels, where one or more class labels may be predicted for each example.
- Consider the example of photo classification, where a given photo may have multiple objects in the scene and a model may predict the presence of multiple known objects in the photo, such as “*bicycle*,” “*apple*,” “*person*,” etc.

Continue…



Classification algorithms used for binary or multi-class classification cannot be used directly for multi-label classification. Specialized versions of standard classification algorithms can be used, so-called multi-label versions of the algorithms, including:

•Multi-label Decision Trees
•Multi-label Random Forests
•Multi-label Gradient Boosting

Imbalanced Classification

- For the imbalanced classification, the number of examples is unevenly distributed in each class, meaning that we can have more of one class than the others in the training data.
- Typically, imbalanced classification tasks are binary classification tasks where the majority of examples in the training dataset belong to the normal class and a minority of examples belong to the abnormal class.
- Using conventional predictive models such as Decision Trees, Logistic Regression, etc. could not be effective when dealing with an imbalanced dataset, because they might be biased toward predicting the class with the highest number of observations, and considering those with fewer numbers as noise.

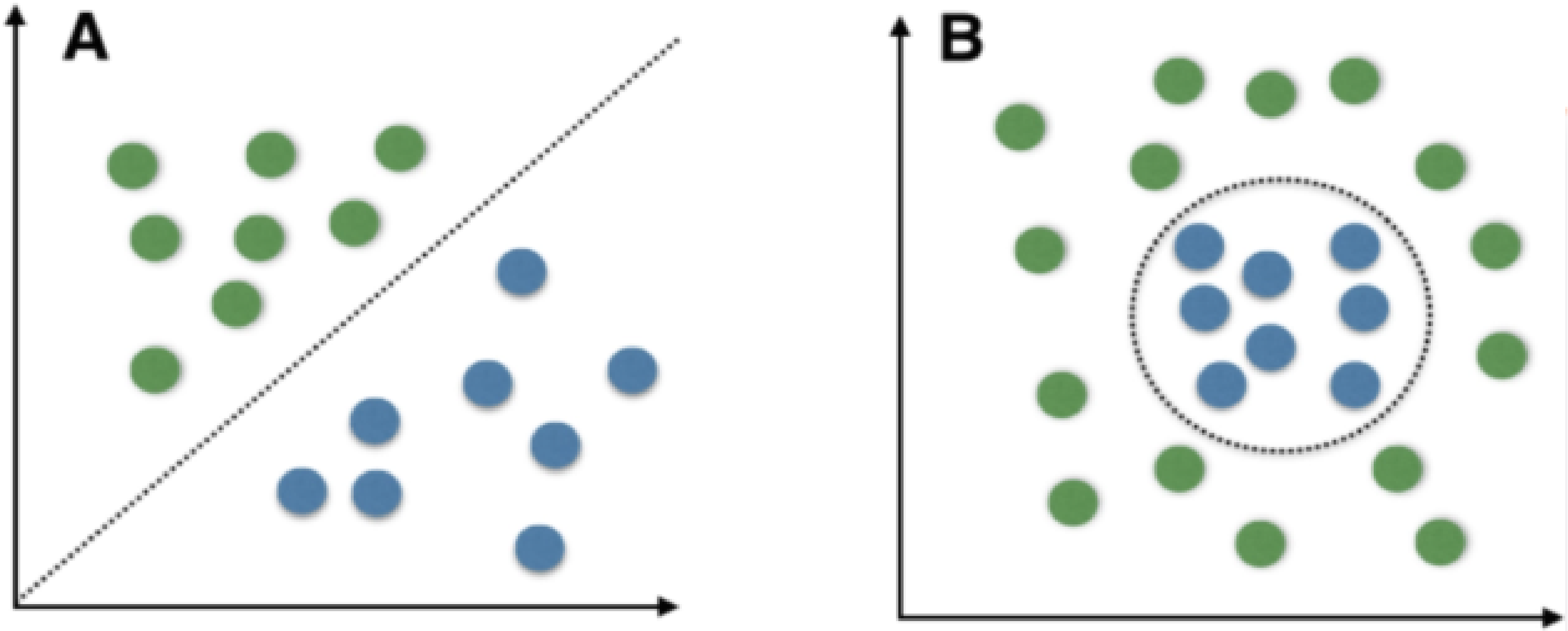
Continue…

- Specialized techniques may be used to change the composition of samples in the training dataset by under sampling the majority class or oversampling the minority class.
- Examples include:
- Random Under sampling…. *random elimination of examples from the majority class.*
- SMOTE Oversampling…. *random replication of examples from the minority class.*

Continue…

- Specialized modeling algorithms may be used that pay more attention to the minority class when fitting the model on the training dataset, such as cost-sensitive machine learning algorithms.
- Examples include:
- Cost-sensitive Logistic Regression.
- Cost-sensitive Decision Trees.
- Cost-sensitive Support Vector Machines.
- Finally, alternative performance metrics may be required as reporting the classification accuracy may be misleading.
- Examples include:
- Precision.
- Recall.
- F-Measure.

Types of ML Classification Algorithms



A: Linearly Separable Data B: Non-Linearly Separable Data

Continue…

- **Linear Models**
  - Logistic Regression
  - Support Vector Machines
- **Non-linear Models**
  - K-Nearest Neighbours
  - Kernel SVM
  - Naïve Bayes
  - Decision Tree Classification
  - Random Forest Classification

Use cases of Classification Algorithms

- Classification algorithms can be used in different places. Below are some popular use cases of Classification Algorithms:
- Email Spam Detection
- Speech Recognition
- Identifications of Cancer tumor cells.
- Drugs Classification
- Biometric Identification, etc.

K-Nearest Neighbour (K-NN)

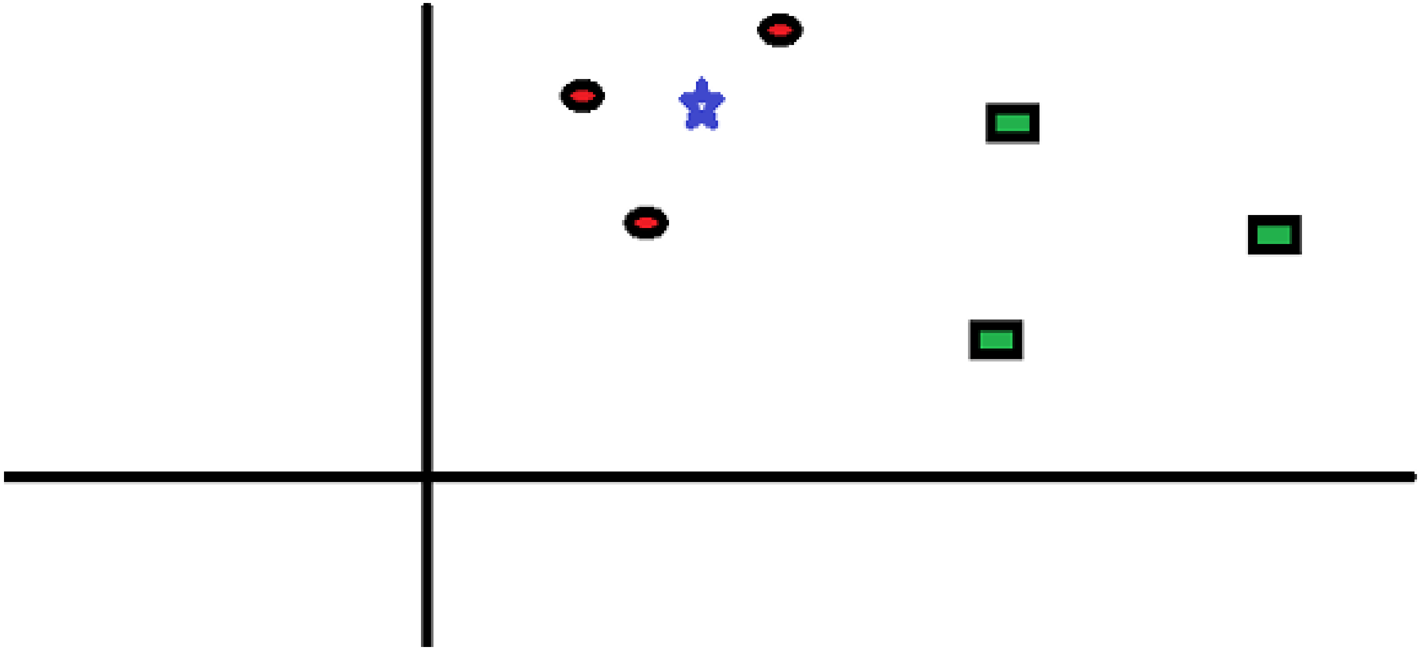
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Continue

- During the training phase, the KNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.
- Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances. In the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point.
- For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the input data point.

How Does the KNN Algorithm Work?

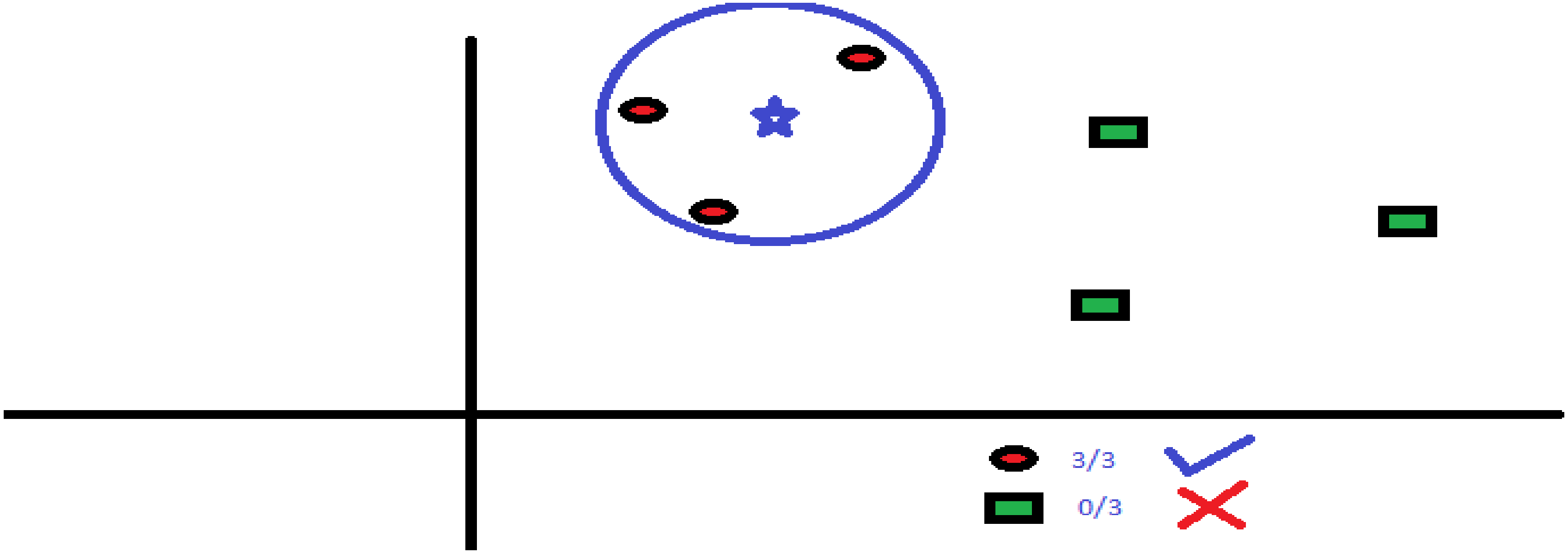
- Let’s take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS).



Continue…

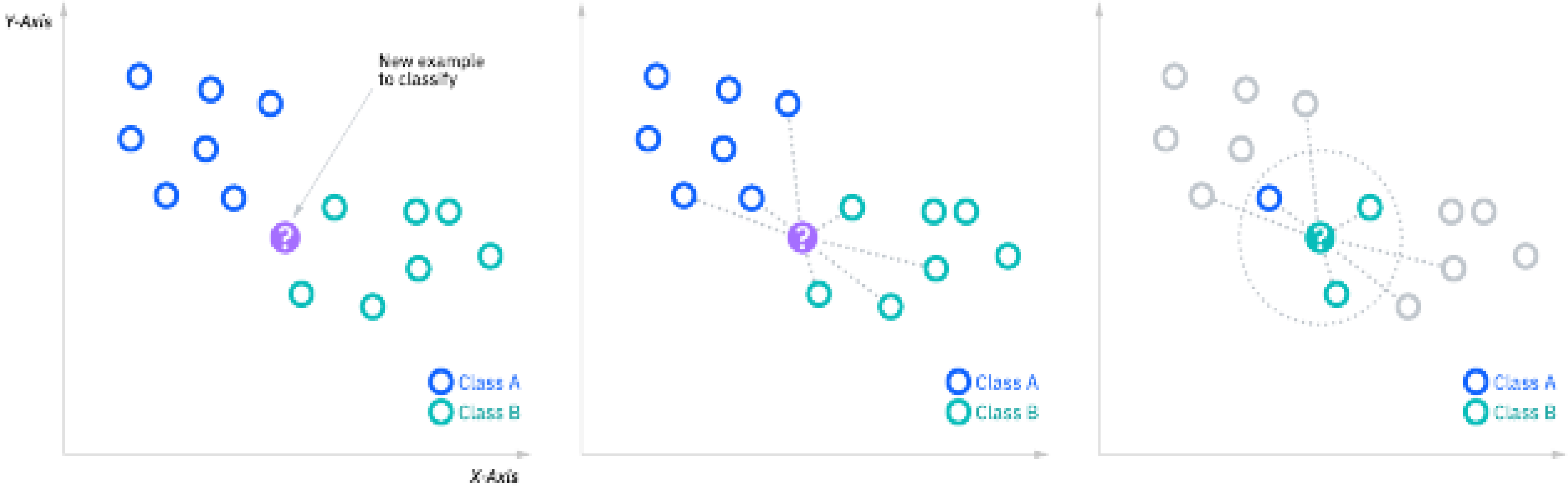
- You intend to find out the class of the blue star (BS). BS can either be RC or GS and nothing else.
- The “K” in KNN algorithm is the nearest neighbor we wish to take the vote from. Let’s say K = 3.
- Hence, we will now make a circle with BS as the center just as big as to enclose only three data points on the plane.

Continue…

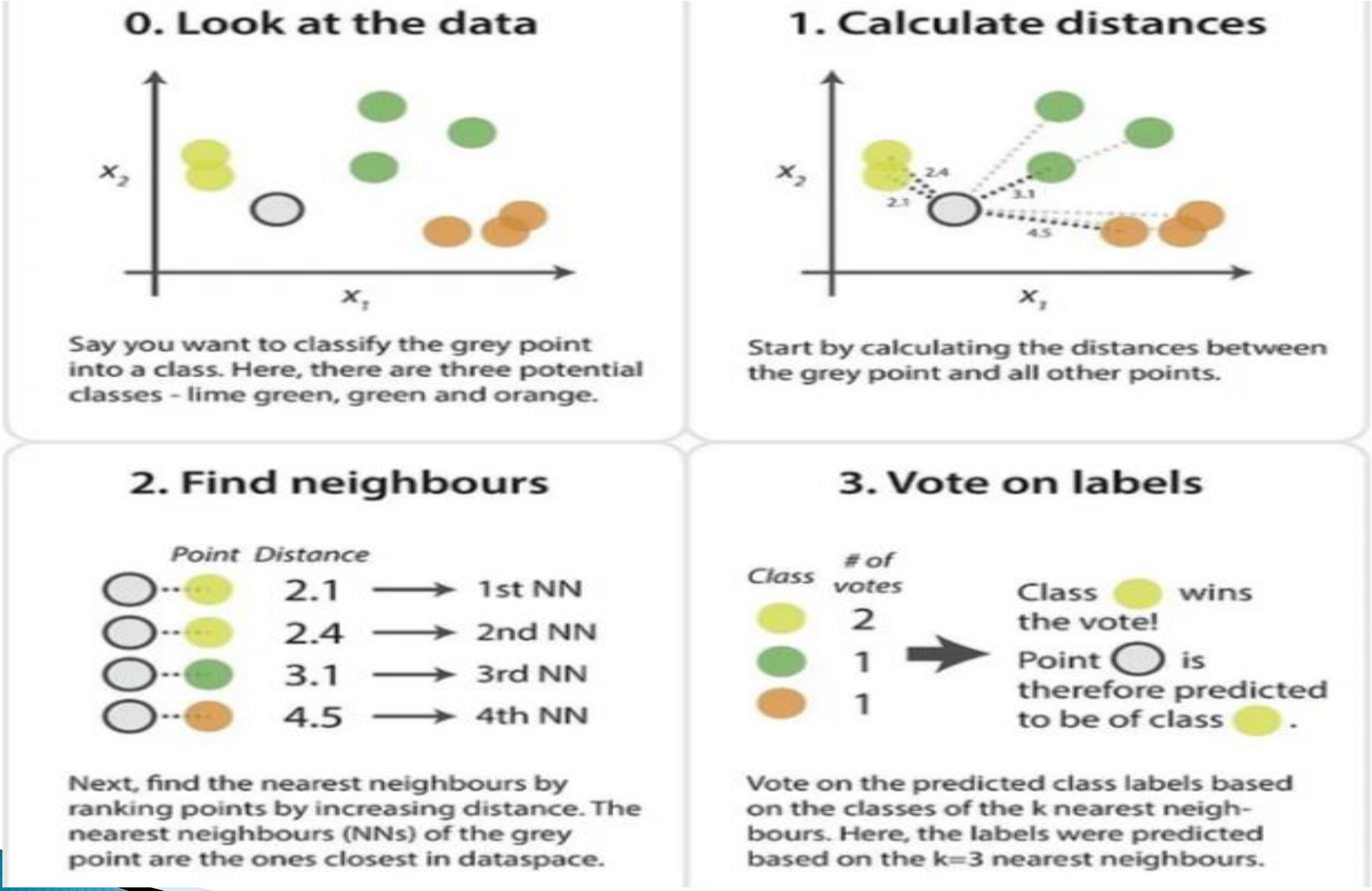


- The three closest points to BS are all RC. Hence, we can say that the BS should belong to the class RC.
- Here, the choice became obvious as all three votes from the closest neighbor went to RC. The choice of the parameter K is very crucial in this algorithm.

Continue…


Y-Axis
New example to classify
?
Class A
Class B
X-Axis
?
Class A
Class B
?
Class A
Class B

Continue...



0. Look at the data

Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances

Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Point	Distance	
○··🟡	2.1	→ 1st NN
○··🟡	2.4	→ 2nd NN
○··🟢	3.1	→ 3rd NN
○··🟠	4.5	→ 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels

Class	# of votes
🟡	2
🟢	1
🟠	1

Class 🟡 wins the vote! Point ○ is therefore predicted to be of class 🟡 .

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

How Do We Choose the Factor K?

- There is no straightforward method to calculate the value of K in KNN. You have to play around with different values to choose the optimal value of K. Choosing a right value of K is a process called Hyperparameter Tuning.
- The value of optimum K totally depends on the dataset that you are using. The best value of K for KNN is highly data-dependent. In different scenarios, the optimum K may vary. It is more or less hit and trail method.
- A small value of k means that noise will have a higher influence on the result and a large value make it computationally expensive.
- Usually choose as an odd number if the number of classes is 2 and another simple approach to select k is set k=sqrt(n).

Continue…

- There is no one proper method of estimation of K value in KNN. No method is the rule of thumb but you should try considering following suggestions:

  1. Square Root Method: Take square root of the number of samples in the training dataset.

  2. Cross Validation Method: We should also use cross validation to find out the optimal value of K in KNN. Start with K=1, run cross validation (5 to 10 fold), measure the accuracy and keep repeating till the results become consistent.

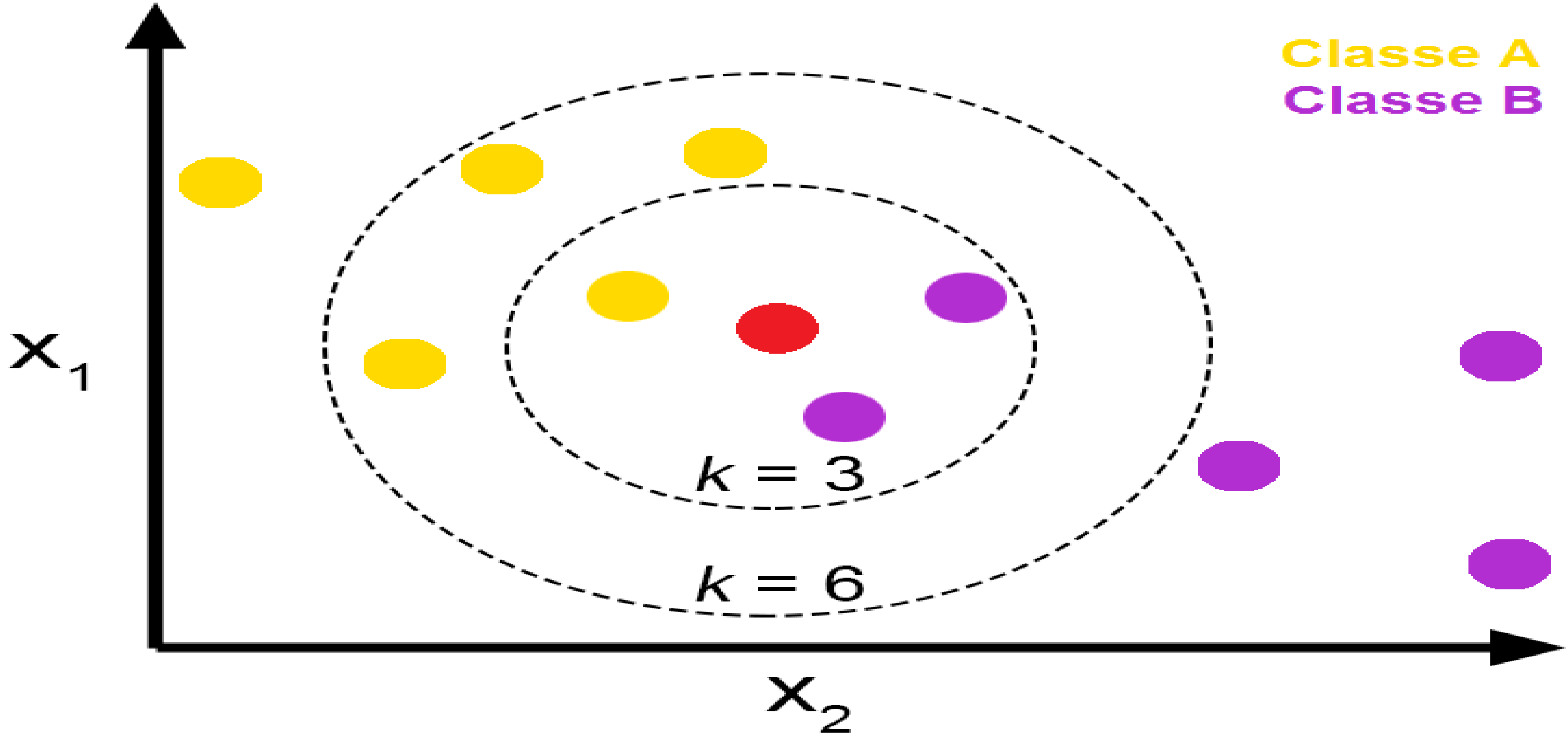
  K=1, 2, 3... As K increases, the error usually goes down, then stabilizes, and then raises again. Pick the optimum K at the beginning of the stable zone. This is also called **Elbow Method.**

  3. Domain Knowledge also plays a vital role while choosing the optimum value of K.

  4. K should be an **odd number**.

Continue…

- With change in value of K you will get the different result.



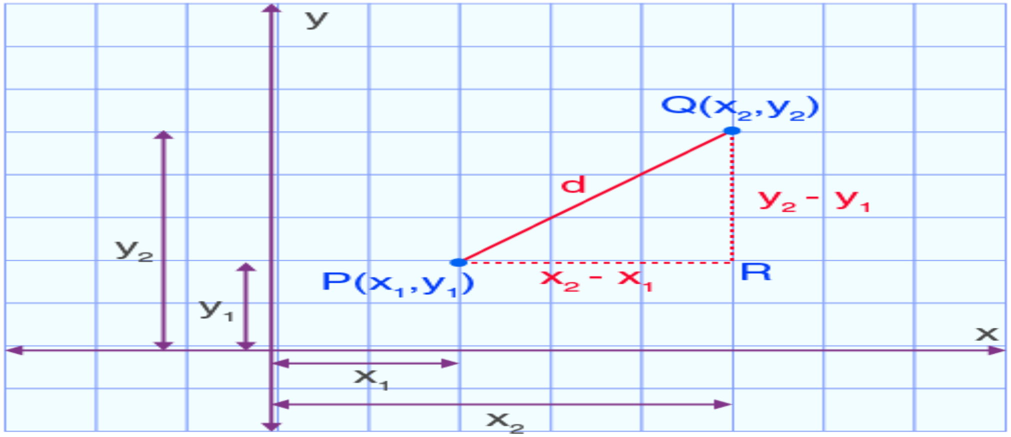
How to calculate Distance in KNN

- To calculate the distance between two points (your new sample and all the data you have in your dataset) Euclidean distance is used.
- In Mathematics, the **Euclidean distance** is defined as the distance between two points. In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.
- Let us assume two points, such as (x_1, y_1) and (x_2, y_2) in the two-dimensional coordinate plane.
- Thus, the Euclidean distance formula is given by:

  $\mathbf{d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}}$ Where,

- "d" is the Euclidean distance
- (x_1, y_1) is the coordinate of the first point
- (x_2, y_2) is the coordinate of the second point.

Continue…



Formula

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space

Applications of KNN

- Agriculture
- Finance
- Medical
- Image recognition
- Recommendation systems
- handwriting detection
- video recognition.

The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, you can use the KNN algorithm for applications that require high accuracy.

KNN is can be used where predictions are not requested frequently but where accuracy is important.

Advantages of KNN

- **No Training Period**: KNN is called **Lazy Learner (Instance based learning)**. It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc.
- Since the KNN algorithm requires no training before making predictions, **new data can be added seamlessly** which will not impact the accuracy of the algorithm.
- **Simple and Easy to Understand :**There are only two parameters required to implement KNN i.e. the value of K and the distance function
- **Can Handle Large Datasets** – The KNN algorithm can handle large datasets without suffering from the curse of dimensionality, which is a common problem in other machine learning algorithms. This makes it a suitable algorithm for problems with high-dimensional data.

Continue…

- **Accurate and Effective** – The KNN algorithm is known for its accuracy and effectiveness, particularly when used with small to medium-sized datasets. It is a robust algorithm that can handle noisy and incomplete data, making it a popular choice in many real-world applications.
- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn’t assume anything about the underlying data.

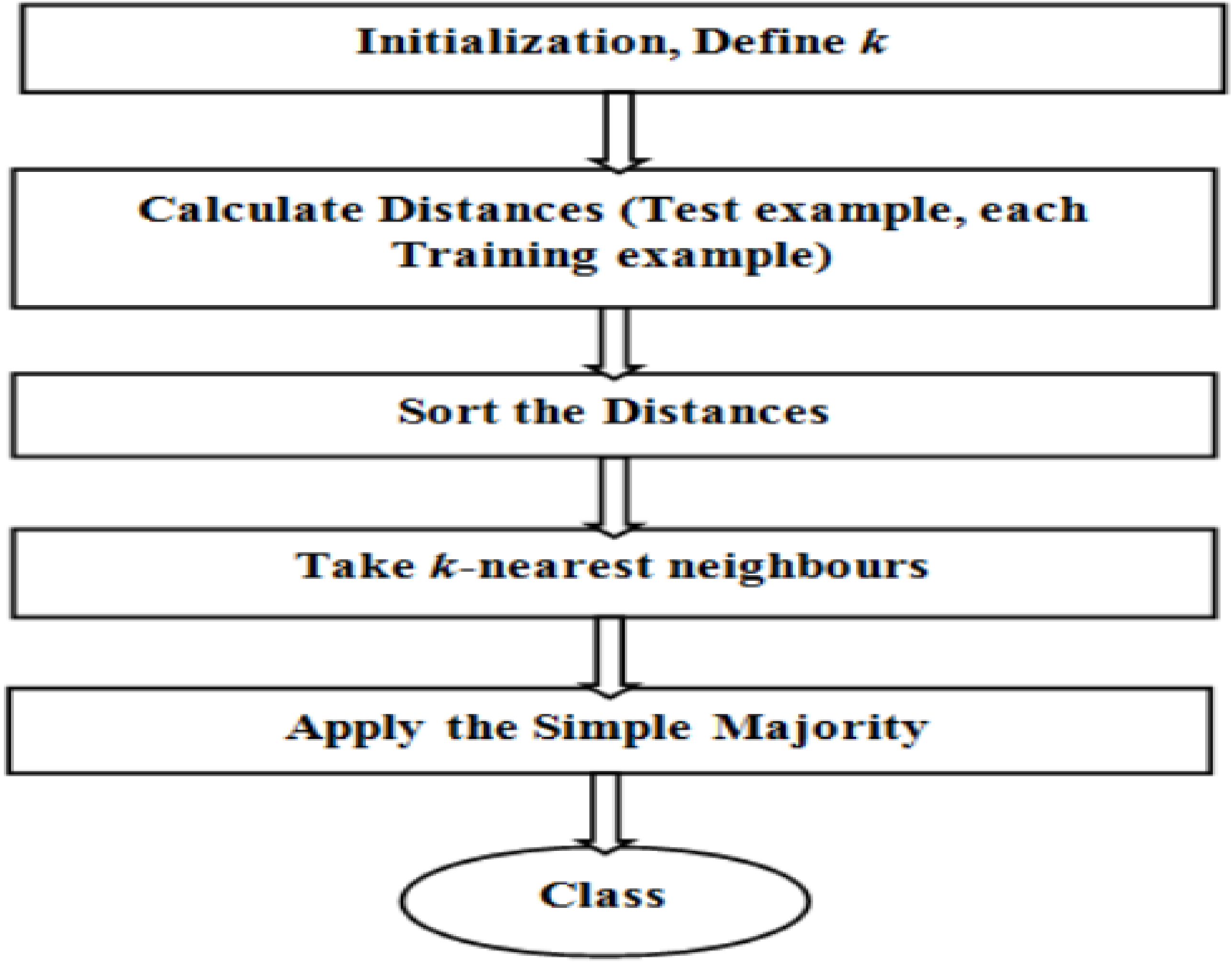
Disadvantages of KNN Algorithm:

- **Requires Good Choice of K** – The KNN algorithm requires a good choice of the K parameter, which determines the number of nearest neighbors used for classification. If K is too small, the algorithm may be too sensitive to noise in the data, while if K is too large, the algorithm may miss important patterns in the data.
- **Computationally Expensive** – The KNN algorithm can be computationally expensive, particularly for large datasets. This is because the algorithm needs to compute the distance between each test data point and every training data point, which can be time-consuming.
- **Sensitive to Outliers** – The KNN algorithm can be sensitive to outliers in the data, which can significantly affect its performance.

Steps of working of K-NN

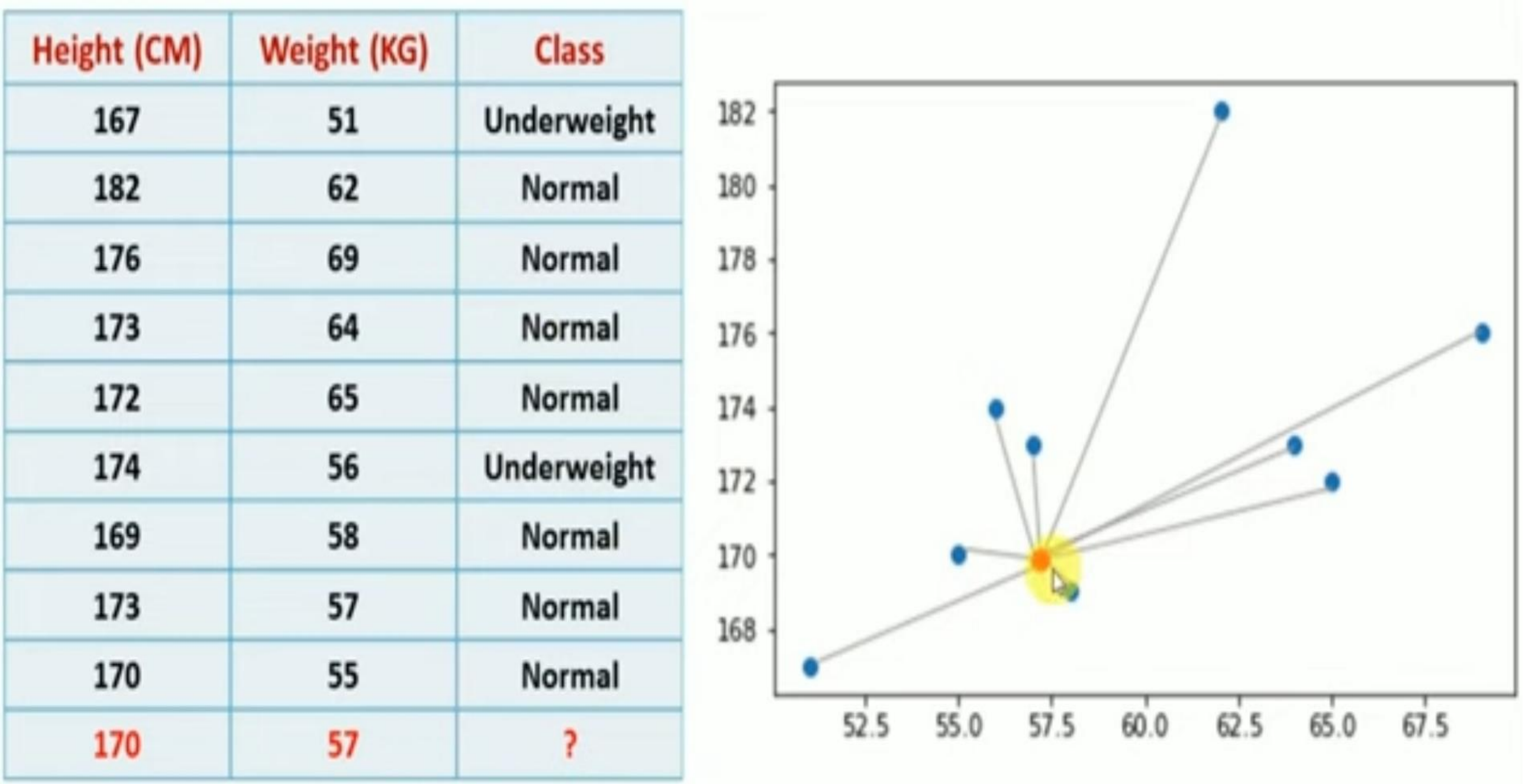
- We can implement a KNN model by following the below steps:
- Load the data
- Initialize the value of k
- For getting the predicted class, iterate from 1 to total number of training data points
  - Calculate the distance between test data and each row of training dataset. Here we will use Euclidean distance as our distance metric since it’s the most popular method.
  - Sort the calculated distances in ascending order based on distance values
  - Get top k rows from the sorted array
  - Get the most frequent class of these rows
  - Return the predicted class

Continue…



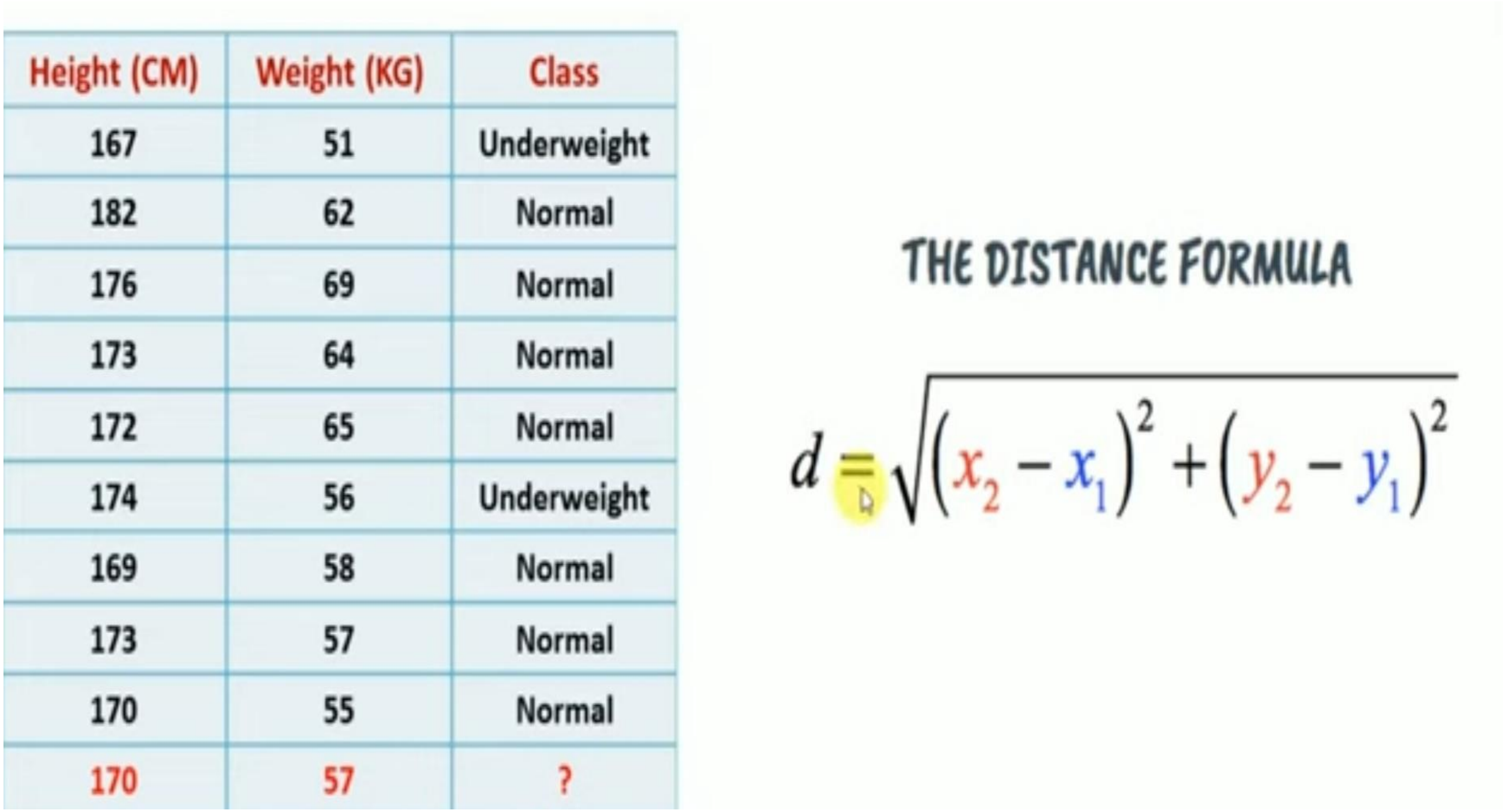
Initialization, Define *k*
Calculate Distances (Test example, each Training example)
Sort the Distances
Take *k*-nearest neighbours
Apply the Simple Majority
Class

KNN Example



Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

Continue…

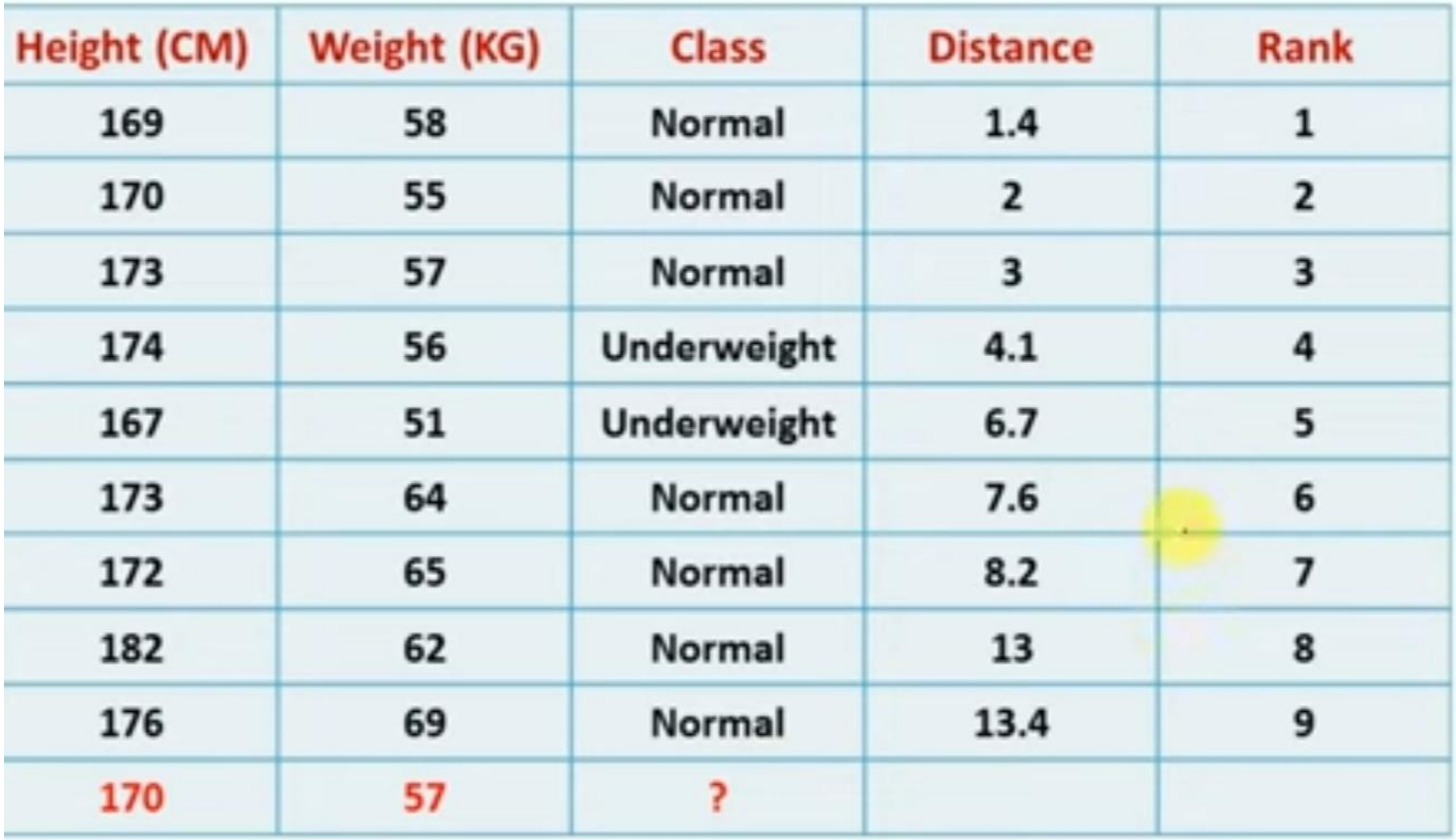


Height (CM)	Weight (KG)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
170	57	?

THE DISTANCE FORMULA

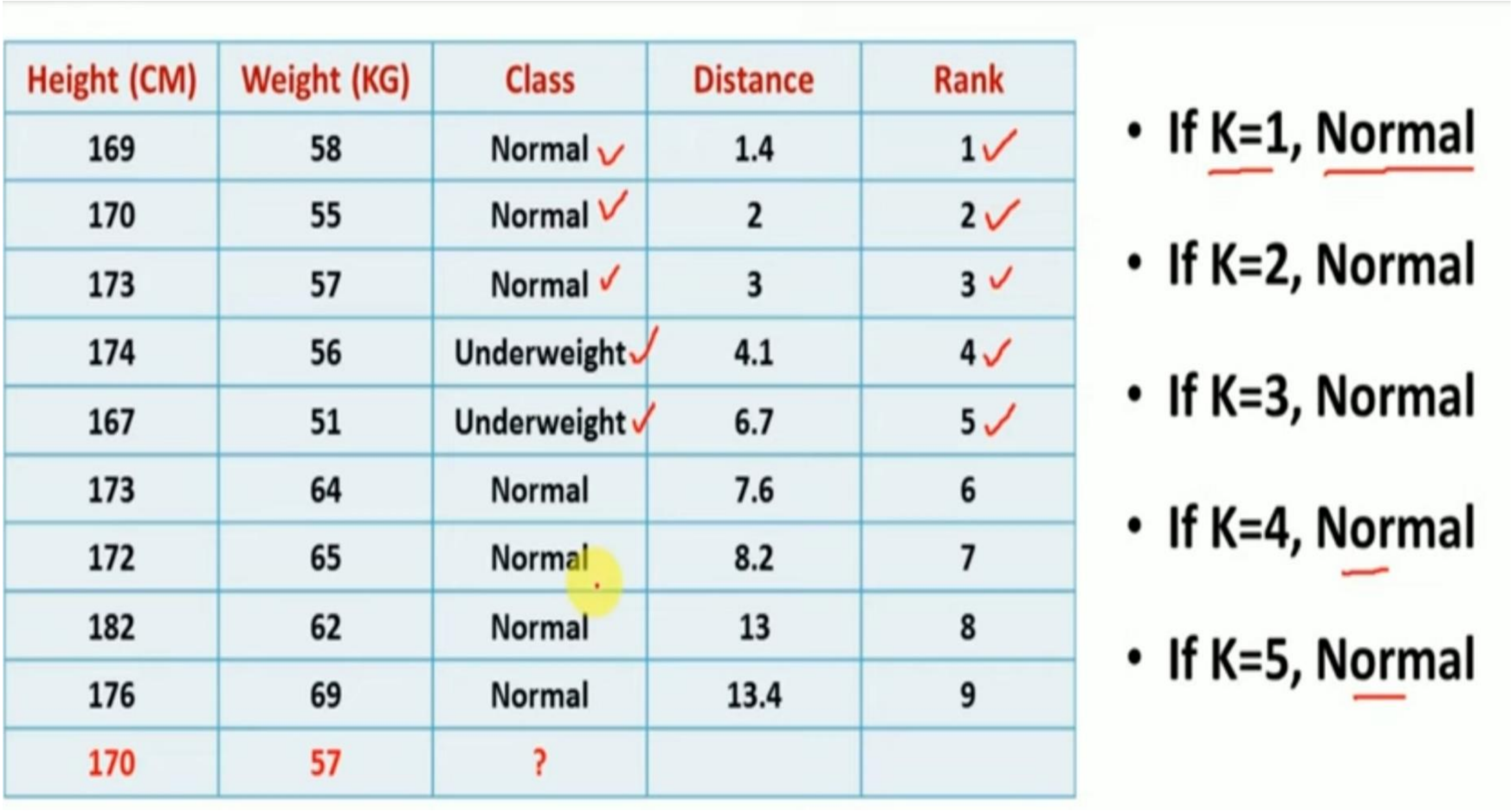
$$d = \sqrt{\left(x_2 - x_1\right)^2 + \left(y_2 - y_1\right)^2}$$

Continue…



Height (CM)	Weight (KG)	Class	Distance	Rank
169	58	Normal	1.4	1
170	55	Normal	2	2
173	57	Normal	3	3
174	56	Underweight	4.1	4
167	51	Underweight	6.7	5
173	64	Normal	7.6	6
172	65	Normal	8.2	7
182	62	Normal	13	8
176	69	Normal	13.4	9
170	57	?		

Continue…

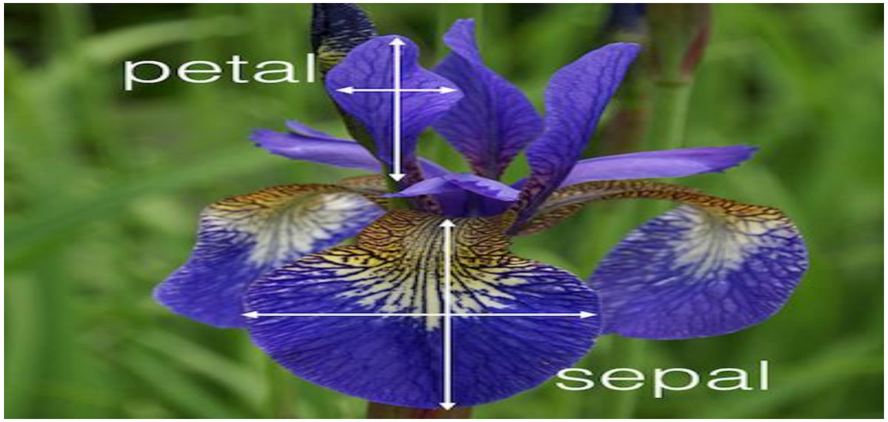


Height (CM)	Weight (KG)	Class	Distance	Rank
169	58	Normal ✓	1.4	1 ✓
170	55	Normal ✓	2	2 ✓
173	57	Normal ✓	3	3 ✓
174	56	Underweight ✓	4.1	4 ✓
167	51	Underweight ✓	6.7	5 ✓
173	64	Normal	7.6	6
172	65	Normal	8.2	7
182	62	Normal	13	8
176	69	Normal	13.4	9
170	57	?		

- If K=1, Normal
- If K=2, Normal
- If K=3, Normal
- If K=4, Normal
- If K=5, Normal

KNN Implementation

- **Iris Species Dataset**
- It includes three iris species (a group of plants) with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.
- Three species are setosa, versicolor, virginica
- The columns in this dataset are:
- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species

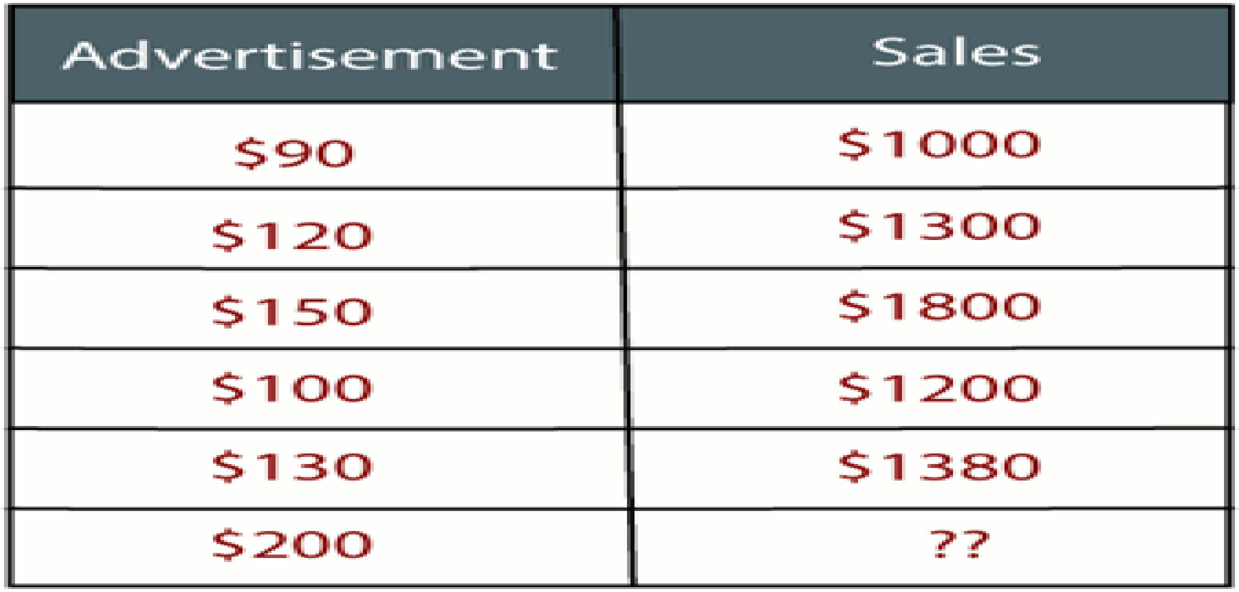


- https://colab.research.google.com/drive/1qMXzY0iCHUp31yw6ixndV7L5Nnpd2DqV#scrollTo=Tp2vka_X3ByT

Regression

- Machine Learning Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome.
- It’s used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.
- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.

Continue...



Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

list shows the advertisement made by the company in the last 5 years and the corresponding sales.

Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.

Types of Regression

- There are various types of regressions which are used in data science and machine learning. all the regression methods analyze the effect of the independent variable on dependent variables.
- **Linear Regression**
- Logistic Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Ridge Regression
- Lasso Regression

Real world examples of regression analysis

- Financial forecasting (like house price estimates, or stock prices)
- Sales and promotions forecasting
- Weather analysis and prediction
- Time series forecasting
- In voting applications to find out whether voters will vote for a particular candidate or not.

Linear regression

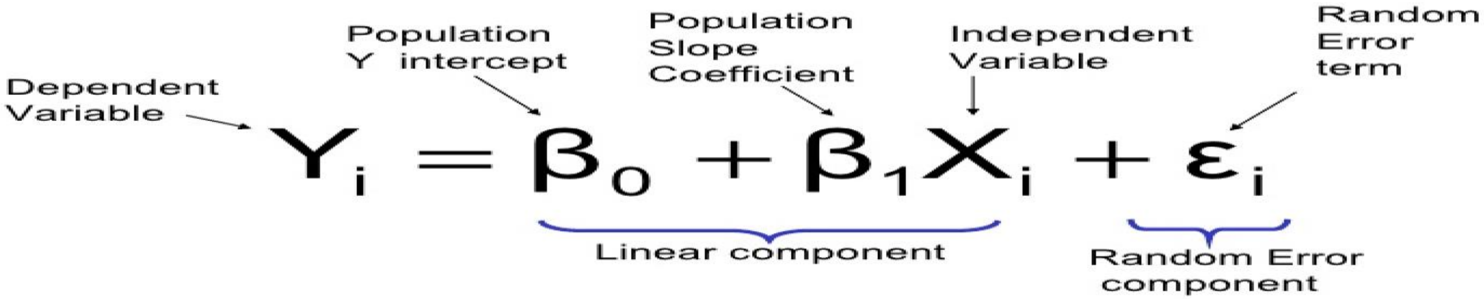
- Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a **dependent variable** and **one or more independent features (variable).**
- When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression.

Types of Linear Regression

- **There are two kinds of Linear Regression Model.**
- **Simple Linear Regression**: A linear regression model with one independent and one dependent variable.
- **Multiple Linear Regression**: A linear regression model with more than one independent variable and one dependent variable.

Mathematical equation of linear regression

- A linear regression line has an equation of the form **Y = a + bX**, where **X** is the explanatory variable and **Y** is the dependent variable. The slope of the line is **b**, and **a** is the intercept/constant (the value of **y** when **X** = 0).



$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

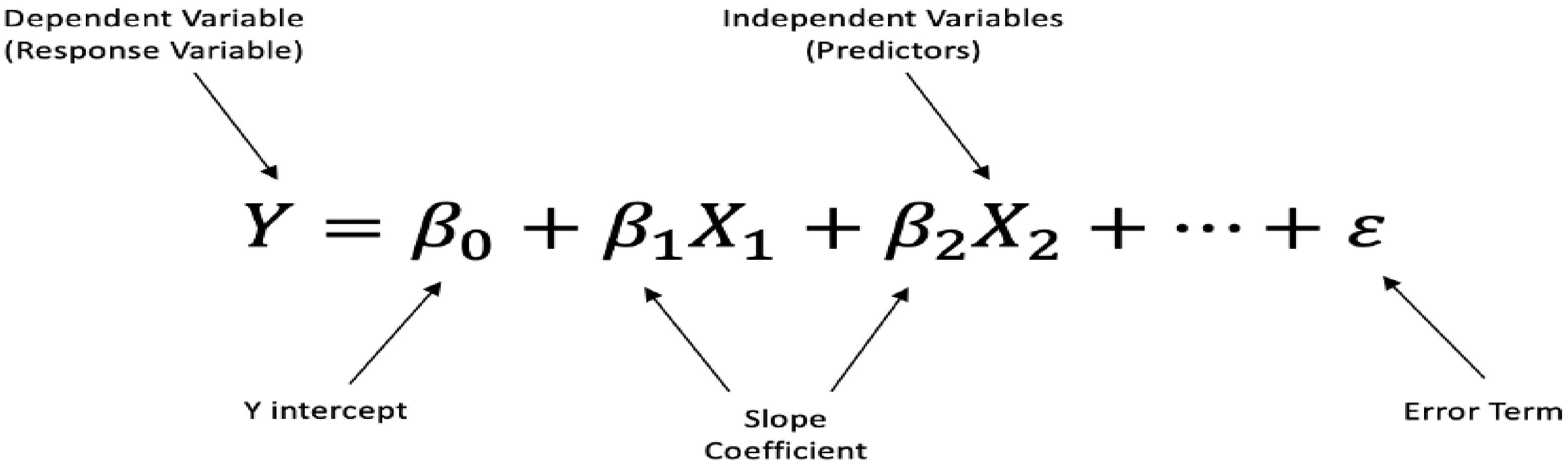
Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Dependent Variable (Response Variable)

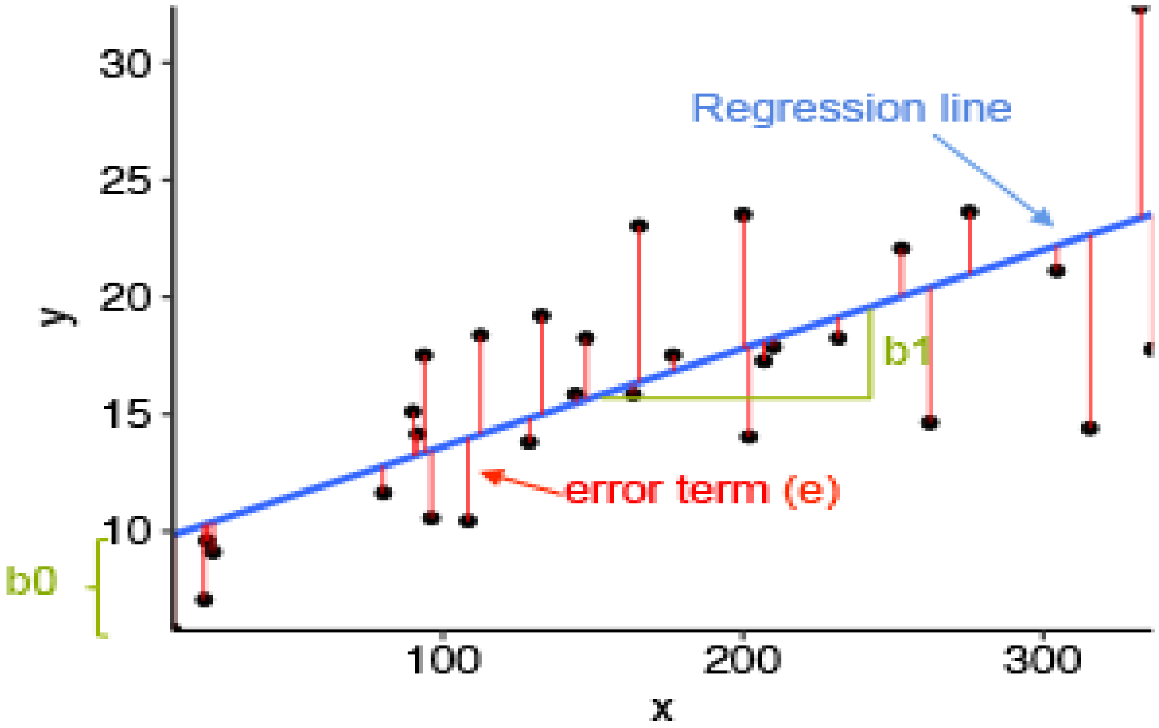
Independent Variables (Predictors)

Y intercept

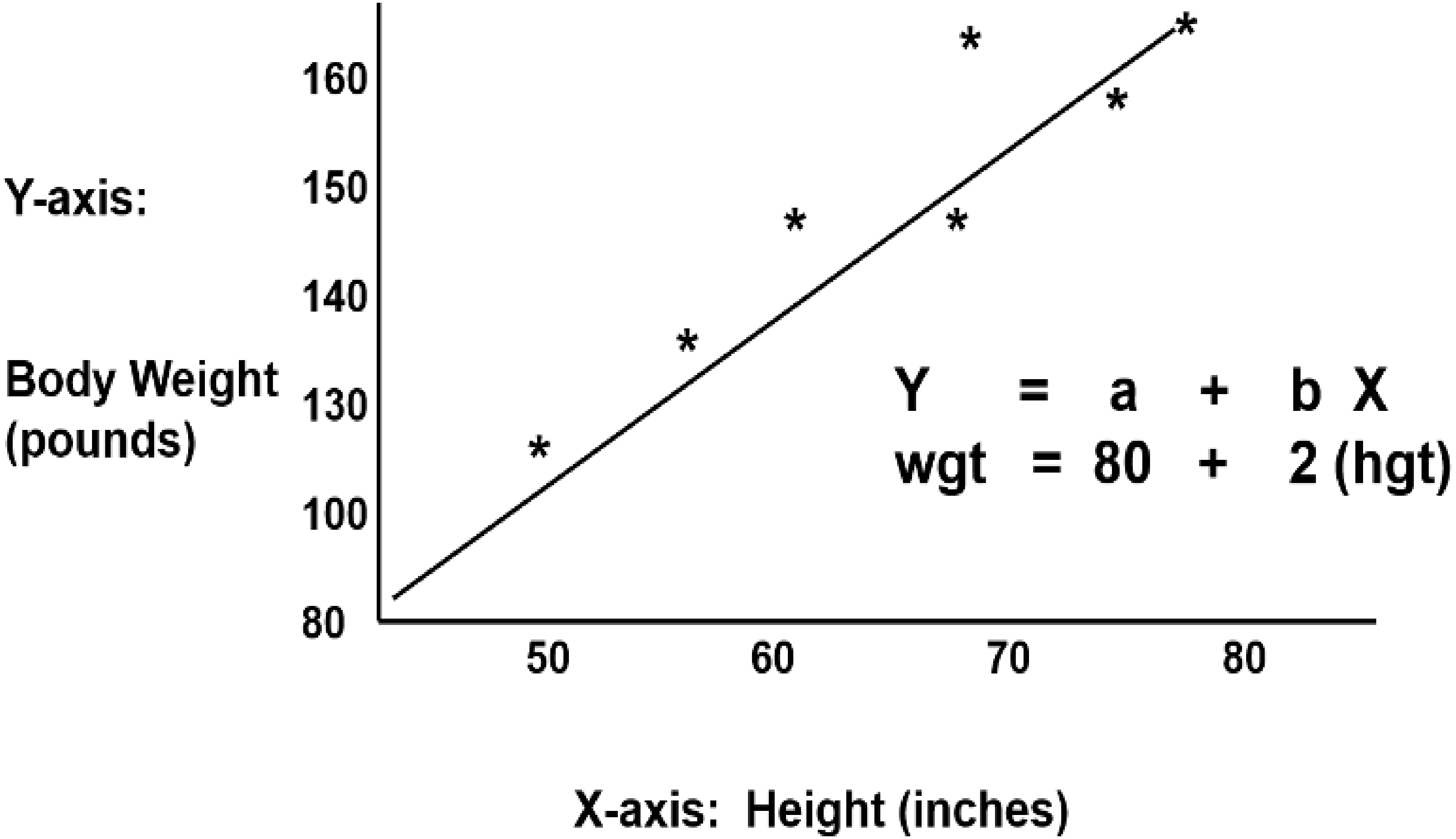
Slope Coefficient

Error Term

Continue…



Continue…



$$Y = a + b\,X$$

$$wgt = 80 + 2\,(hgt)$$

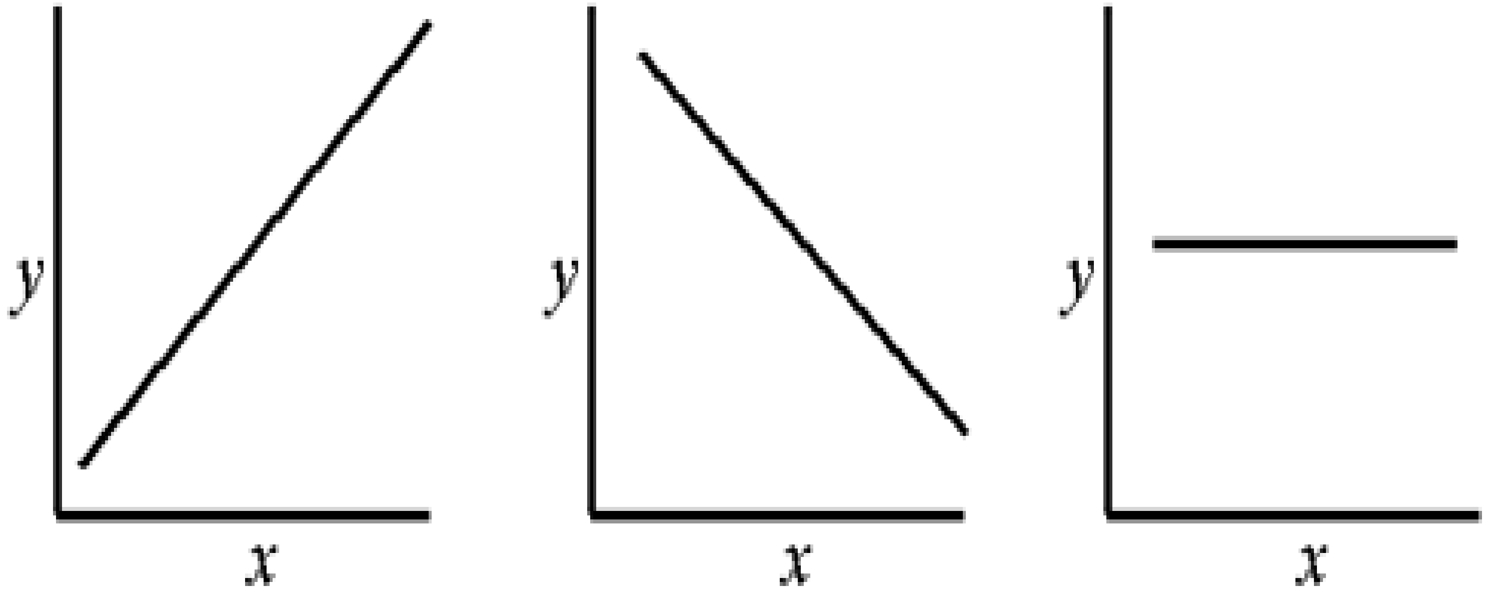
linear regression line

- A regression line indicates a linear relationship between the dependent variables on the y-axis and the independent variables on the x-axis.
- The correlation is established by analyzing the data pattern formed by the variables.
- The regression line is plotted closest to the data points in a regression graph.
- A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

Types of regression line

- The sign of a linear regression coefficient tells you whether there is a positive or negative correlation between each independent variable and the dependent variable.
- A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase (You will get **Positive regression Line**).
- A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease (You will get **Negative regression Line**).

Continue…



Positive slope

Negative slope

Zero slope

Applications of linear regression

- Linear regression is a statistical measure that establishes the relationship between variables that businesses use to develop forecasts and make informed decisions.
- It has applications in **Market analysis , Financial analysis, Business planning ,Sports analysis, Environmental health , Health and Medicine, Agriculture.**

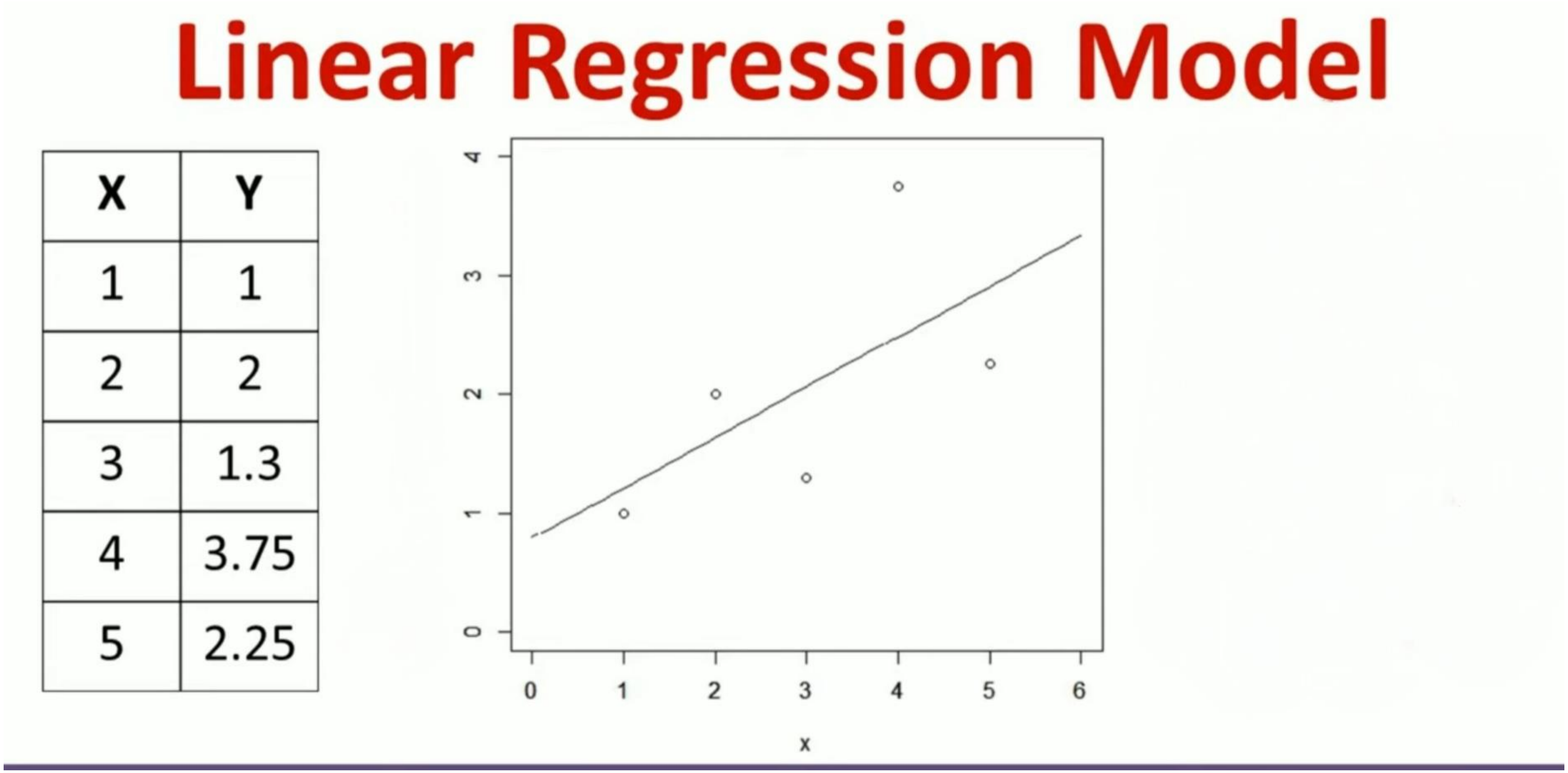
Advantages of Linear Regression

- Simple implementation
- Good Performance on linearly seperable datasets
- Overfitting can be reduced by regularization

Dis Advantages of Linear Regression

- Prone to underfitting
- Sensitive to outliers
- Limited to Linear Relationships

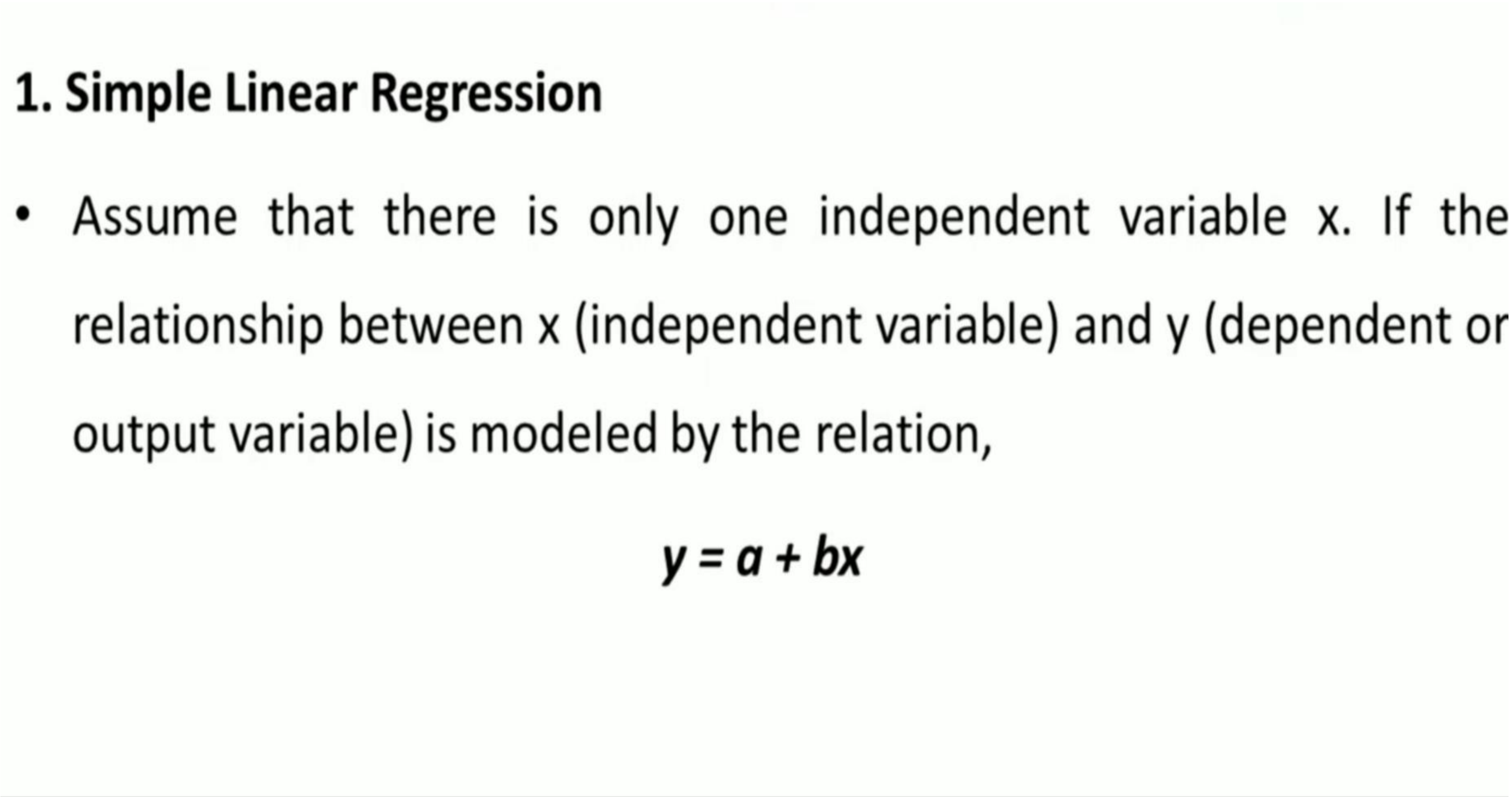
Simple Linear Regression Example



Linear Regression Model

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Continue…

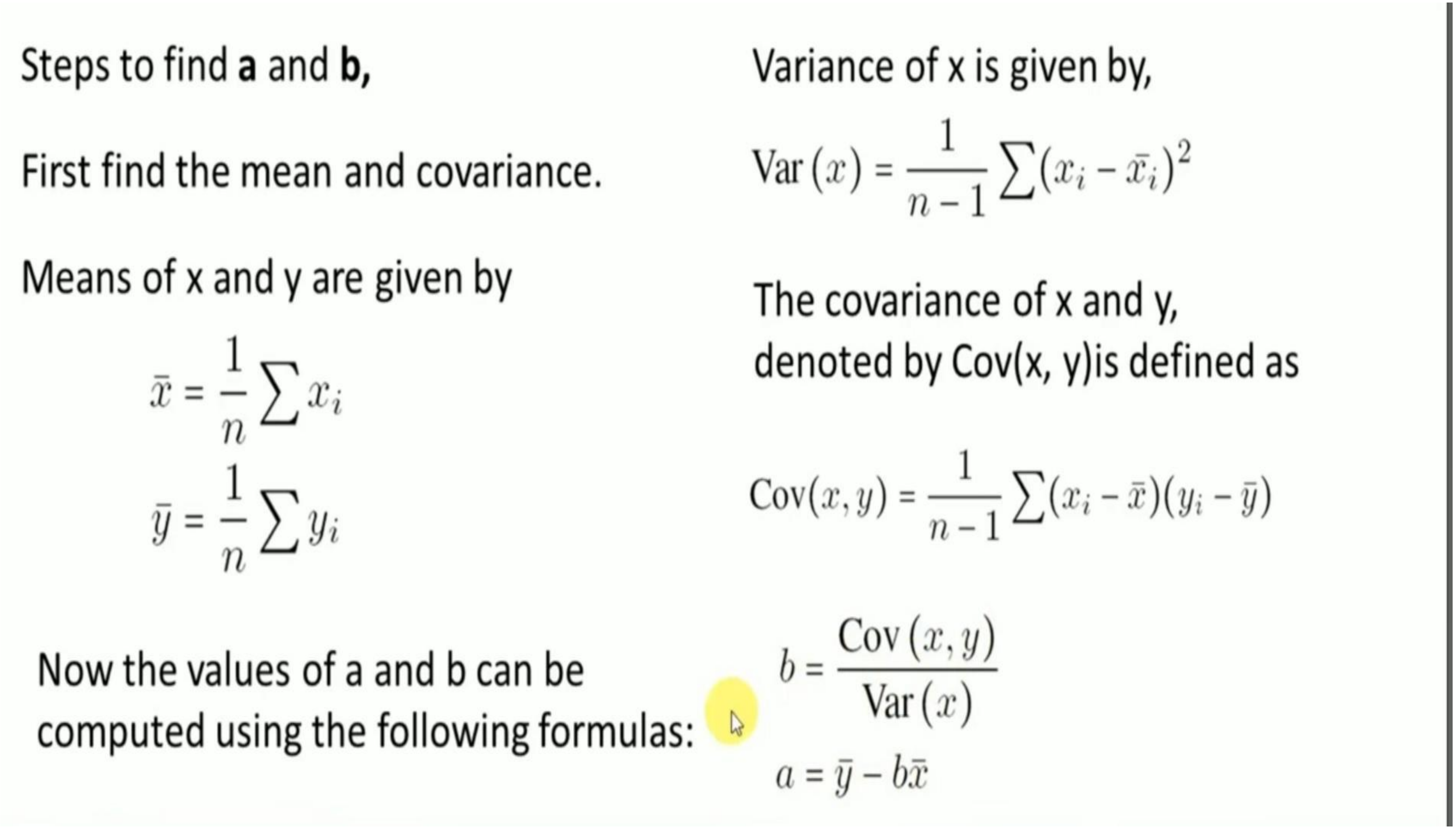


1. Simple Linear Regression

- Assume that there is only one independent variable x. If the relationship between x (independent variable) and y (dependent or output variable) is modeled by the relation,

$$y = a + bx$$

Continue…



Steps to find **a** and **b**,

First find the mean and covariance.

Means of x and y are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

Variance of x is given by,

$$\text{Var}\,(x) = \frac{1}{n-1}\sum(x_i - \bar{x}_i)^2$$

The covariance of x and y, denoted by Cov(x, y)is defined as

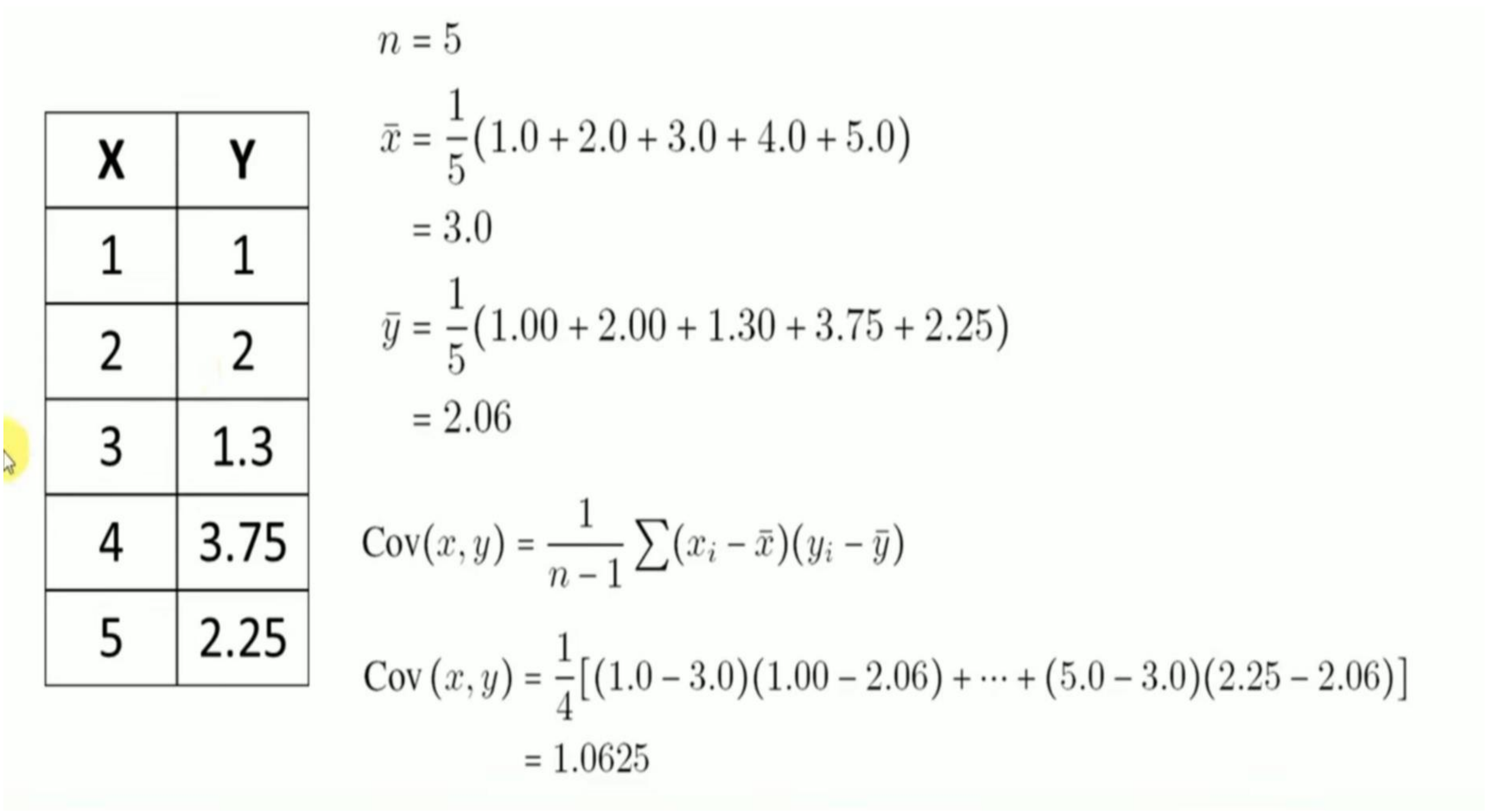
$$\text{Cov}(x, y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

Now the values of a and b can be computed using the following formulas:

$$b = \frac{\text{Cov}\,(x, y)}{\text{Var}\,(x)}$$

$$a = \bar{y} - b\bar{x}$$

Continue…



X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

$$n = 5$$

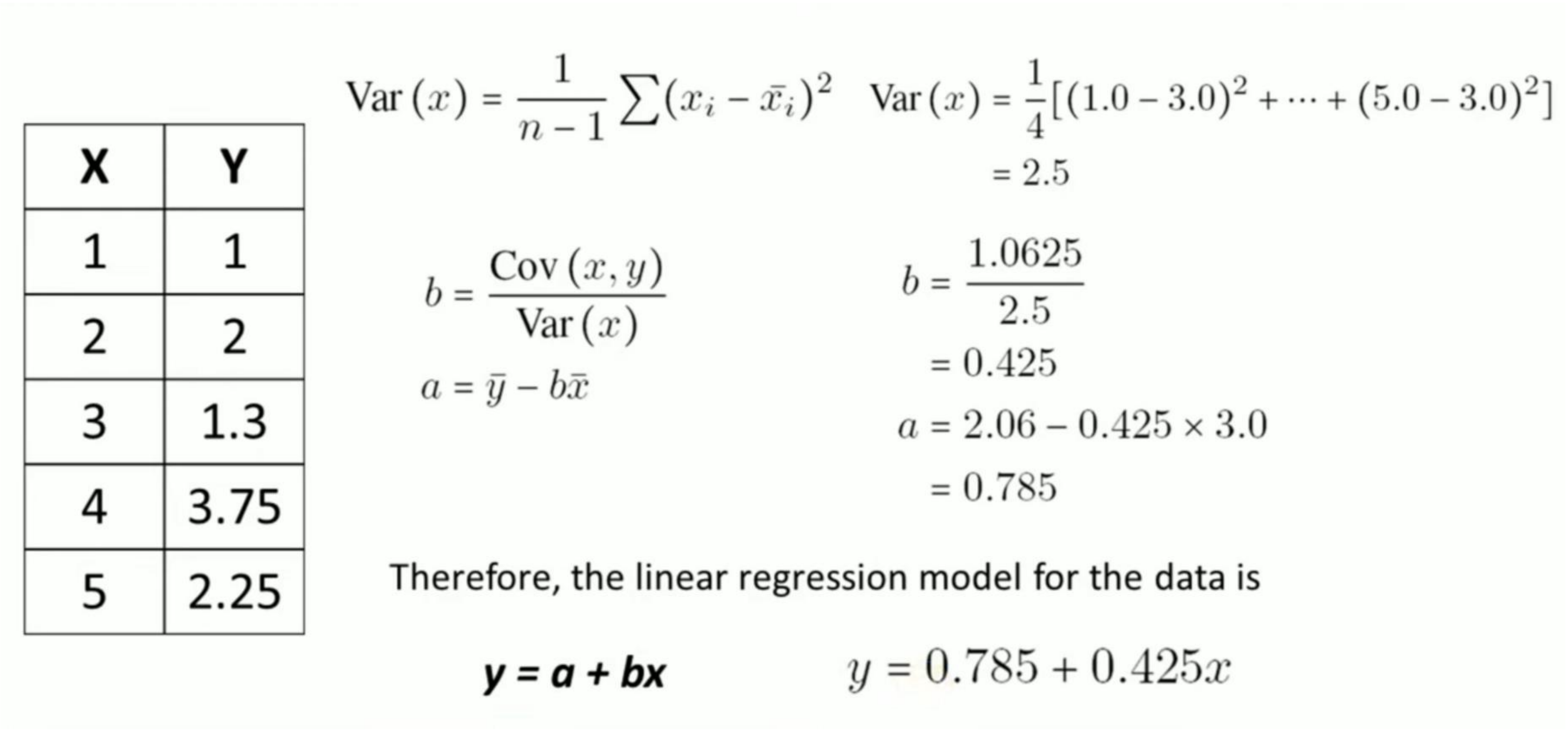
$$\bar{x} = \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0)$$
$$= 3.0$$

$$\bar{y} = \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25)$$
$$= 2.06$$

$$\mathrm{Cov}(x, y) = \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Cov}\,(x, y) = \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)]$$
$$= 1.0625$$

Continue…



X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

$$\text{Var}\,(x) = \frac{1}{n-1}\sum(x_i - \bar{x_i})^2 \quad \text{Var}\,(x) = \frac{1}{4}[(1.0 - 3.0)^2 + \cdots + (5.0 - 3.0)^2]$$

$$= 2.5$$

$$b = \frac{\text{Cov}\,(x, y)}{\text{Var}\,(x)} \qquad b = \frac{1.0625}{2.5}$$

$$= 0.425$$

$$a = \bar{y} - b\bar{x} \qquad a = 2.06 - 0.425 \times 3.0$$

$$= 0.785$$

Therefore, the linear regression model for the data is

y = a + bx $\qquad y = 0.785 + 0.425x$

Simple Linear Regression Implementation

- Salary prediction based on Experience.
- Dataset contain 30 entry for experience in years (X) and related salary (Y).
- By using Simple Linear regression model we can predict the value of salary for given experience.

https://colab.research.google.com/drive/1JH3f11GaJcd1FuksljiJlhjjlF9XEnKU

Evaluation metric for simple linear regression

- **Mean Squared Error (MSE) = $\Sigma(yi - \hat{y}i)^2 / n$**
  - yi is the observed (actual) value for the i th data point.
  - $\hat{y}i$ is the predicted value for the i th data point.
  - n is the number of data points.

- **Root Mean Squared Error (RMSE) =$\sqrt{}$(MSE)**

- **Mean Absolute Error (MAE) = $\Sigma|yi - \hat{y}i| / n$**

Happy Learning