Parul® University

Data Visualization and Data Analytics

Prof. Khushbu Chauhan, Assistant Professor
Information Technology (PIET)

CHAPTER 5

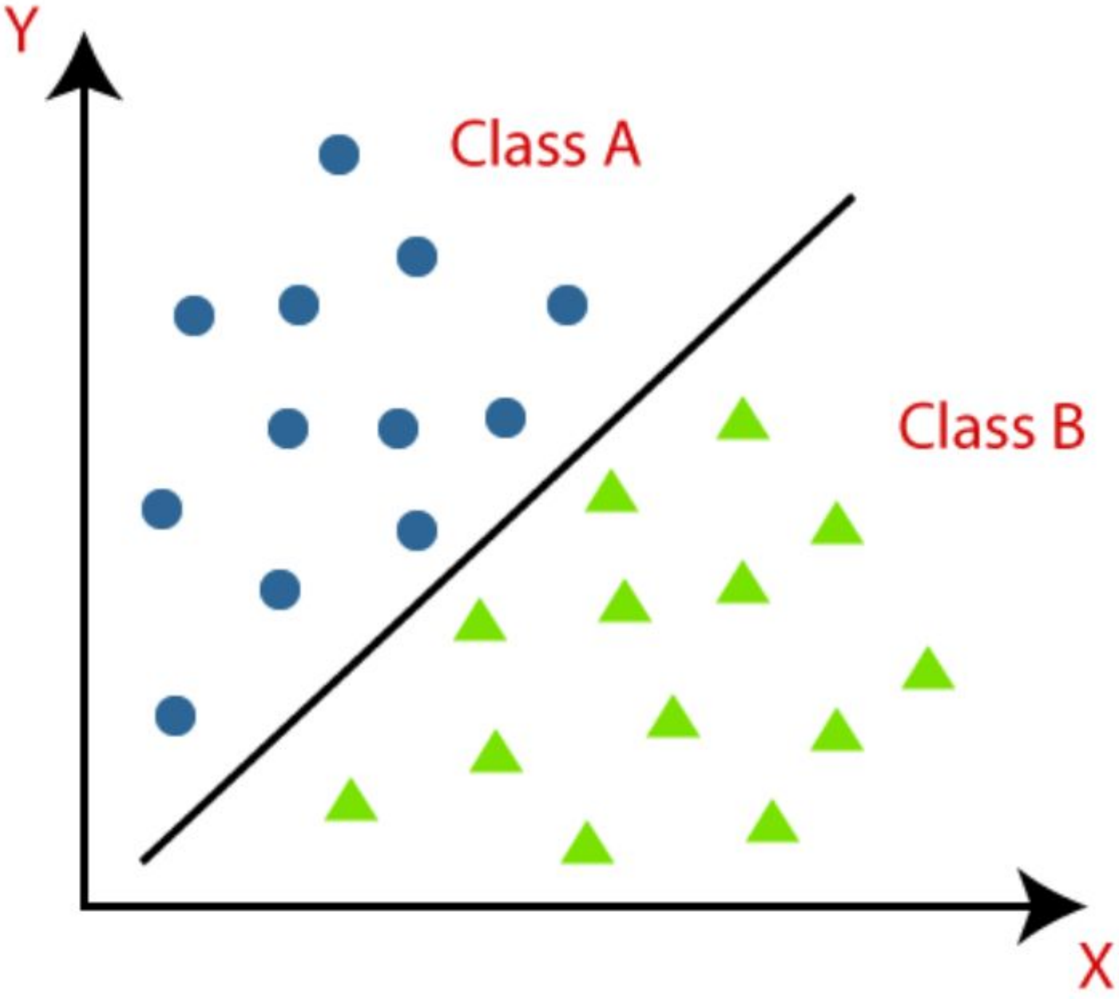
Classification & Clustering

Outlines

- Introduction – Classification & Clustering
- Use of classification & clustering for insights
- KNN
- Decision Tree
- K-Means clustering
- Cluster Analysis
- Introduction to analytics tool - PowerBI

Introduction

- **Classification** :
- classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Examples of classification problems include: Given an example, classify if it is spam or not. Given a handwritten character, classify it as one of the known characters.
- 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 observations into a number of classes or groups.


Y
Class A
Class B
X

Binary Classification

- Binary classification refers to those classification tasks that have two class labels.
- Examples include:
- Email spam detection (spam or not).
- Churn prediction (churn or not).
- Conversion prediction (buy or not).
- Typically, binary classification tasks involve one class that is the normal state and another class that is the abnormal state.

Multi class classification

- Multi-class classification refers to those classification tasks that have more than two class labels.
- Examples include:
- Face classification.
- Plant species classification.
- Optical character recognition.
- Unlike binary classification, the multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a range of known classes.

Multi Label Classification

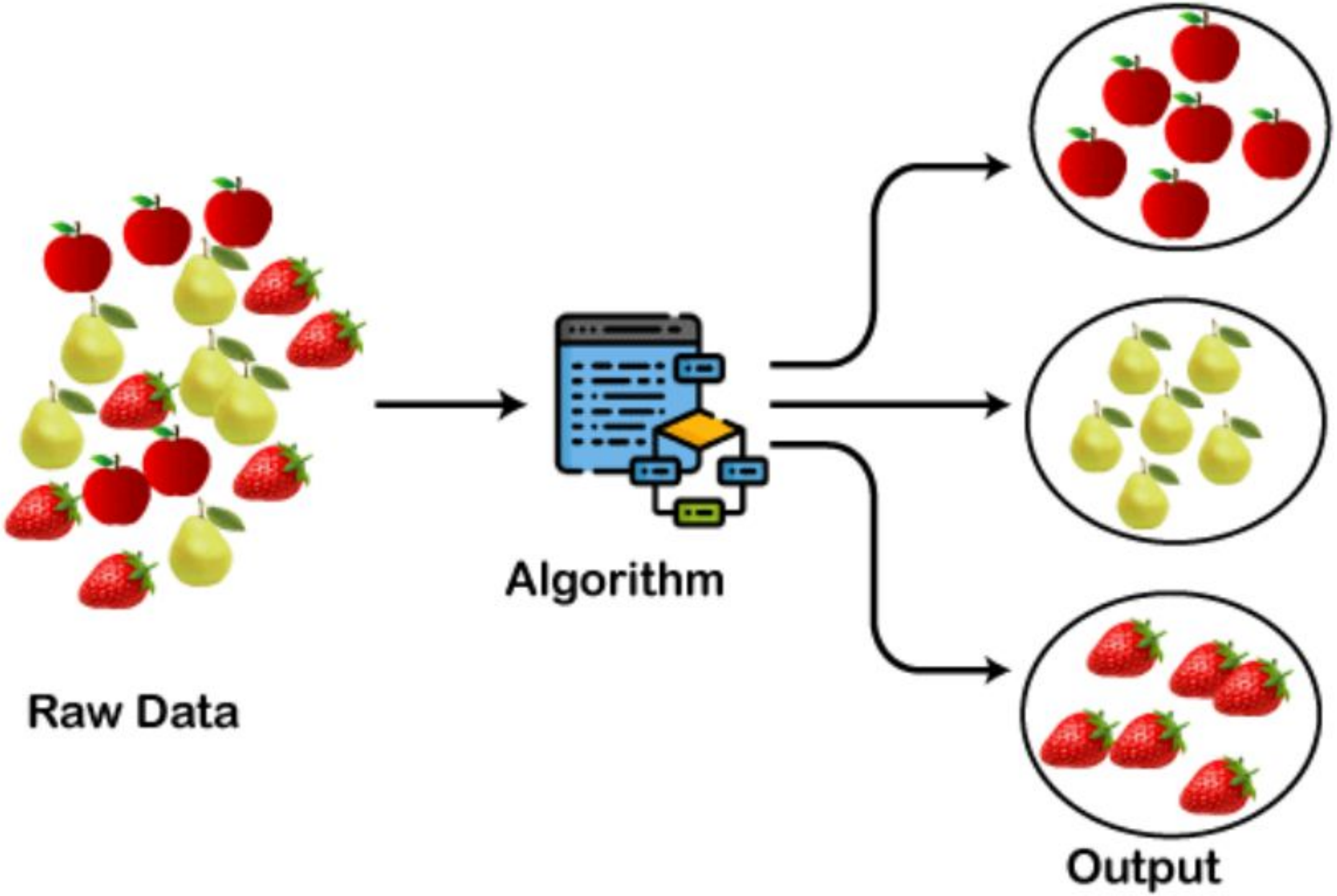
- Multi-label classification refers to those classification tasks that have two or more class labels, where one or more class labels may be predicted for each example.
- Consider the example of photo classification, where a given photo may have multiple objects in the scene and a model may predict the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

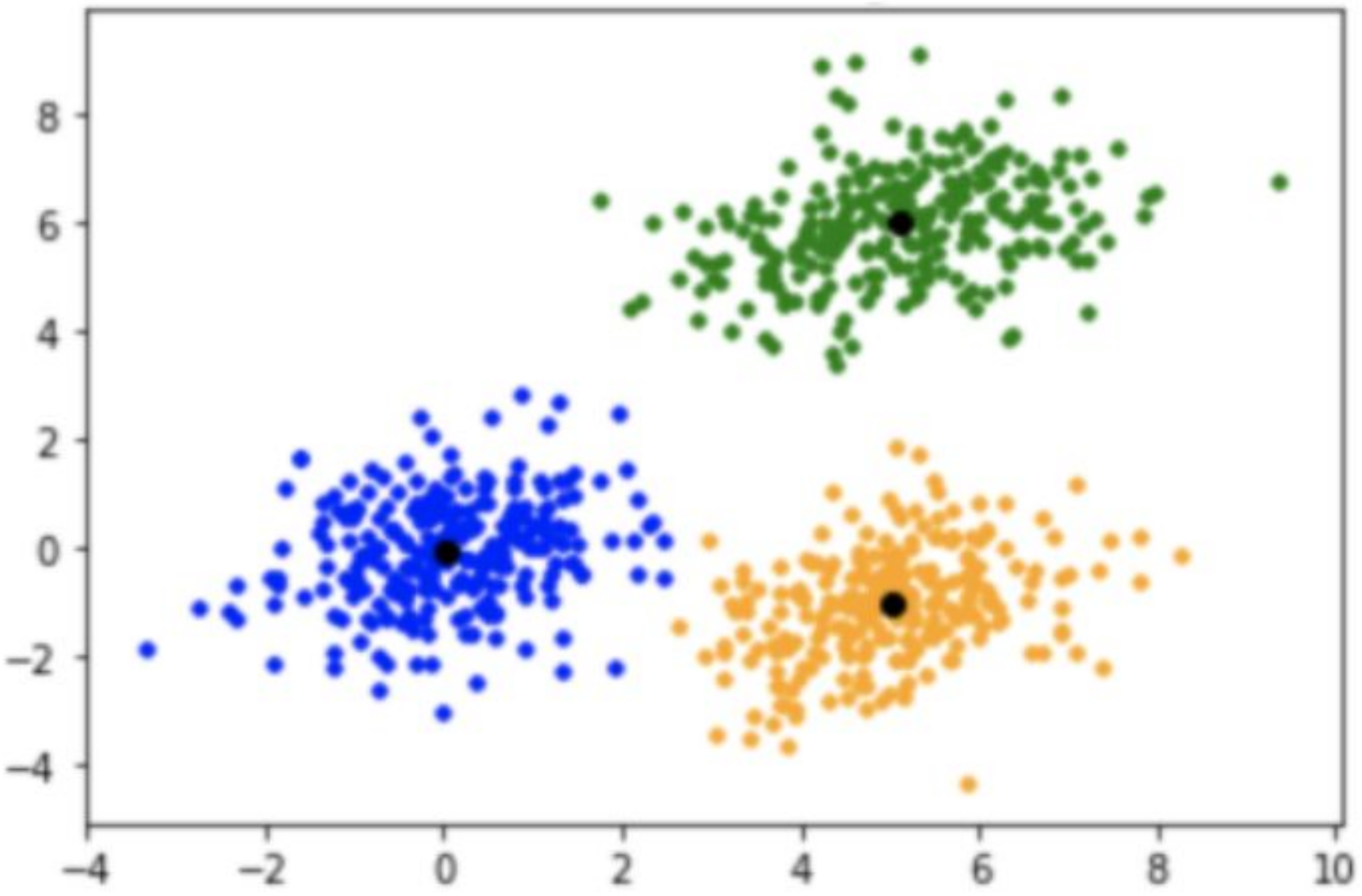
Imbalanced classification

- Imbalanced classification refers to classification tasks where the number of examples in each class is unequally distributed.
- 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.
- Examples include:
- Fraud detection.
- Outlier detection.
- Medical diagnostic tests.
- These problems are modeled as binary classification tasks, although may require specialized techniques.

Clustering

- **Clustering** is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.
- data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.


Raw Data
Algorithm
Output

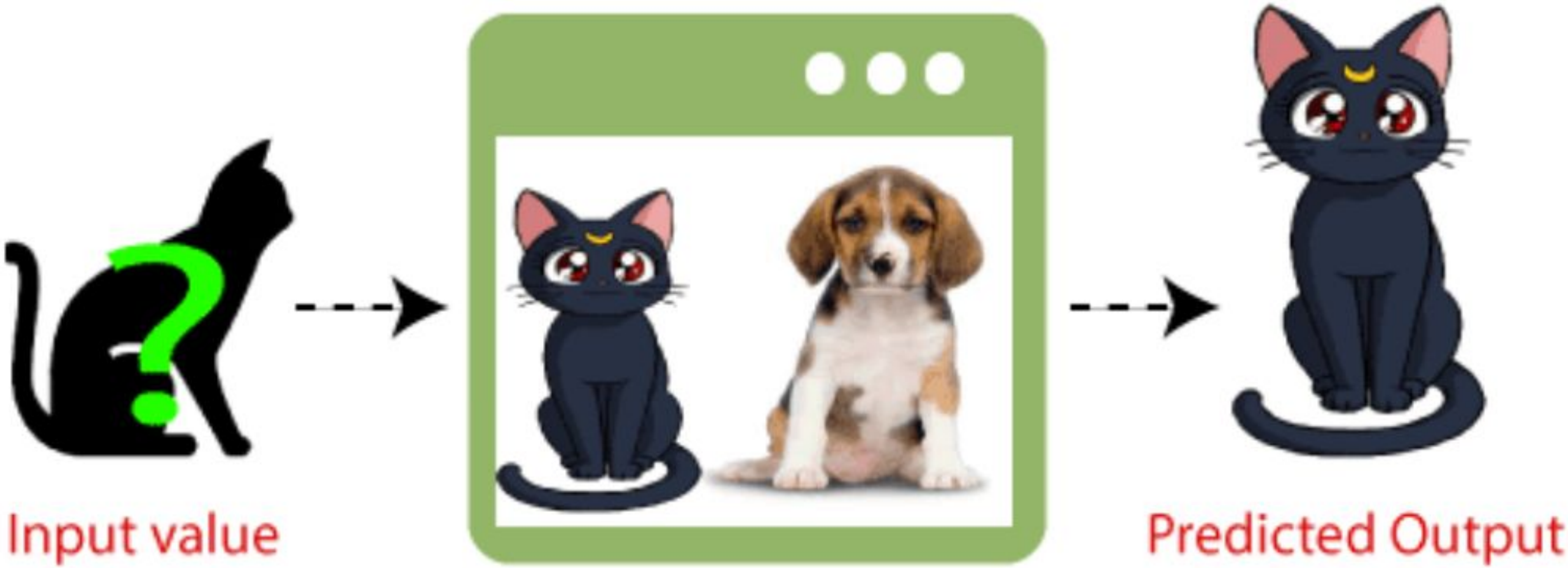

8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10

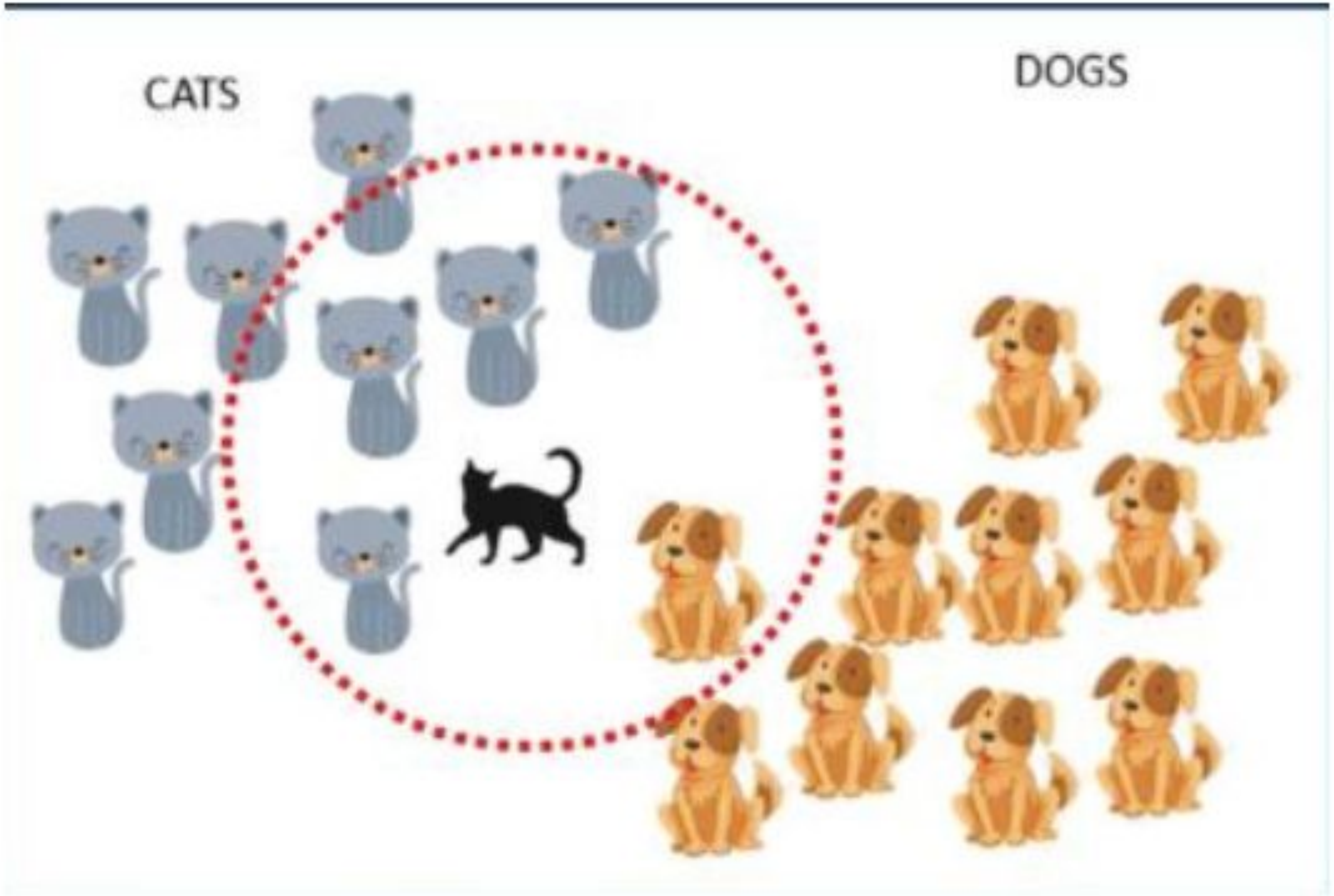
KNN

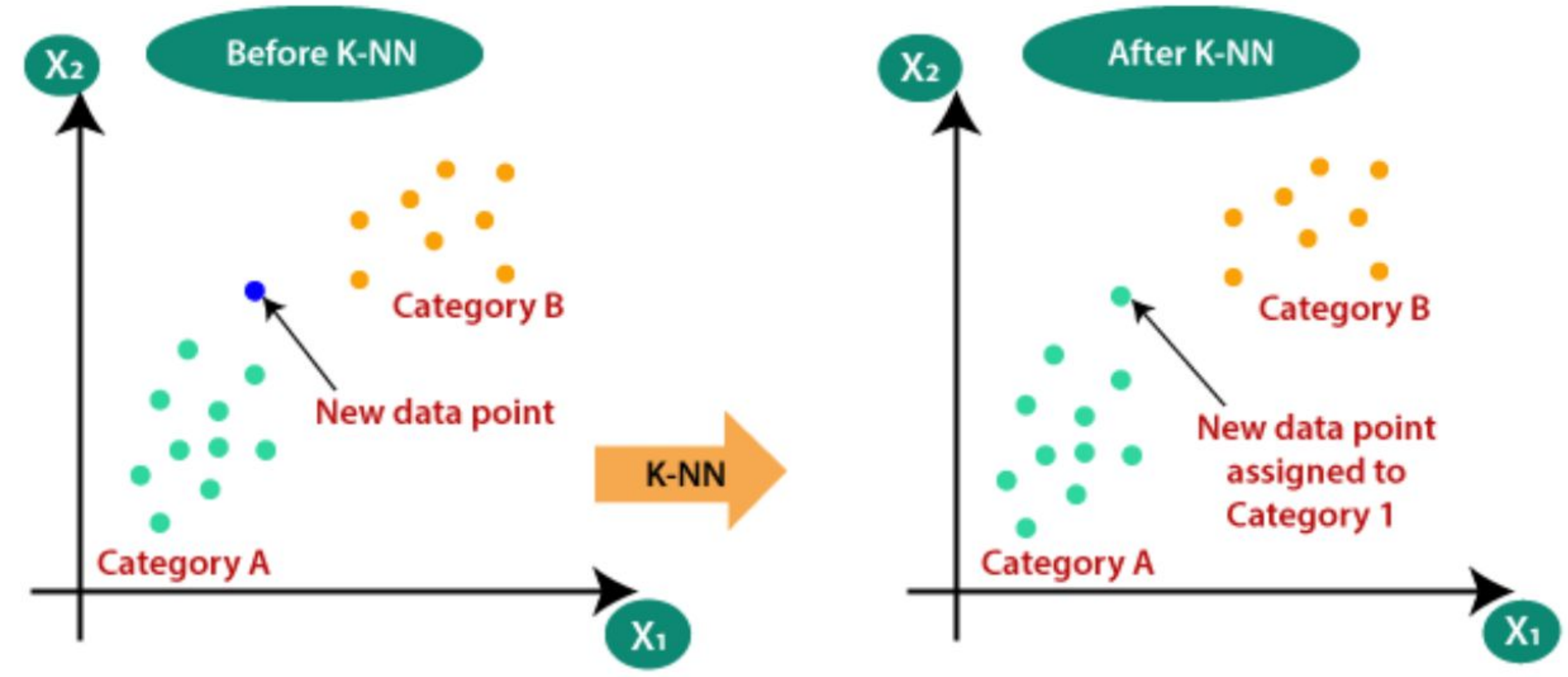
- **K Nearest Neighbour :**
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on the Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available 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 good 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 Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- 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.

KNN Classifier




CATS
DOGS


Before K-NN
X_2
Category B
New data point
Category A
X_1
K-NN
After K-NN
X_2
Category B
New data point assigned to Category 1
Category A
X_1

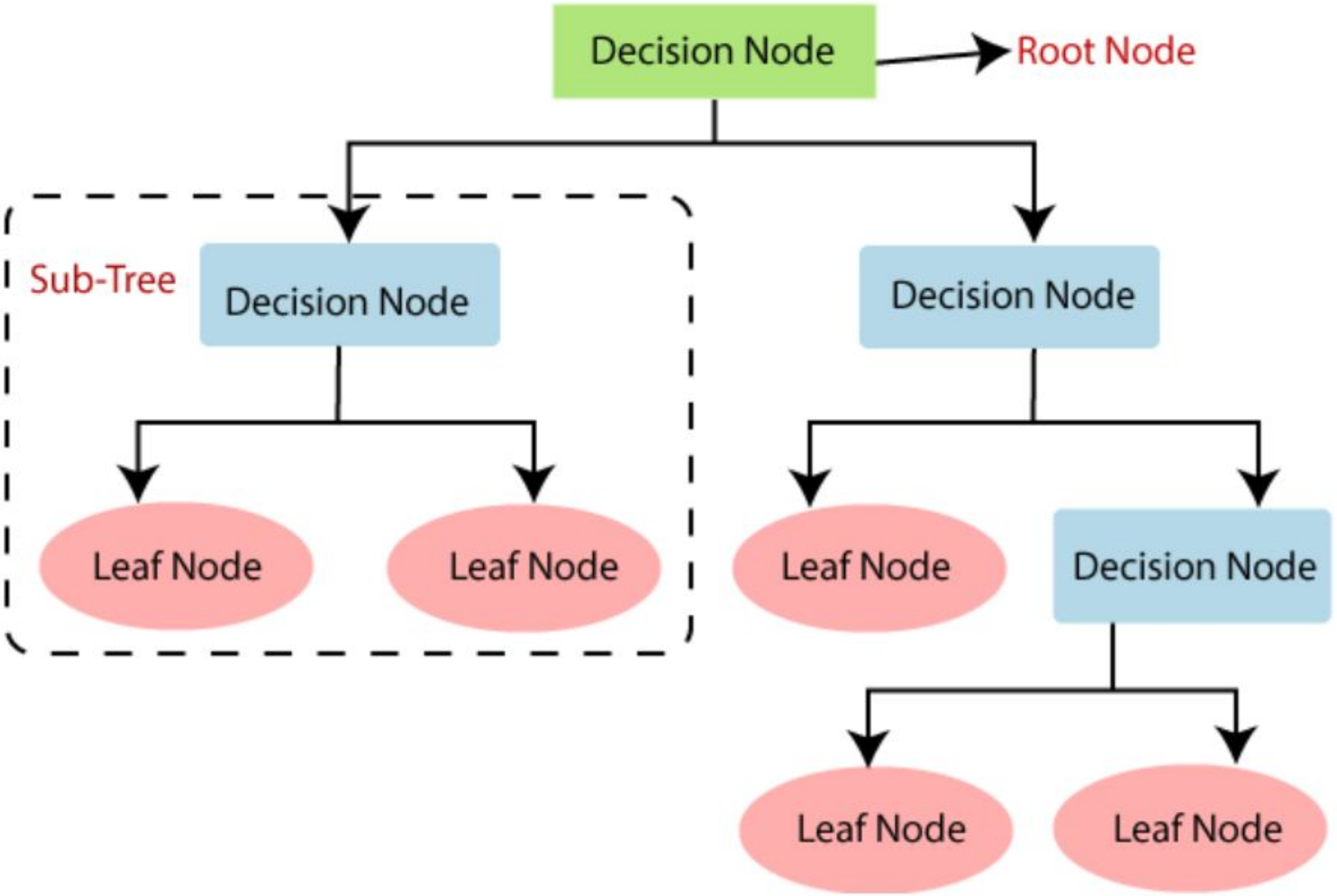
How does KNN works

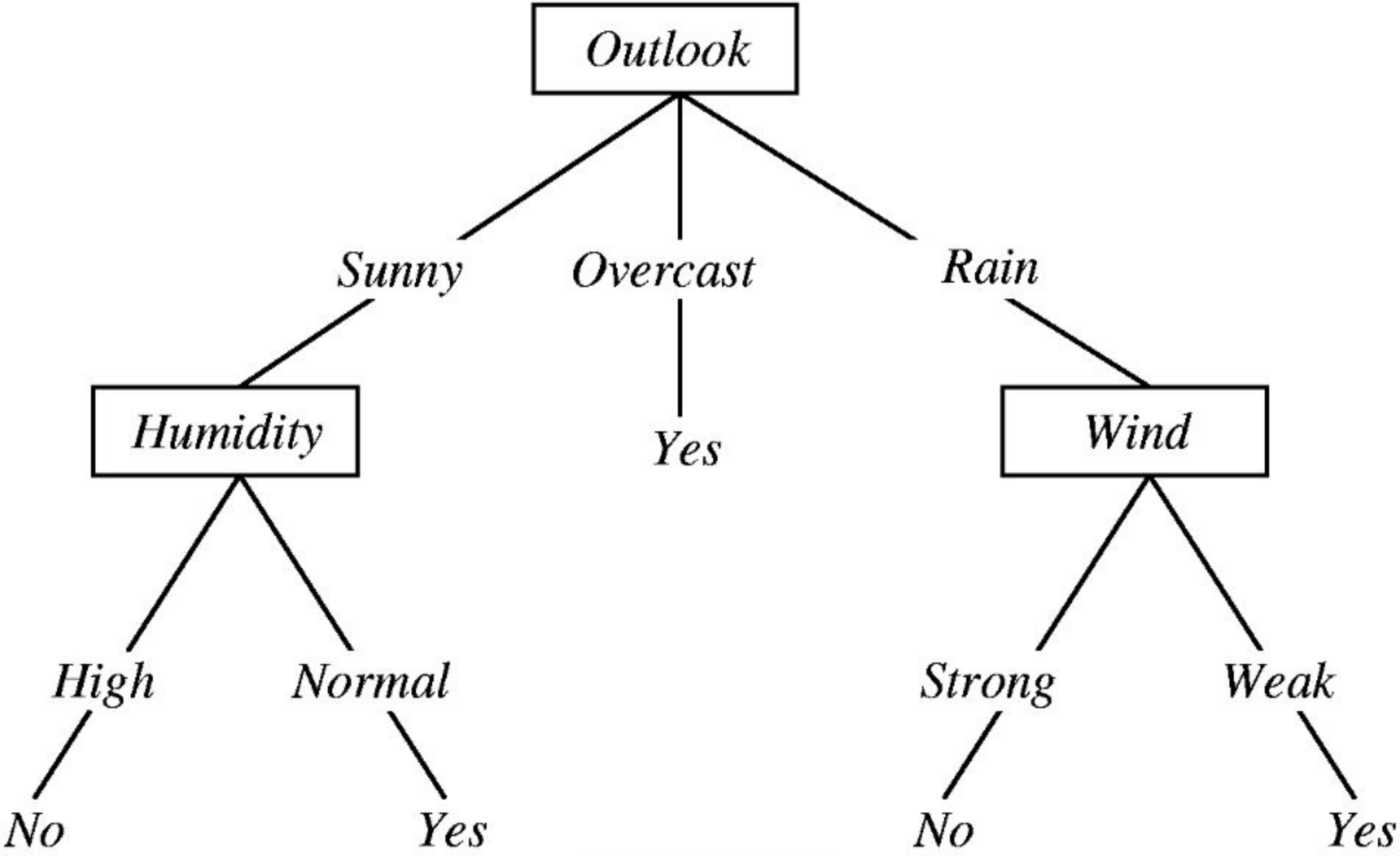
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

- **Advantages of KNN Algorithm:**
- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.
- **Disadvantages of KNN Algorithm:**
- Always needs to determine the value of K which may be complex sometimes.
- The computation cost is high because of calculating the distance between the data points for all the training samples

Decision Tree

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.


Decision Node
Root Node
Sub-Tree
Decision Node
Decision Node
Leaf Node
Leaf Node
Leaf Node
Decision Node
Leaf Node
Leaf Node


Outlook
Sunny
Overcast
Rain
Humidity
Yes
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes

Decision tree terminologies

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

How does decision tree works

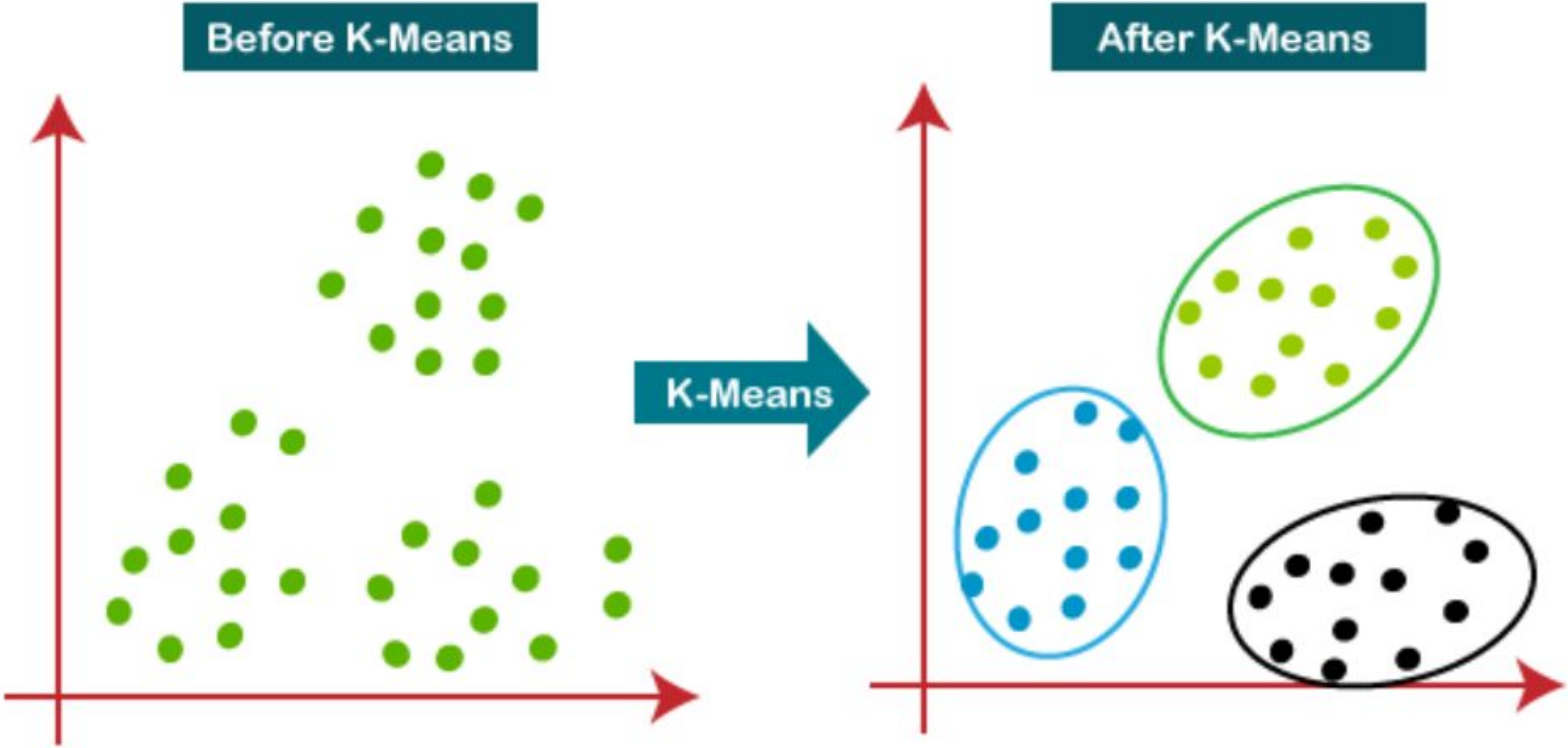
- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using the **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contain possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

- Advantages of the Decision Tree
- It is simple to understand as it follows the same process that a human follows while making any decision in real life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement for data cleaning compared to other algorithms.

- 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.

K-Means clustering

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.


Before K-Means
K-Means
After K-Means

How does K-means algorithm works

- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to its closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid in each cluster.
- **Step-5:** Repeat the third step, which means reassigning each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step 4 else go to FINISH.
- **Step-7**: The model is ready.

Cluster Analysis

- In cluster analysis, the K-means algorithmcan be used to partition the input dataset into k-partitions (clusters).
- The parameter k is known to be hard to choose when not given by external constraints.
- Another initiation is that it cannot be used with arbitrary distance functions or on non-numerical data. For these use cases, many other algorithms are superior.

Feature Learning

- K-means clustering has been used as a feature learning step, in either supervised or unsupervised learning.
- Approach is to first train a k-means clustering representation, using the input training data. Then to project any input datum into the new feature space, an “encoding function” , such as the threshold matrix-product of the datum with the centroid locations, computes the distance from the datum to each centroid, or simply an indicator function for the nearest cenroid.

Application

- *k*-means clustering is rather easy to apply to even large data sets, particularly when using heuristics such as Lloyd's algorithm. It has been successfully used in market segmentation, computer vision, and astronomy among many other domains. It often is used as a preprocessing step for other algorithms, for example, to find a starting configuration.

Clustering applications

- **In Identification of Cancer Cells:** The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
- **In Search Engines:** Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

- **Customer Segmentation:** It is used in market research to segment the customers based on their choice and preferences.
- **In Biology:** It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- **In Land Use:** The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

Analysis Tool – Power BI

- Power BI is a Business Intelligence and Data Visualization tool for converting data from various data sources into interactive dashboards and analysis reports. Power BI offers cloud-based services for interactive visualizations with a simple interface for end-users to create their own reports and dashboards
- Different Power BI versions like Desktop, Service-based (SaaS), and mobile Power BI apps are used for different platforms. It provides multiple software connectors and services for business intelligence.

Why use Power BI

- Pre-built dashboards and reports for SaaS Solutions
- Power BI allows real-time dashboard updates.
- Offers Secure and reliable connection to your data sources in the cloud or on-premises
- Power BI offers Quick deployment, hybrid configuration, and a secure environment.
- Allows data exploration using natural language query
- Offers feature for dashboard visualization regularly updated with the community.

Types of Power BI Tools

- **Power BI Desktop**
- Power BI desktop is the primary authoring and publishing tool for Power BI. Developers and power users use it to create brand new models and reports from scratch.
- **Power BI service**
- Online Software as a Service (SaaS) where Powe BI data models, reports, and dashboards are hosted. Administration, sharing, and collaboration happen in the cloud.

- **Power BI Data Gateway**
- Power BI Data Gateway works as the bridge between the Power BI Service and on-premise data sources like DirectQuery, Import, and Live Query. It is Installed by BI Admin.
- **Power BI Report Server**
- It can host paginated reports, KPIs, mobile reports, & Power BI Desktop reports. It is updated every 4 months and installed/managed by the IT team. The users can modify Power BI reports and other reports created by the development team.

- **Power BI Mobile Apps**
- Power BI mobile app is available for iOS, Android, and Windows. It can be managed using Microsoft Intune. You can use this tool to view reports and dashboards on the Power Bl Service Report Server.

Key terms used in Power BI

Visualizations

Datasets

Reports

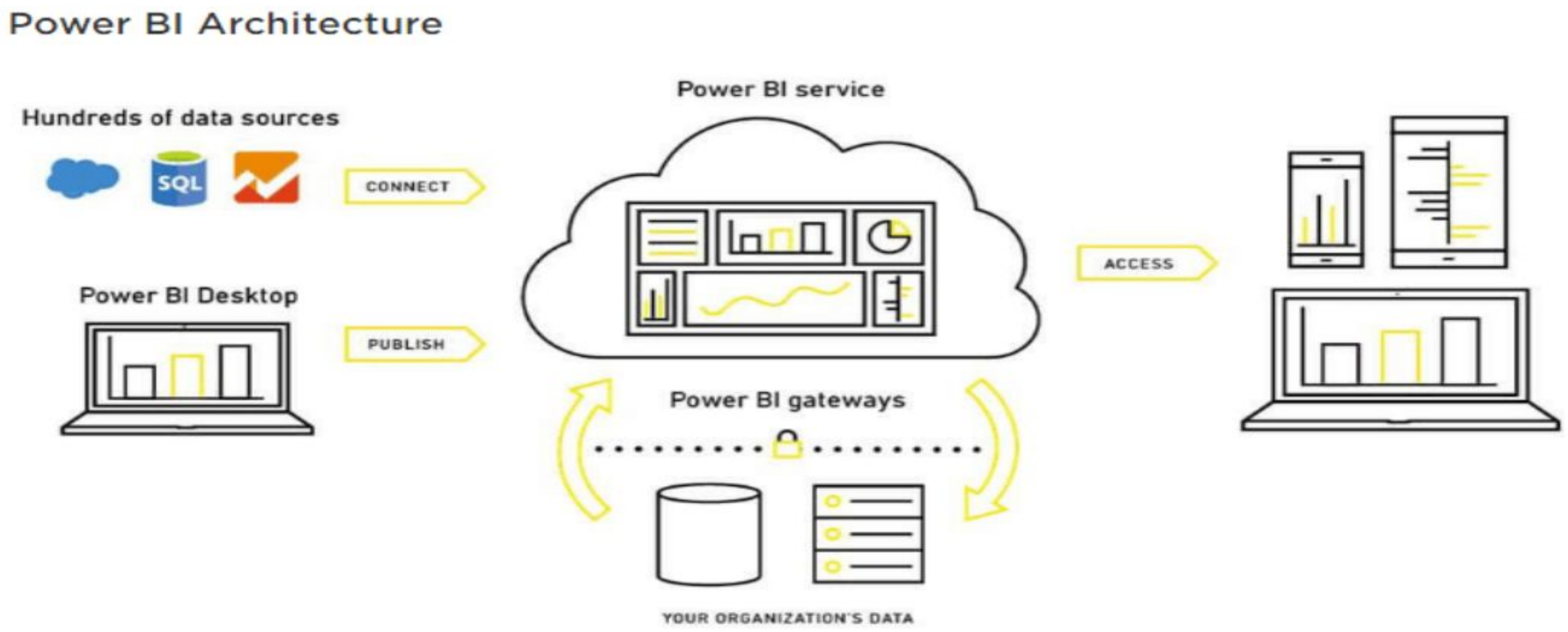
Dashboards

Tiles

Term	Description
Visualization	A visual display of information to achieve one or more objective. It offers a single-screen display of information. It alerts users on issues or problems Operational, Performance, Personal, etc.
Datasets	A dataset is something which you import or connect to. Datasets can be renamed, refreshed, removes, and explored.
Dashboard	The dashboard is a collection which contains zero or more tiles and widgets. It is used to represent a customized view of some subset of the underlying datasets.

Reports	A Power BI report is one or multiple pages of visualizations. It can be created from scratch, imported to a dashboard, and created using datasets.
Tile	It a single visualization found in a report or on a rectangular dashboard box which contains each visual.

Architecture of Power BI



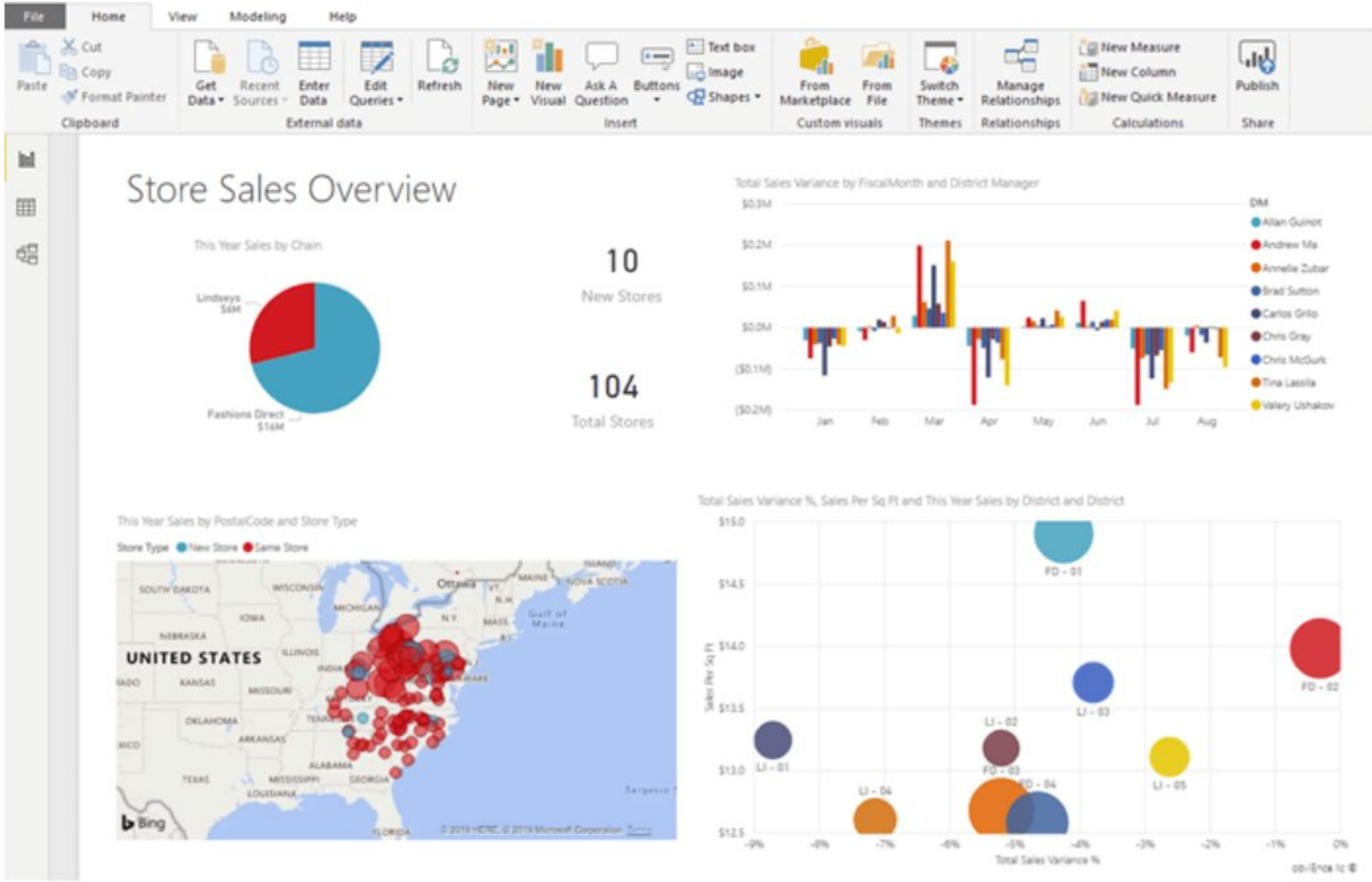
Architecture of Power BI

- **Data Integration:**
- An organization needs to work with data that comes from different sources in various file formats. The data should be extracted from a different source which can be from different servers or databases. This data is integrated into one standard format in a common staging area.

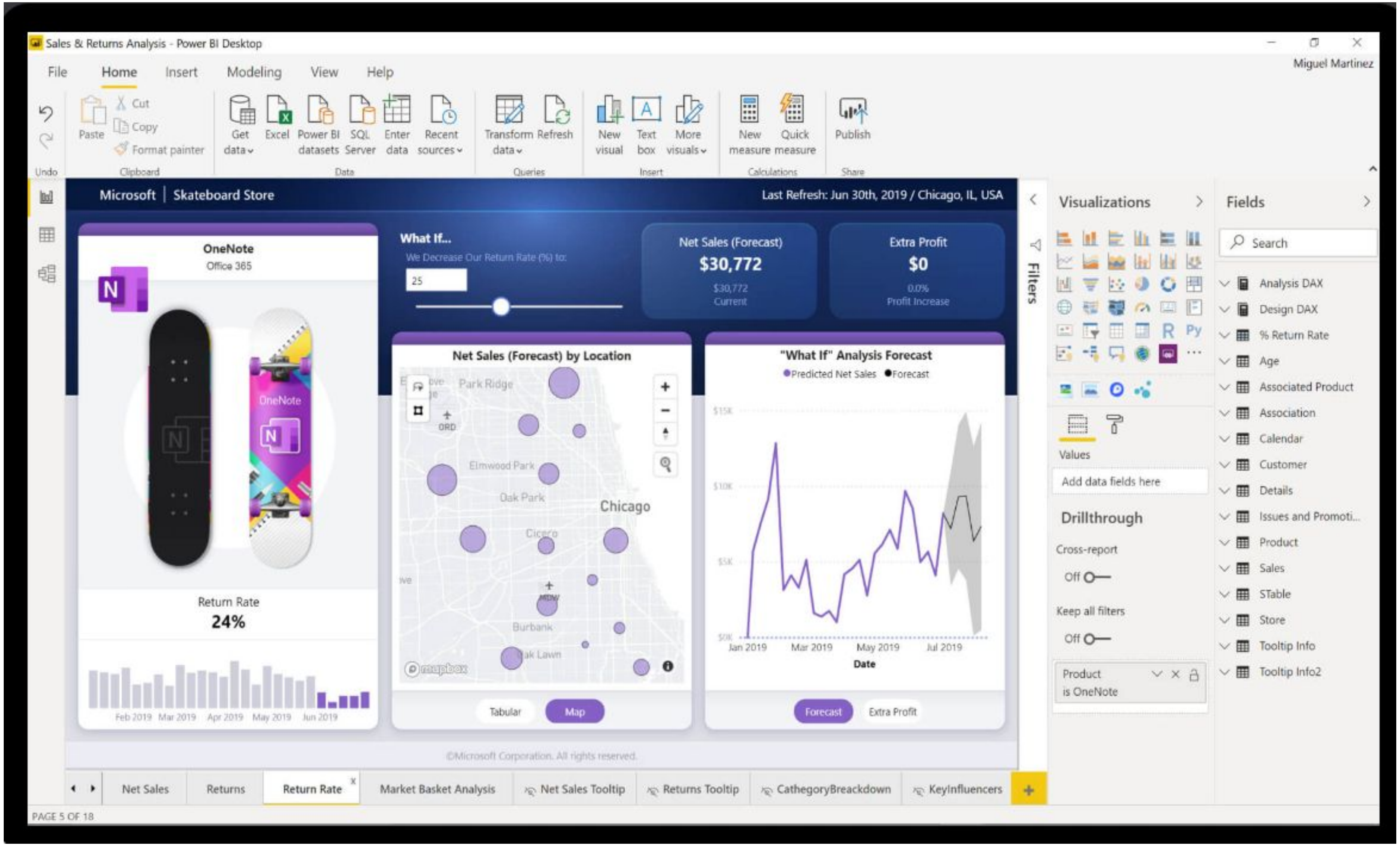
- **Data Processing:**
- In this stage, the integrated data is still not prepared for visualization as the data needs processing. This data is pre-processed. For example, redundant values, and missing values will be removed from the data set.
- The business rule should be applied to the data when the data is cleaned. You can load that data back to Data Warehouse.

- **Data Presentation**
- Once the data is loaded and processed, it can be visualized much better with the use of various visualization that Power Bi has to offer. The use of a dashboard and report helps one represent data more intuitively. This visual report helps business end-users to take business decisions based on the insights.

Power BI deshboard



Power BI desktop



DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in