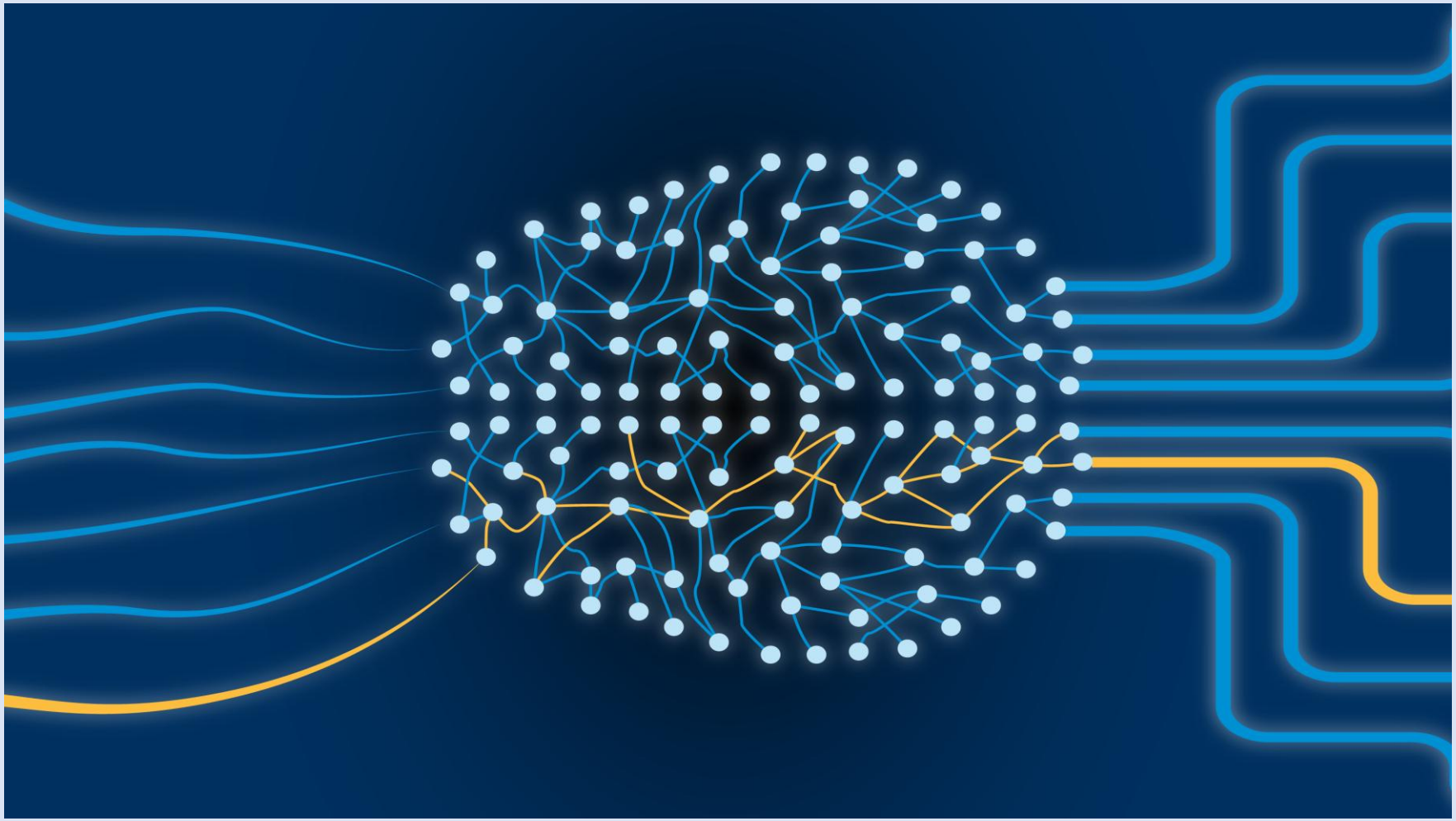
SUPERVISED LEARNING



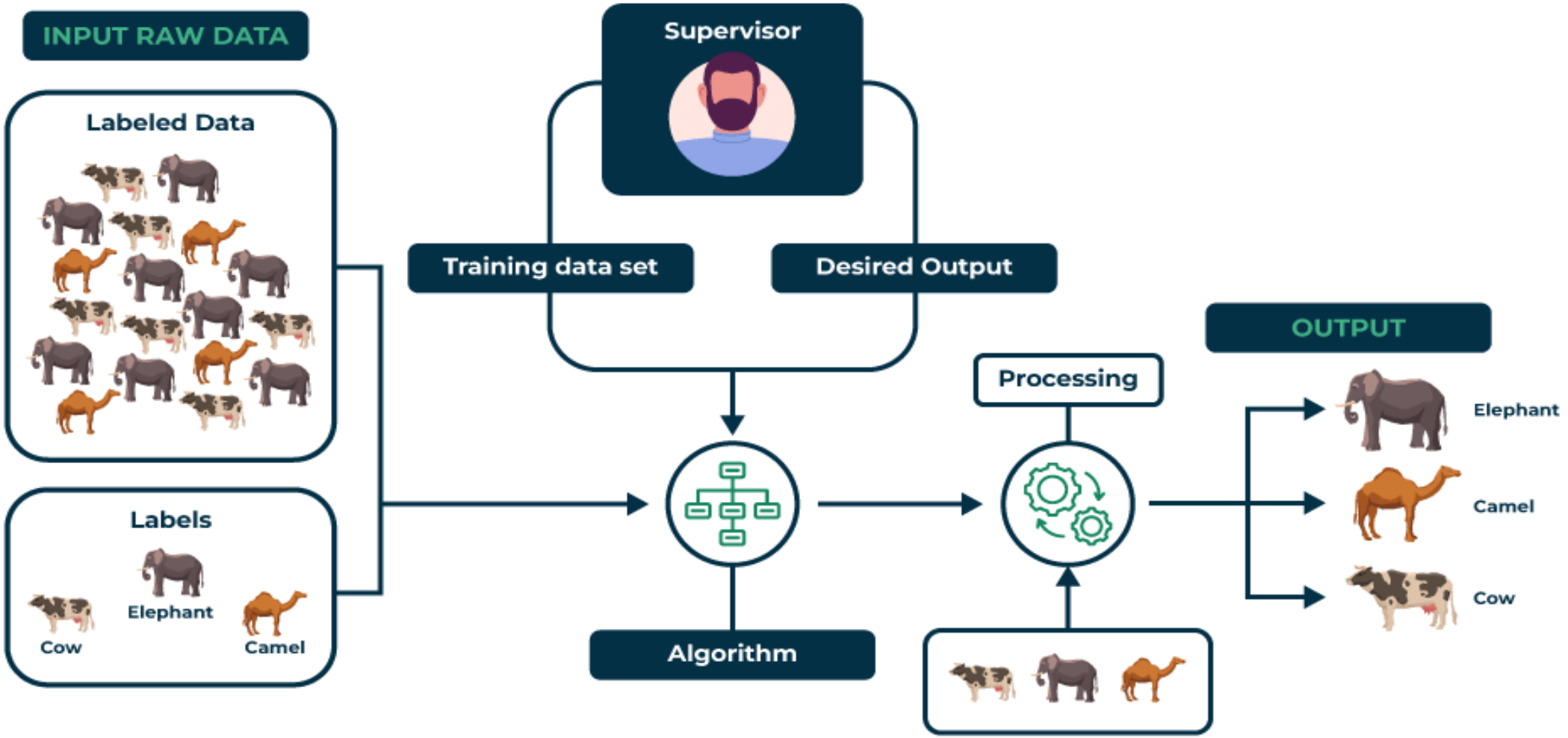
By Sithara Madhubhashini

What is Supervised Learning?

- **supervised learning** is a type of machine learning where a model is trained on labeled data.
- Meaning each input is paired with the correct output.
- The model learns by comparing its predictions with the actual answers provided in the training data.
- Over time, it adjusts itself to minimize errors and improve accuracy.

The goal of supervised learning is to make accurate predictions when given new, unseen data

Supervised Learning



How Supervised Machine Learning Works?

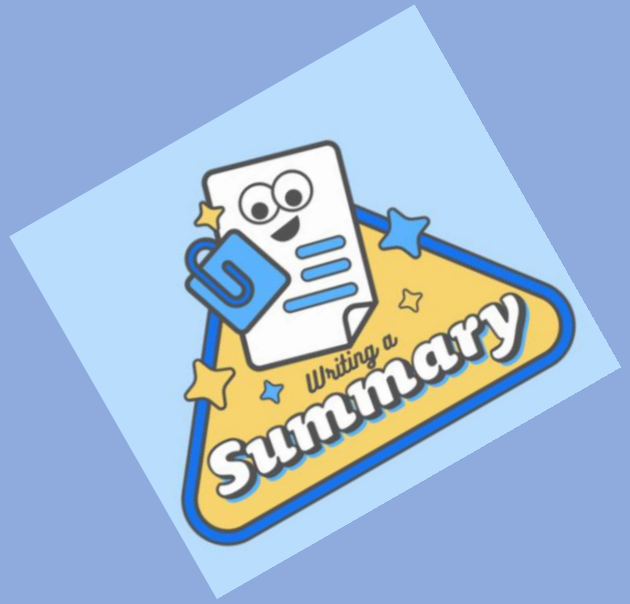
Training Data: The model is provided with a training dataset that includes input data (features) and corresponding output data (labels or target variables).

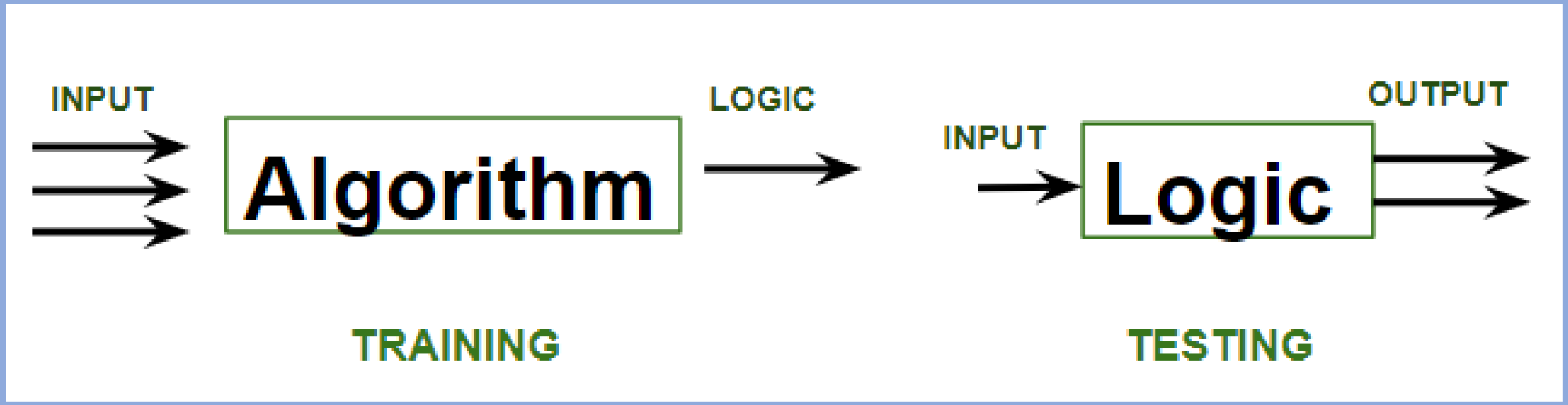
Learning Process: The algorithm processes the training data, learning the relationships between the input features and the output labels.

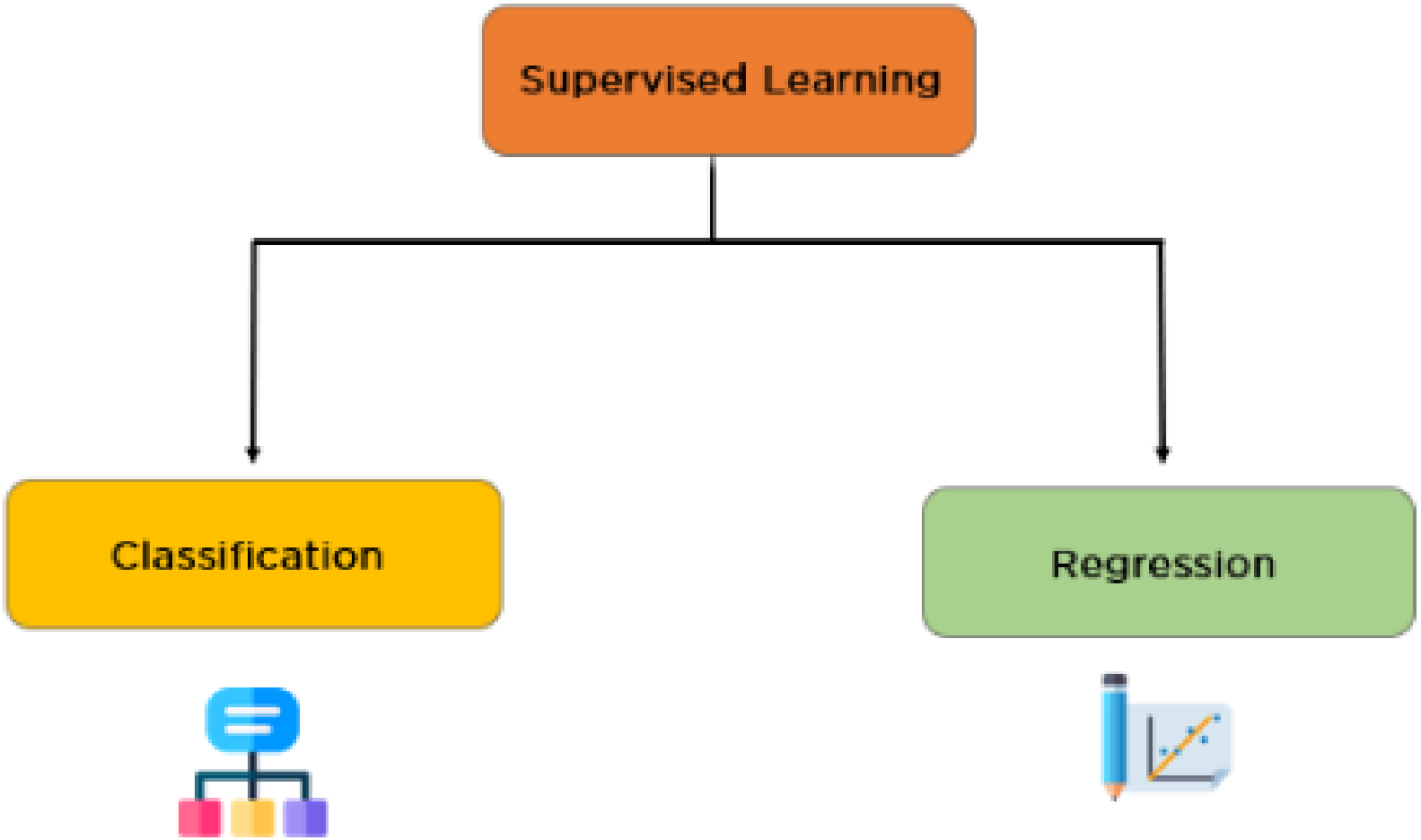
This is achieved by adjusting the model's parameters to minimize the difference between its predictions and the actual labels.

How Supervised Machine Learning Works?

- After training, the model is evaluated using a test dataset to measure its accuracy and performance.
- Then the model's performance is optimized by adjusting parameters and using techniques like cross-validation to balance bias and variance.


Writing a
Summary


INPUT
Algorithm
LOGIC
TRAINING
INPUT
Logic
OUTPUT
TESTING


Supervised Learning
Classification
Regression

Classification:

Where the output is a categorical variable (e.g., spam vs. non-spam emails, yes vs. no).

Regression:

Where the output is a continuous variable (e.g., predicting house prices, stock prices).

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

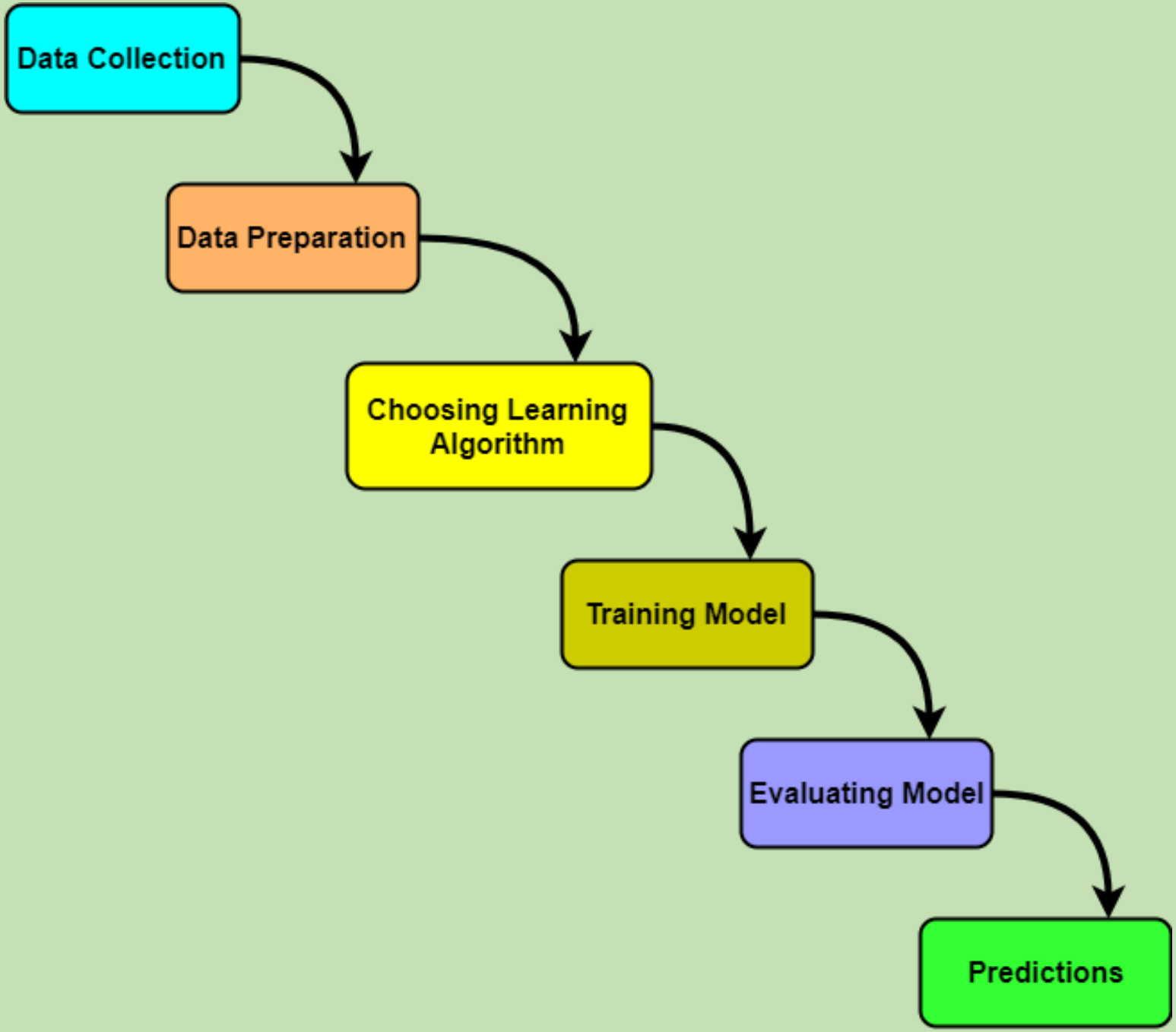
Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

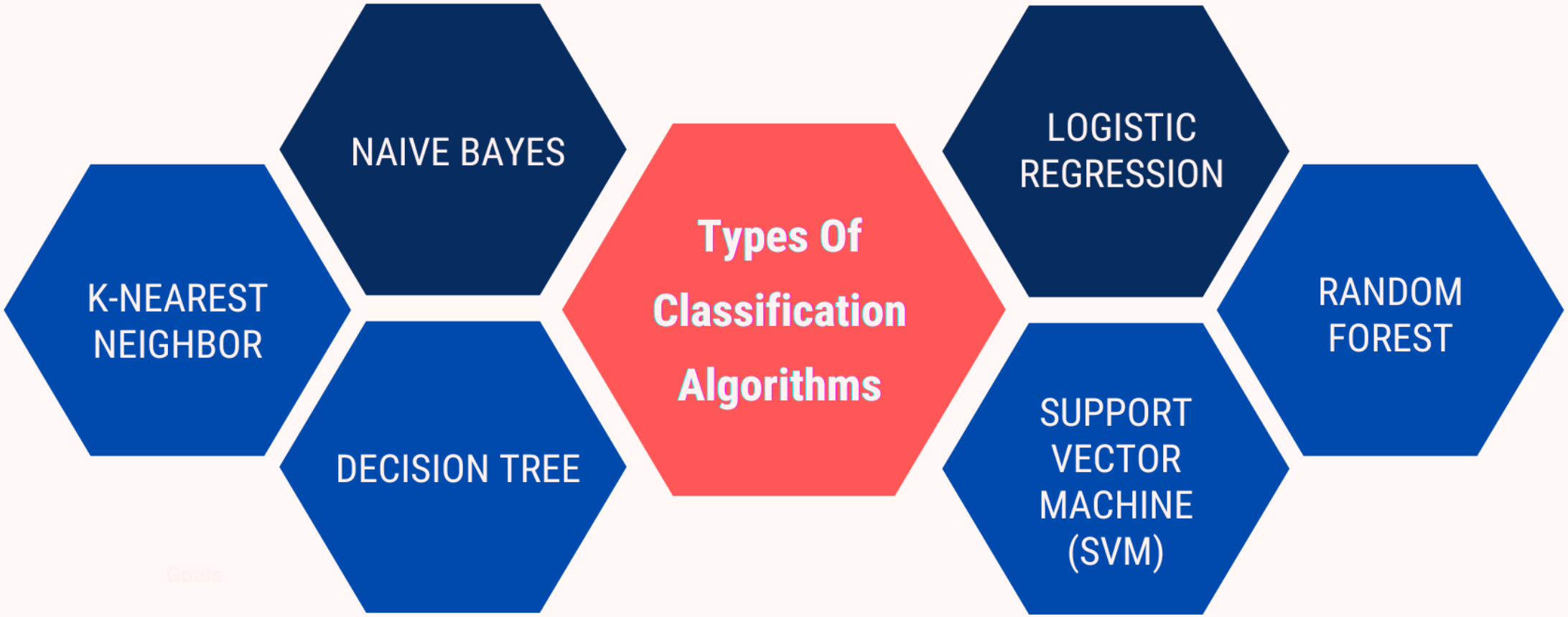
Figure B: REGRESSION

Training a Supervised Learning Model

Key Steps ????


Data Collection
Data Preparation
Choosing Learning Algorithm
Training Model
Evaluating Model
Predictions

Classification Algorithms



Classification Algorithms

Classification teaches a machine to sort things into categories.

Example:

A classification model might be trained on dataset of images labeled as either **dogs** or **cats** and it can be used to predict the class of new and unseen images as dogs or cats based on their features such as color, texture and shape.

TYPES OF CLASSIFICATION

1. Binary Classification

Classifying into two categories

Spam
Not Spam

Positive
/ Negative

Default
No Default

Default /
No Default

2. Multiclass Classification

Involves more than two classes

Digit
Recognition

Sentiment
Analysis

Animal
Classification

Dog,
Bird

3. Multilabel Classification

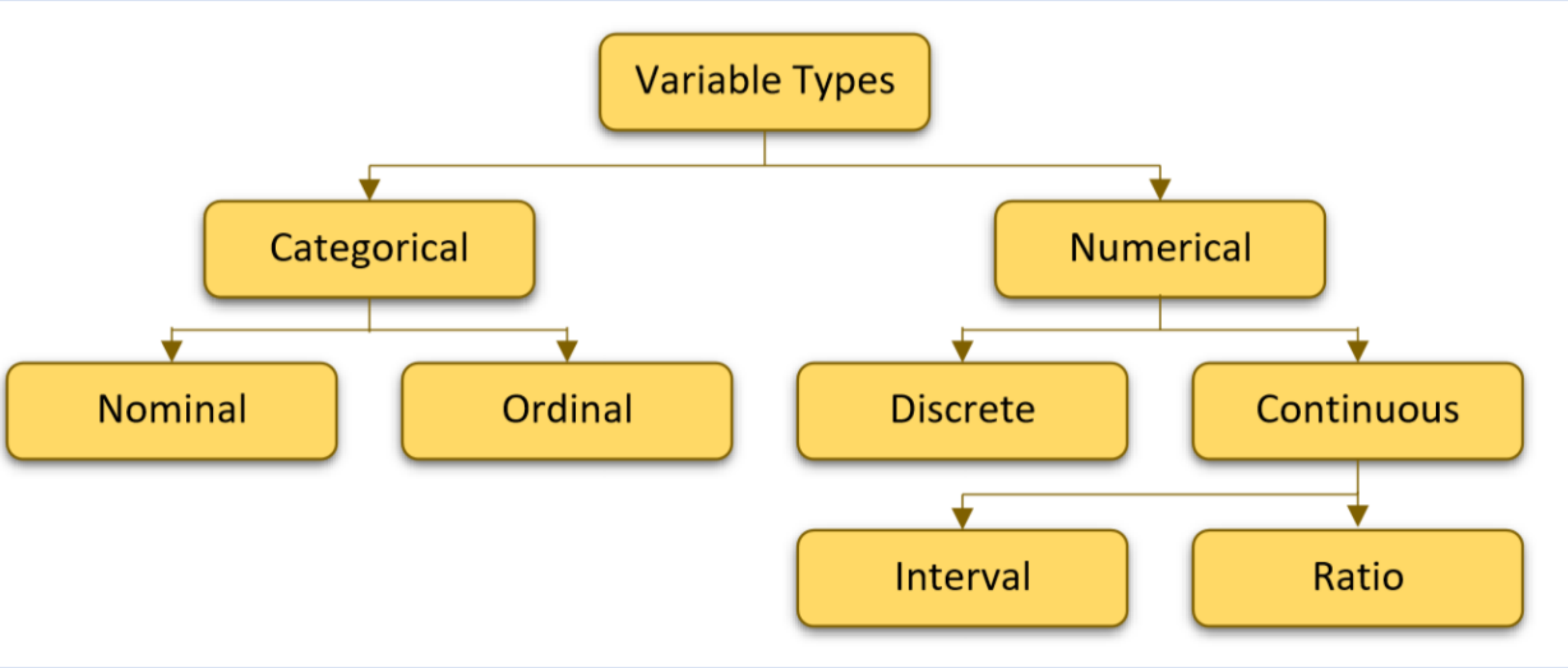
Each instance has multiple classes

Tagging

Image
Tagging

Music
Genre

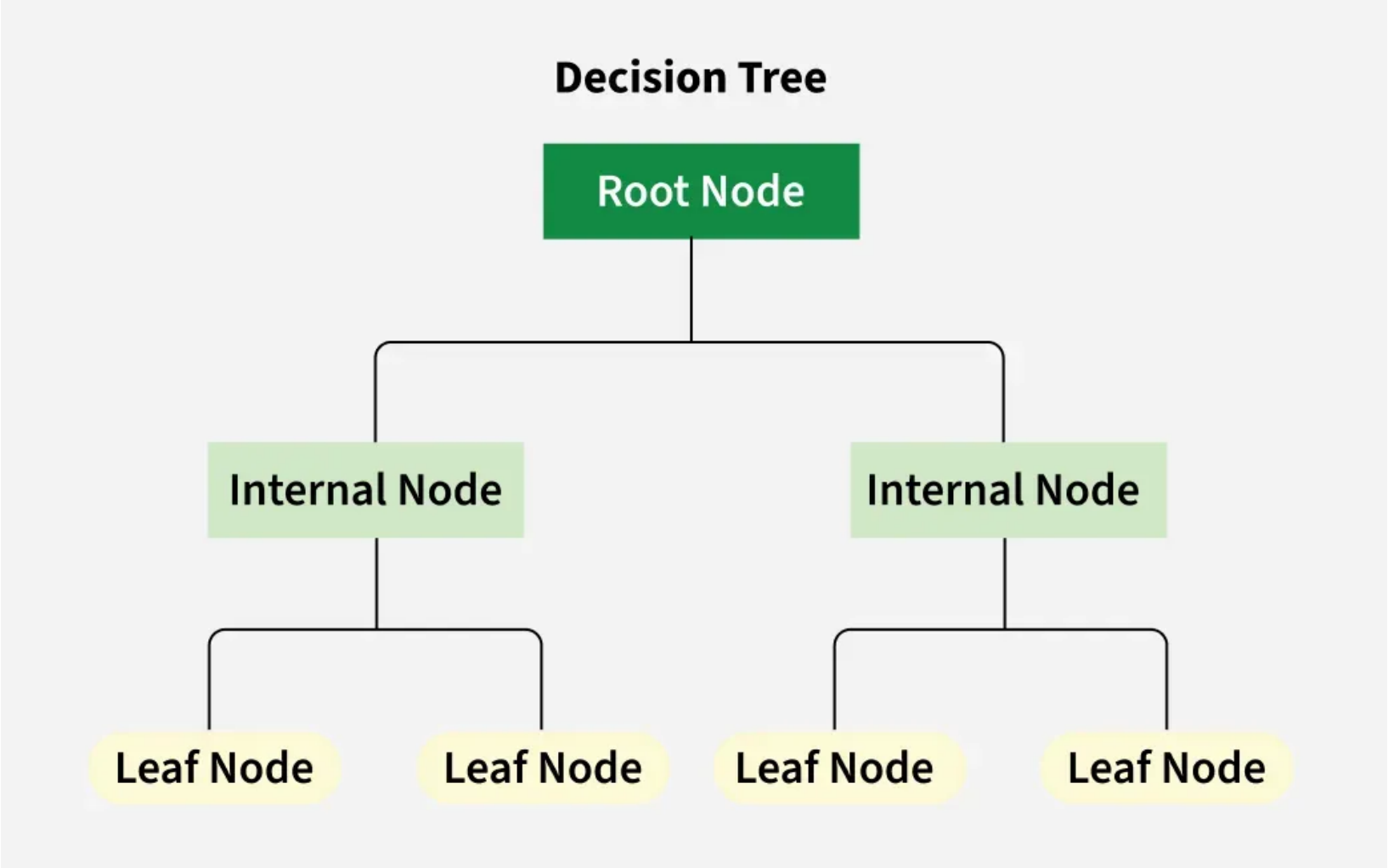
Image
Tagging


Variable Types
Categorical
Numerical
Nominal
Ordinal
Discrete
Continuous
Interval
Ratio

Activity

Examples of Machine Learning Classification in Real Life?????

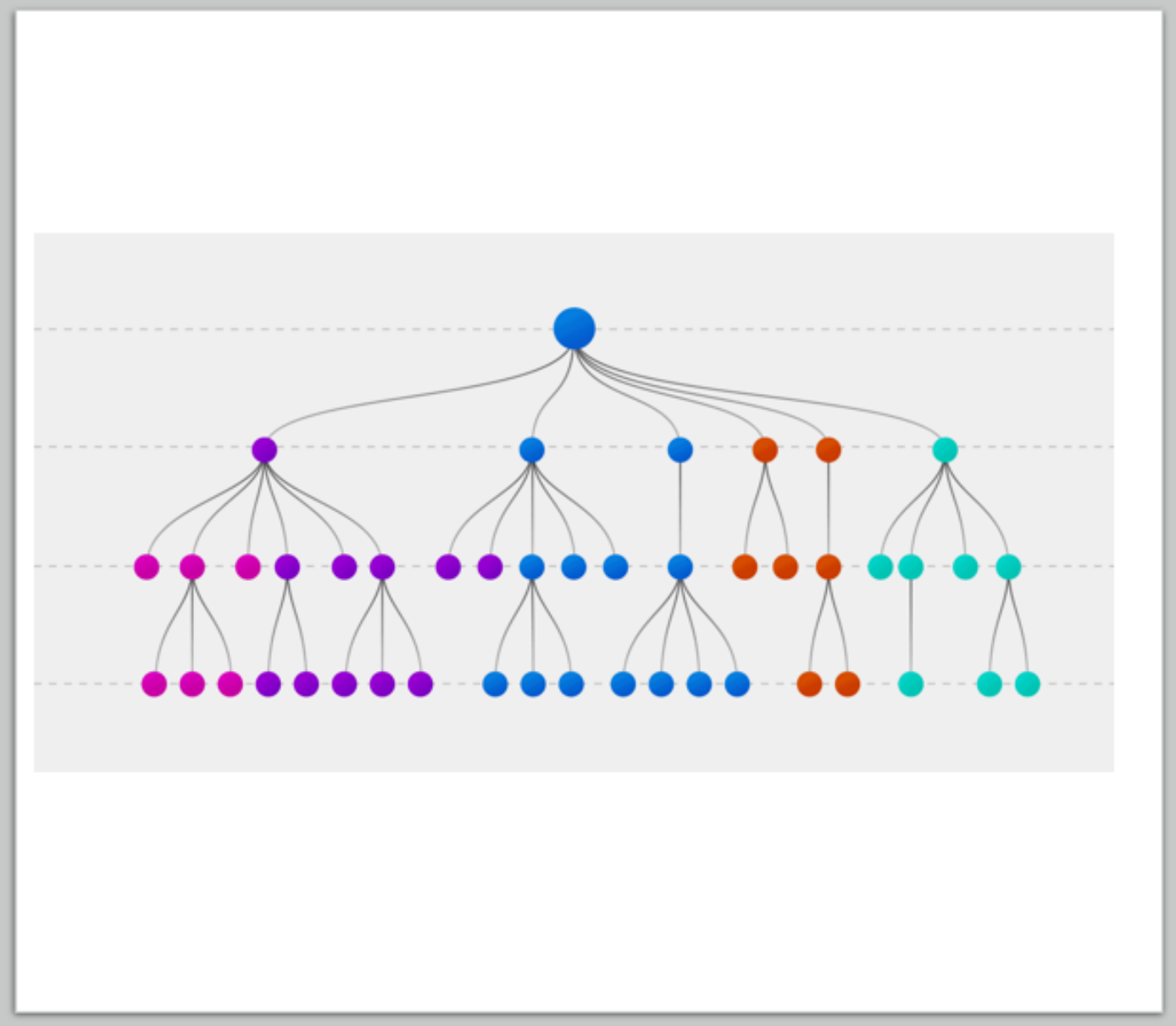
Decision Tree Algorithm



What is Decision Tree algorithm?

- A non-parametric supervised learning algorithm.
- Utilized for both classification and regression tasks.
- Has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.
- Based on the available features, Internal nodes conduct evaluations to form homogenous subsets, which are denoted by leaf nodes.
- Leaf nodes represent all the possible outcomes within the dataset.
- White box type ML algorithm

- Each leaf node represents a Class Label
- The paths from root to leaf represent Classification Rules
- Decision tree is one of the predictive modelling approaches used in Statistics & Data Mining apart from ML.



- **Root Node** : It represents the entire population or sample, and this further gets divided into two or more homogeneous sets.

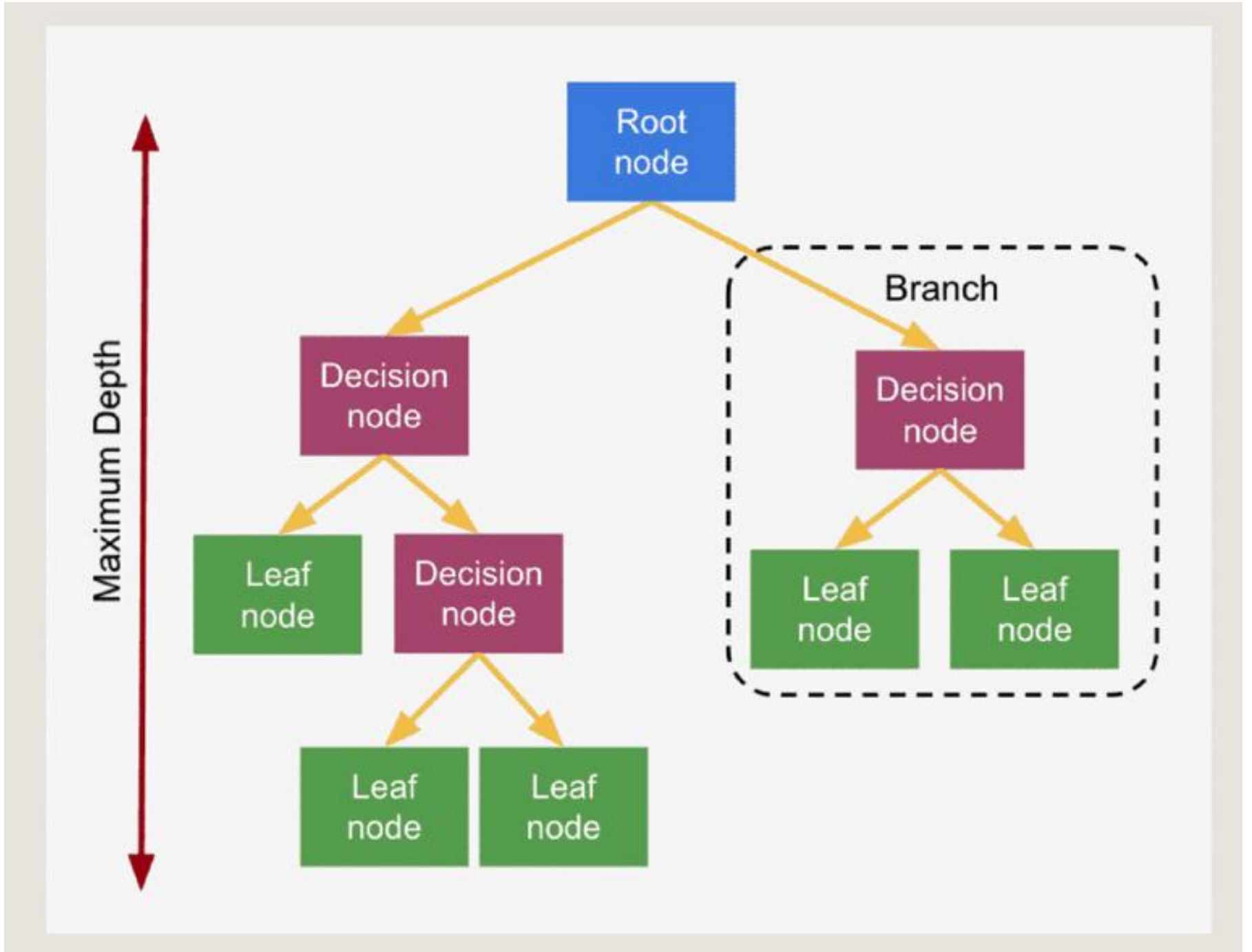
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.

- **Leaf / Terminal Node:** Nodes do not split is called Leaf or Terminal node.

- **Splitting:** It is a process of dividing a node into two or more sub- nodes.

- **Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.

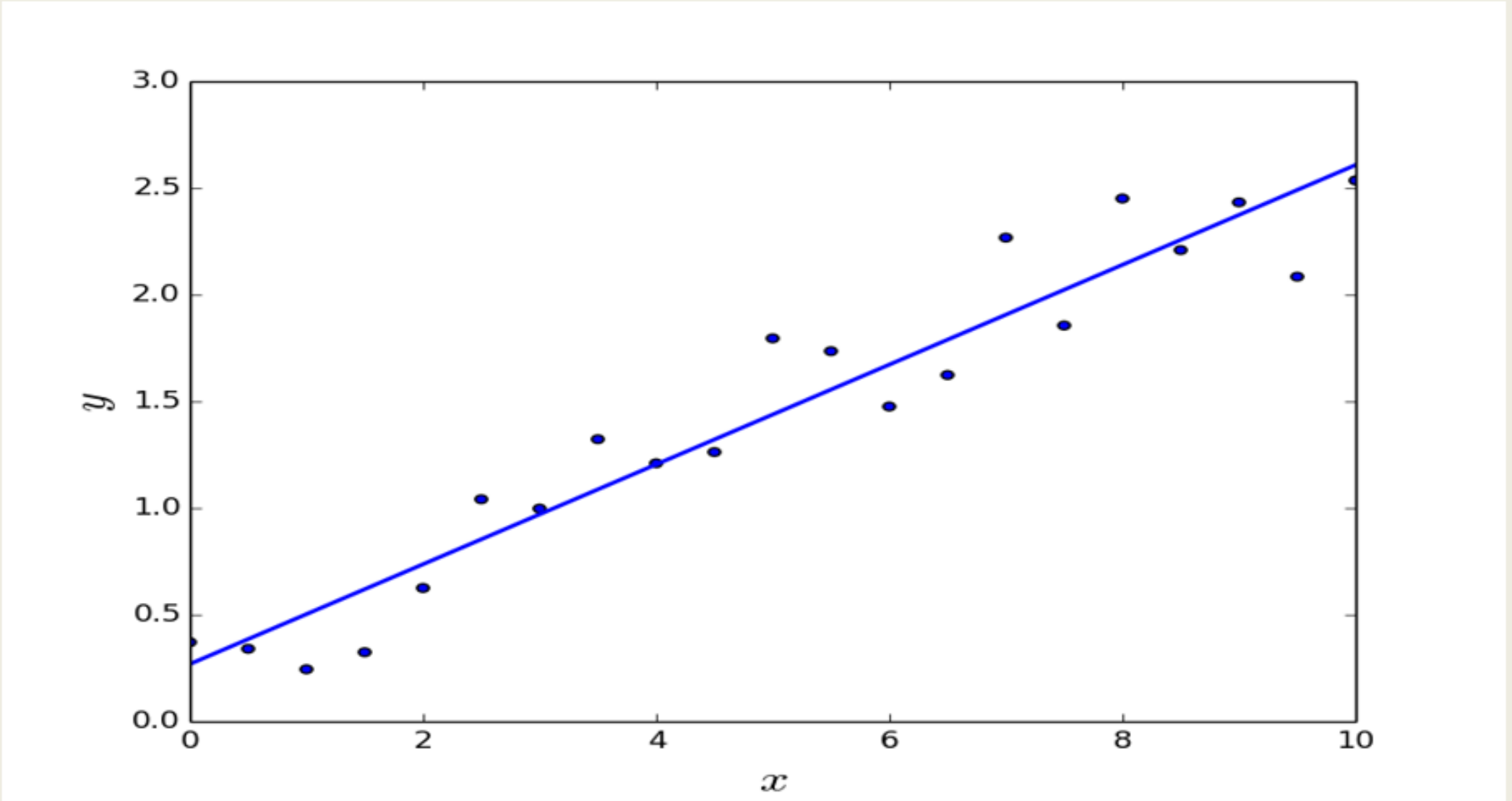
- **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.


Root node
Decision node
Branch
Decision node
Leaf node
Decision node
Leaf node
Leaf node
Leaf node
Leaf node
Maximum Depth

- Parametric Models : Parametric models are those that require the specification of some parameters before they can be used to make predictions.

- Non-parametric models : do not rely on any specific parameter settings and therefore often produce more accurate results

Parametric vs Non-Parametric Models



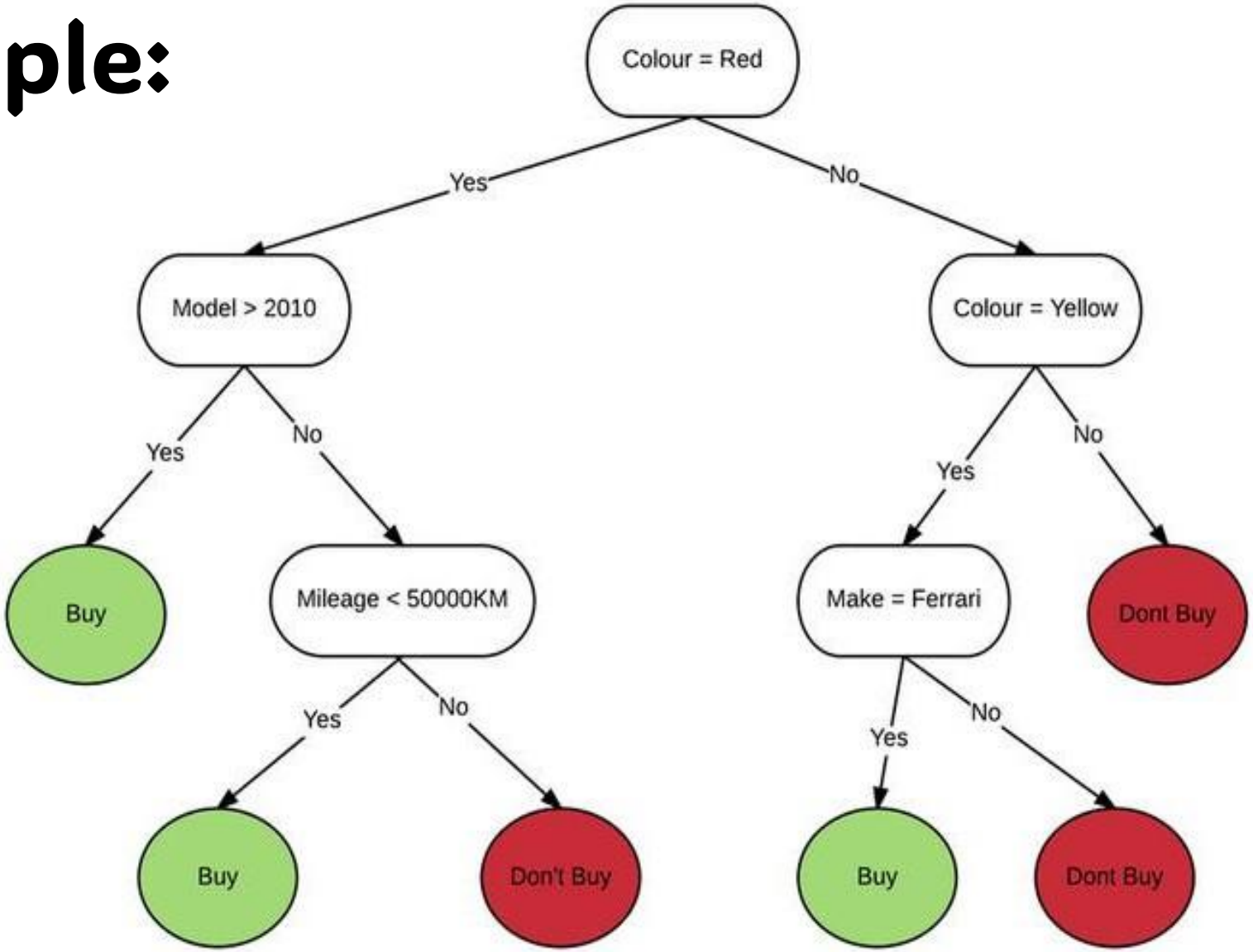
Classification Trees:

- Used for predicting categorical outcomes like spam or not spam.
- These trees split the data based on features to classify data into predefined categories.

Regression Trees:

- Used for predicting continuous outcomes like predicting house prices.
- Instead of assigning categories, it provides numerical predictions based on the input features.

Example:


Colour = Red
Yes
No
Model > 2010
Colour = Yellow
Yes
No
Yes
No
Buy
Mileage < 50000KM
Make = Ferrari
Dont Buy
Yes
No
Yes
No
Buy
Don't Buy
Buy
Dont Buy

How Decision Trees Work?

1. Start with the Root Node: It begins with a main question at the root node which is derived from the dataset's features.

2. Ask Yes/No Questions: From the root, the tree asks a series of yes/no questions to split the data into subsets based on specific attributes.

3. Branching Based on Answers: Each question leads to different branches:

- If the answer is yes, the tree follows one path.
- If the answer is no, the tree follows another path.

4. Continue Splitting: This branching continues through further decisions helps in reducing the data down step-by-step.

5. Reach the Leaf Node: The process ends when there are no more useful questions to ask leading to the leaf node where the final decision or prediction is made.

Splitting Criteria in Decision Trees

- In a Decision Tree, the process of splitting data at each node is important.

- The splitting criteria finds the best feature to split the data on.

Attribute Selection Measures

- Entropy
- Information Gain
- Gini Index
- Reduction in Variance
- Chi Square

Common splitting criteria

- **Gini Impurity**: This criterion measures how "impure" a node is. The lower the Gini Impurity the better the feature splits the data into distinct categories.

- **Entropy**: This measures the amount of uncertainty or disorder in the data. The tree tries to reduce the entropy by splitting the data on features that provide the most information about the target variable.

Overfitting in Decision Trees

- Overfitting occurs when a tree becomes too deep and starts to memorize the training data rather than learning general patterns.
- This leads to poor performance on new, unseen data.
- Pruning is an important technique used to prevent overfitting in Decision Trees.
- This technique reduces the complexity of the tree by removing branches that have little predictive power.

- It improves model performance by helping the tree generalize better to new data. It also makes the model simpler and faster to deploy.

- It is useful when a Decision Tree is too deep and starts to capture noise in the data.

Reasons for Overfitting:

- High variance and low bias.
- The model is too complex.
- The size of the training data.

Pre – Pruning

- This technique refers to the early stopping mechanism
- do not allow the training process to go through, consequently preventing the overfitting of the model.

Post Pruning

- This technique allows decision trees to grow to their full depth in the training process.
- then starts removing the branches of the trees to prevent the model from overfitting.

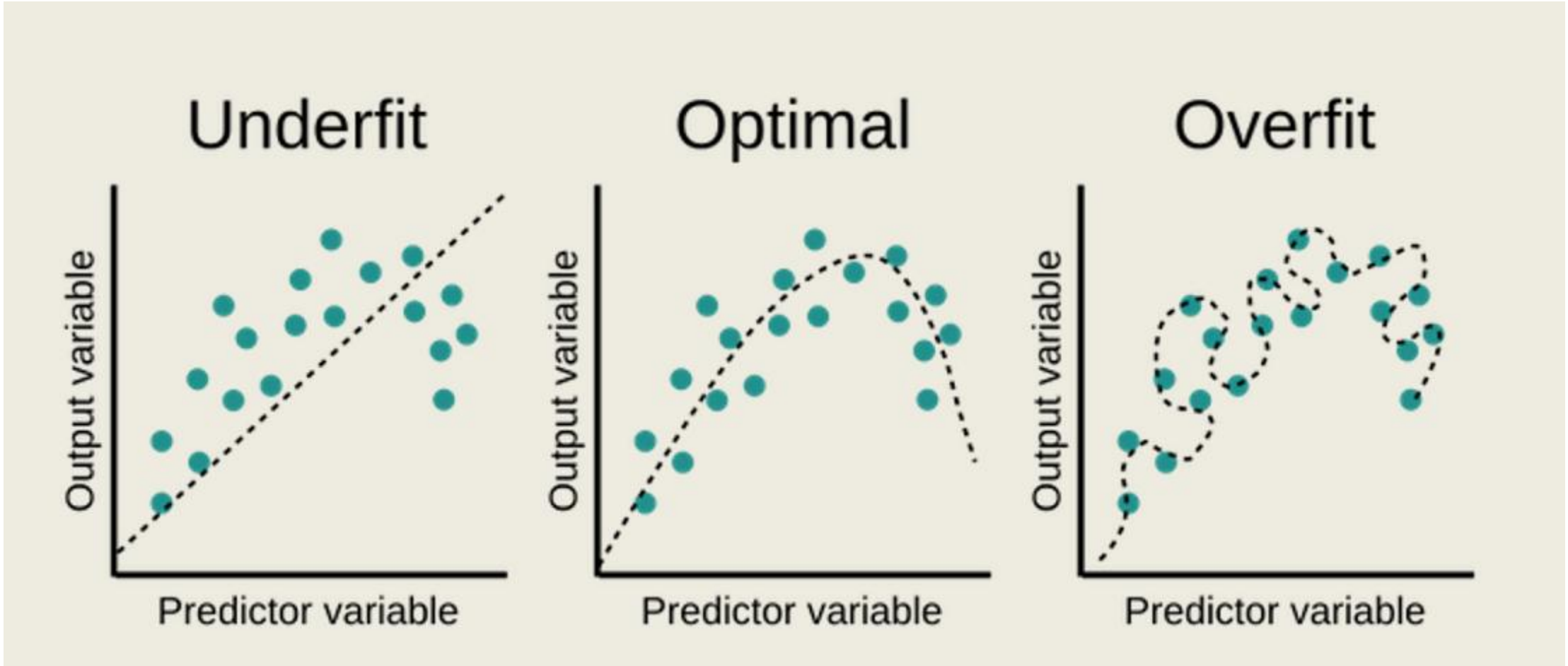
Underfitting in Decision Trees

Underfitting is the opposite of overfitting.

It happens when a model is too simple to capture what's going on in the data.

Reasons for Underfitting:

- The model is too simple, So it may be not capable to represent the complexities in the data.
- The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
- The size of the training dataset used is not enough.
- Features are not scaled.


Underfit
Optimal
Overfit
Output variable
Predictor variable
Output variable
Predictor variable
Output variable
Predictor variable