Introduction to Machine Learning

Quantitative Data Analysis for Education Research

Chaitanya Jadhav

AI Engineer, PhillipCapital

Julian Tan

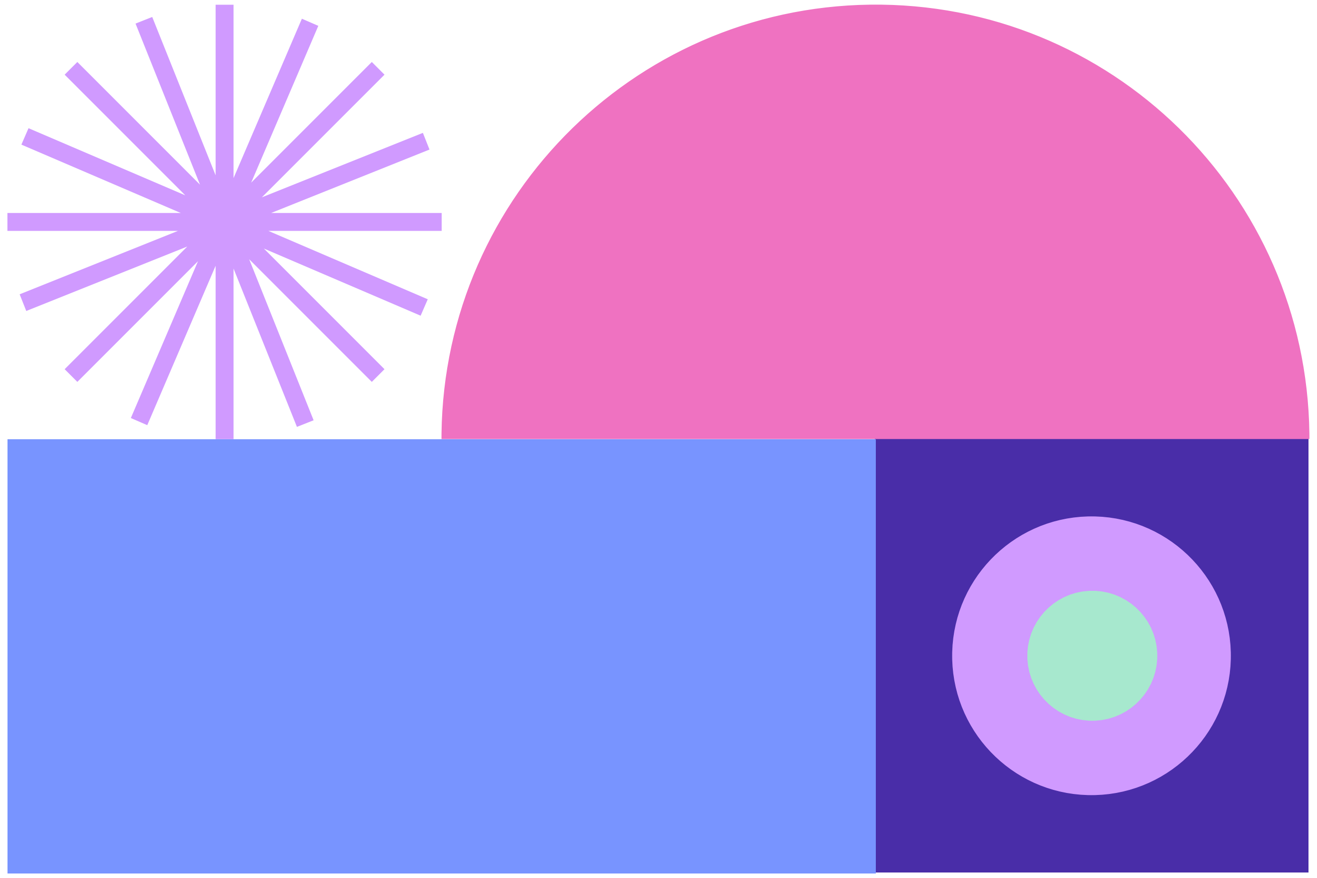
PhD Student, PESS

John Komar

Assistant Dean, Research Support

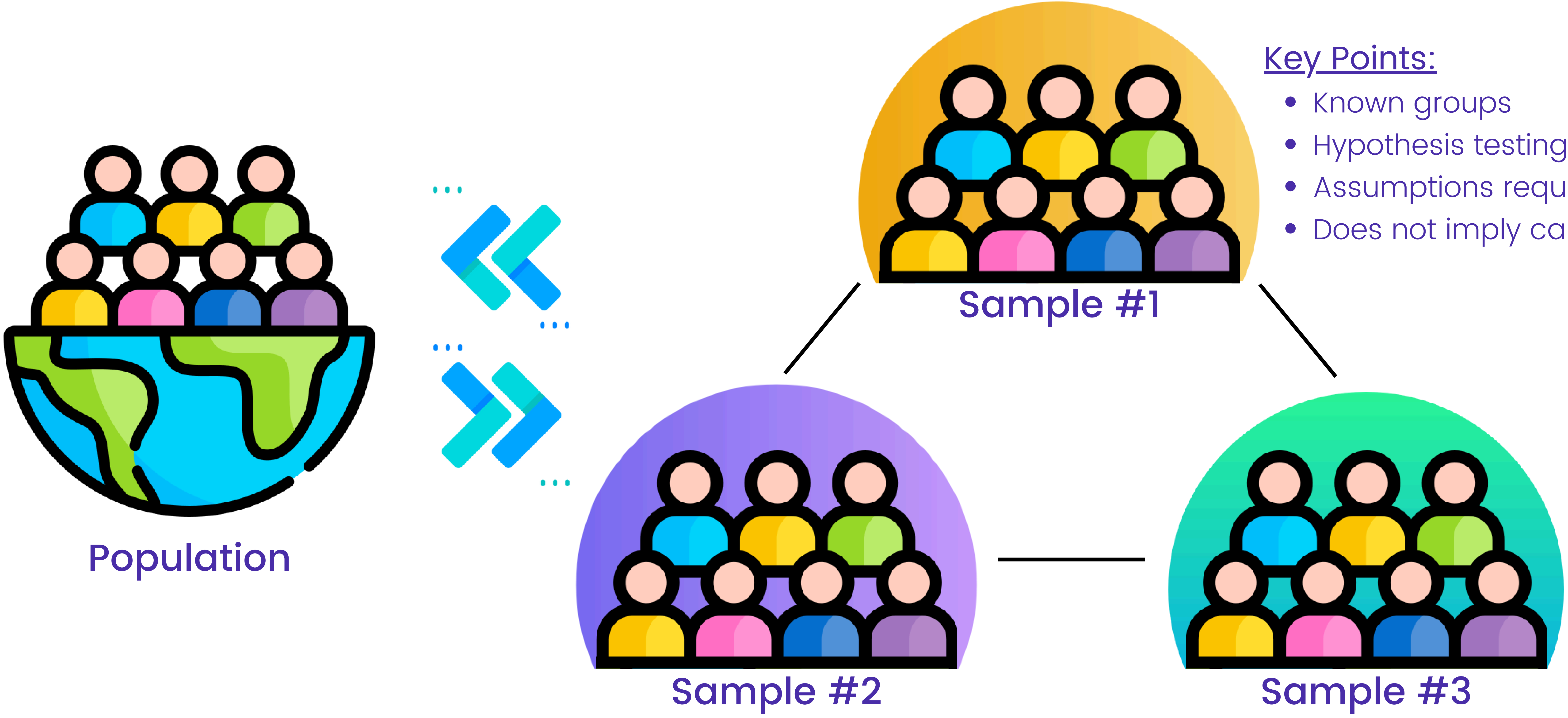
Workshop Structure

Agenda



- → Inferential Statistics vs. Machine Learning
- → Supervised vs. Unsupervised Learning
- → Introduction to K Means
- → Introduction to rule mining with decision trees
- → Apply techniques through Python and Google Colab

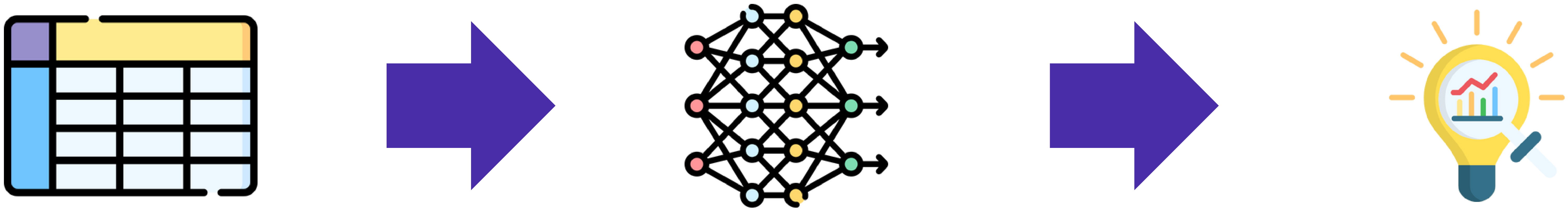
Inferential Statistics



Key Points:

- Known groups
- Hypothesis testing
- Assumptions required
- Does not imply cause & effect

Motivation



Data

Model

Insights

Use Cases:

Identify student learning patterns

Predict academic performance

Support targeted interventions

Enhance curriculum design

Enable evidence-based policy decisions

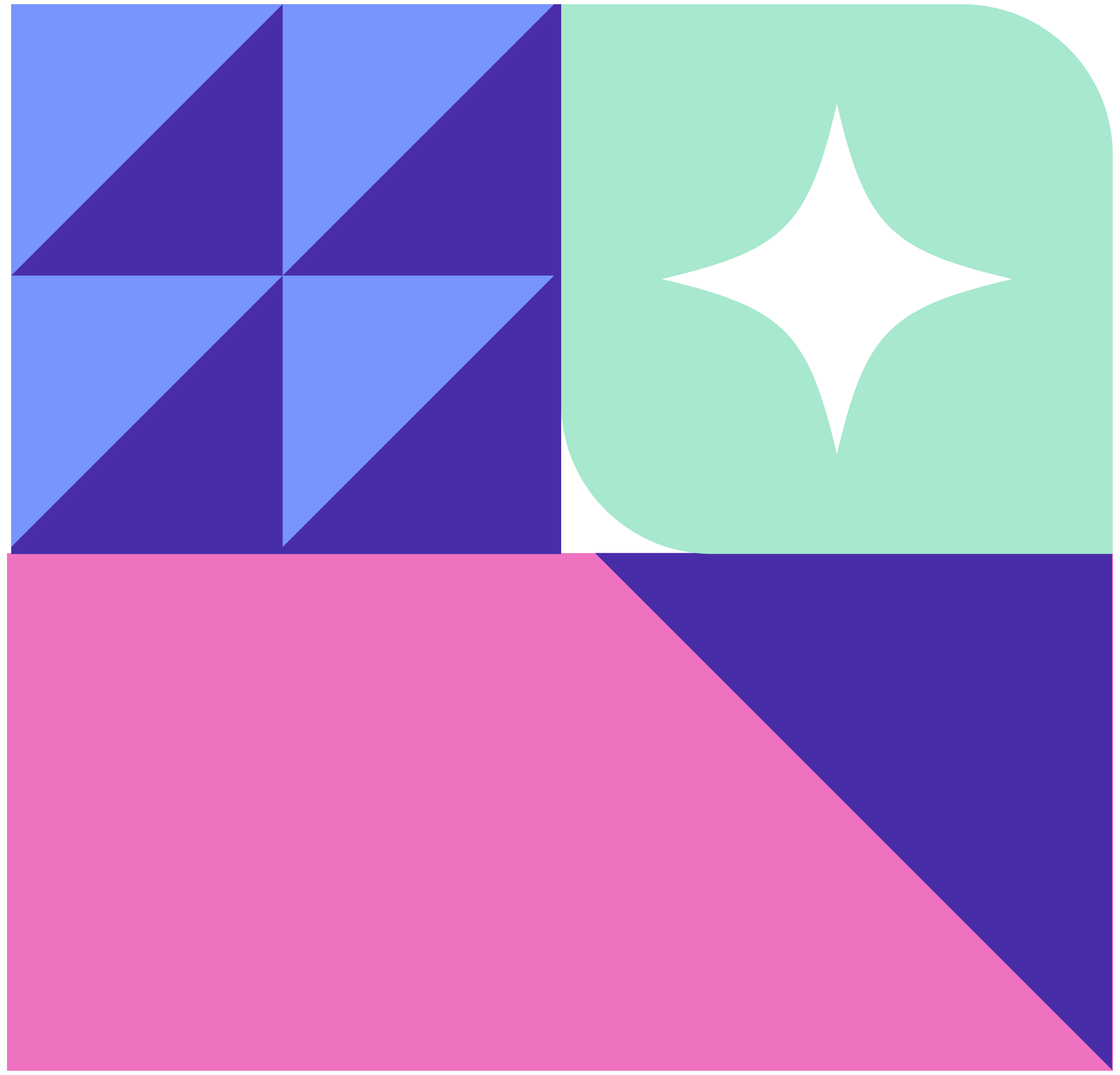
Supervised Learning

"I see both inputs and outputs, so I learn to map one to the other."

Unsupervised Learning

"I only see inputs, so I have to figure out the structure myself."

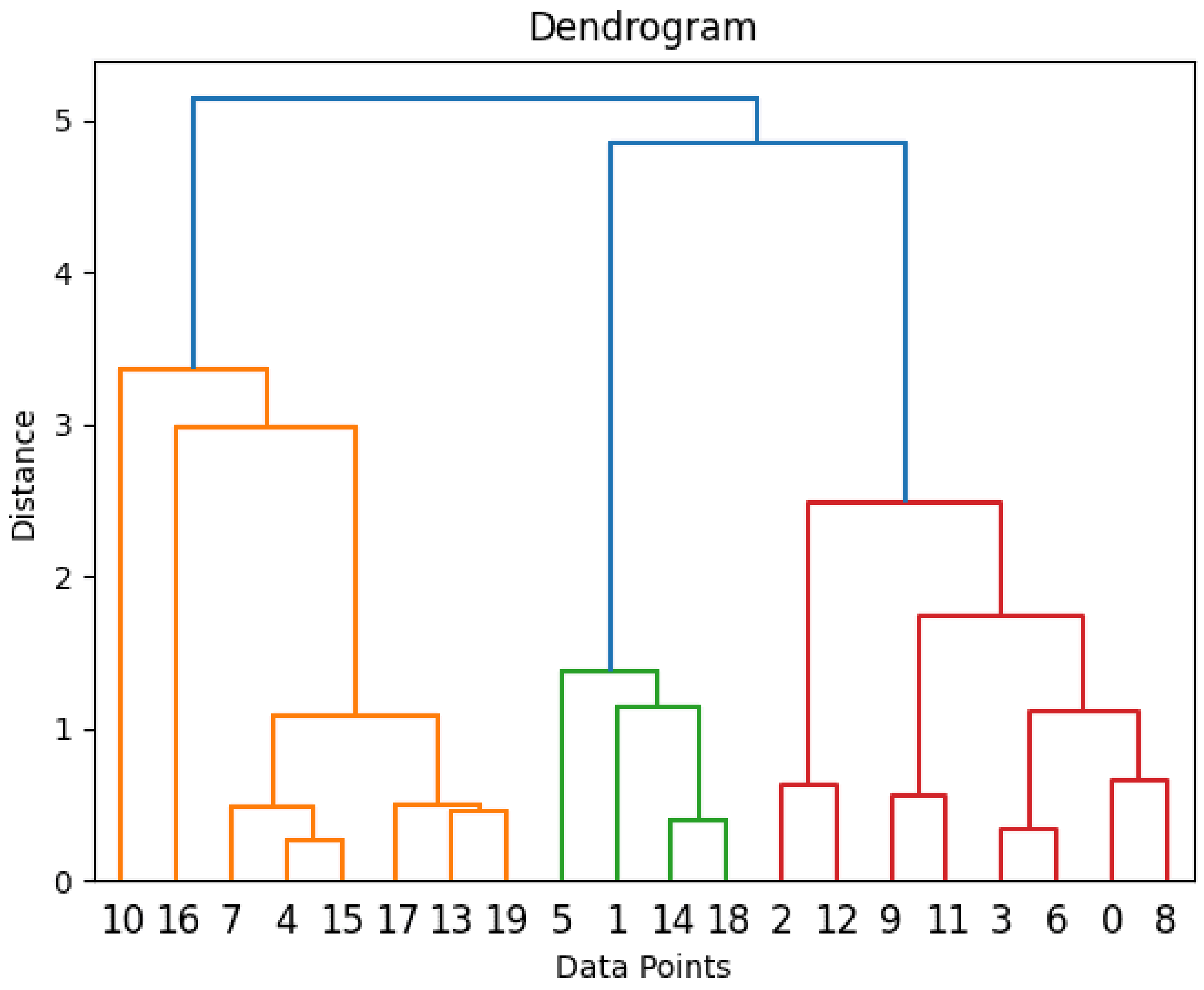
https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning



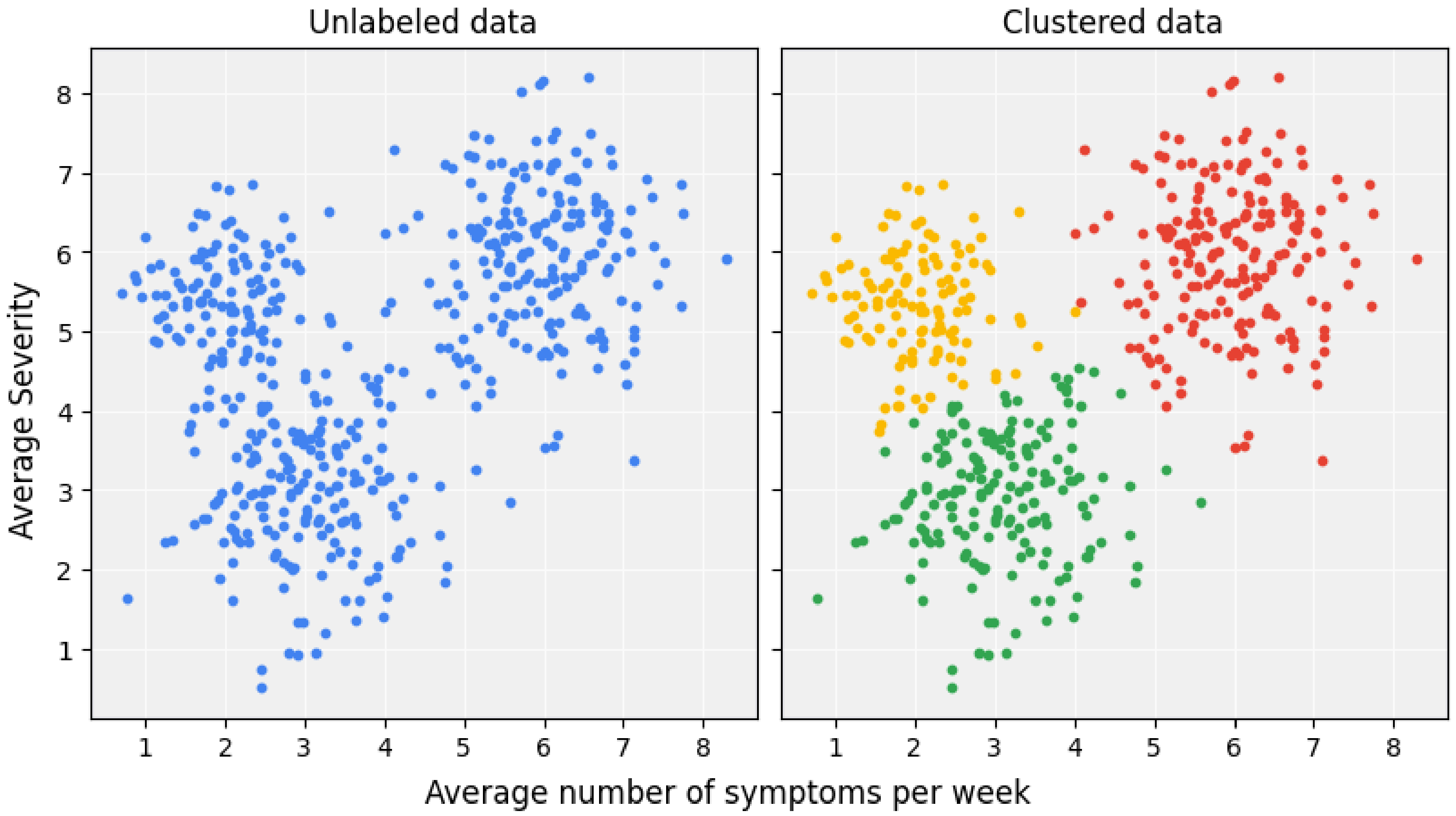
Applications of Unsupervised Learning

1. Student Segmentation
2. Curriculum Evaluation
3. Early Risk Detection
4. Personalized Learning
5. Data Exploration

Clustering



Hierarchical Clustering (HCA)



K-Means Clustering

https://developers.google.com/machine-learning/clustering/overview
https://developers.google.com/machine-learning/clustering/clustering-algorithms

Clustering Methods

Aspect	HCA	K-Means
Use Case	Small to medium datasets	Large Datasets
Pros	No need to define the number of clusters	Fast, scales well with large datasets
Cons	Slow for large datasets Sensitive to noise.	Must choose K Best suited for spherical datasets

https://developers.google.com/machine-learning/clustering/overview

Data Pre-Processing

01

Handle NULL values

02

Check Data Types

03

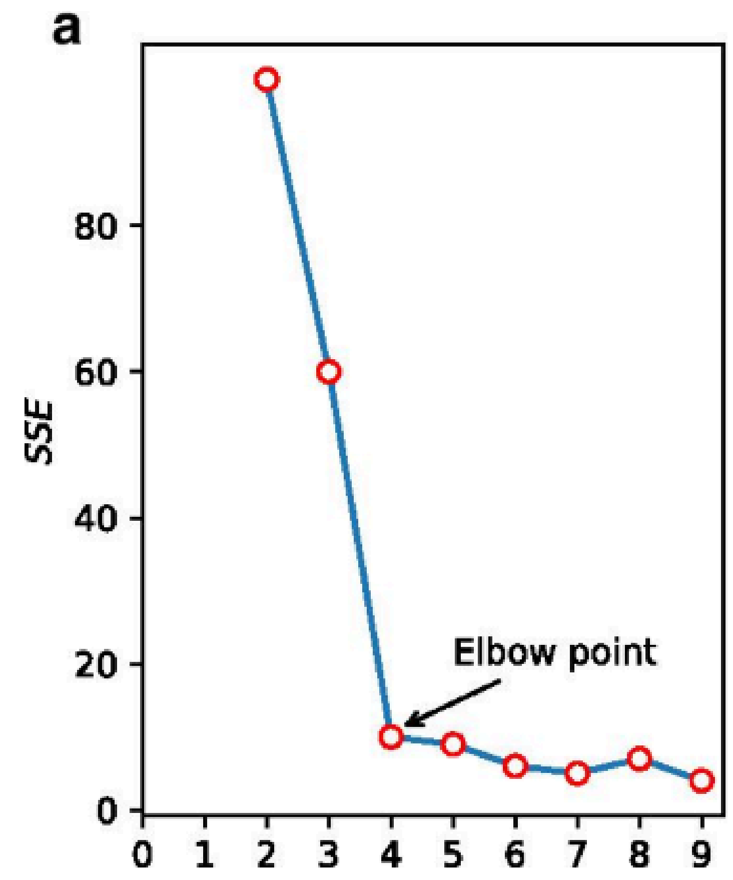
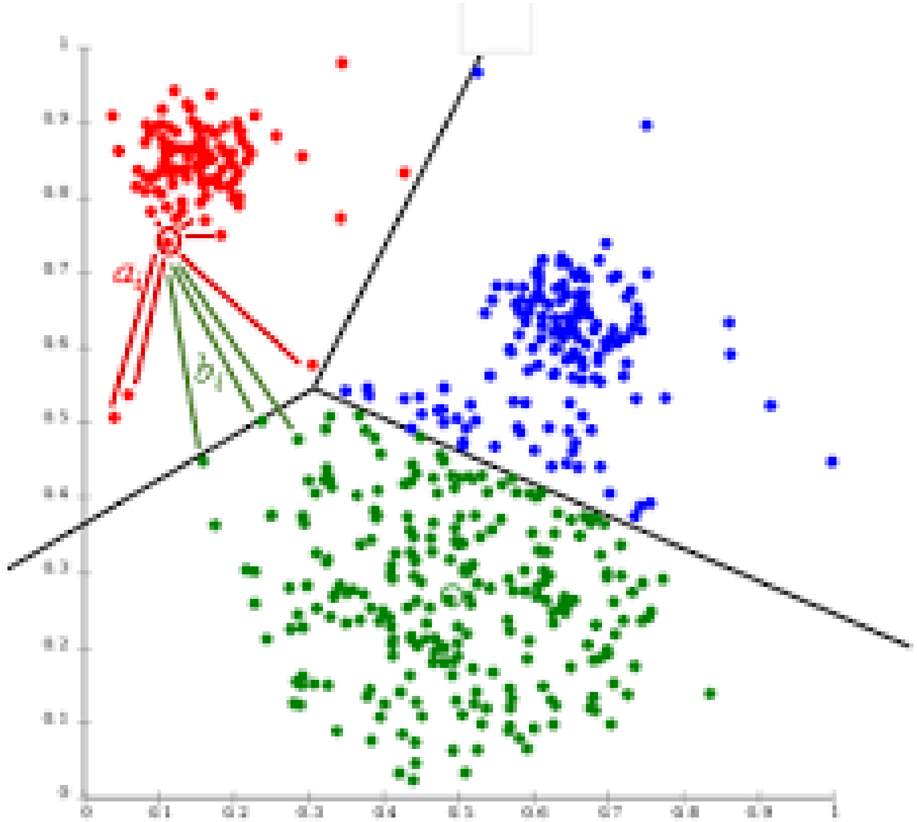
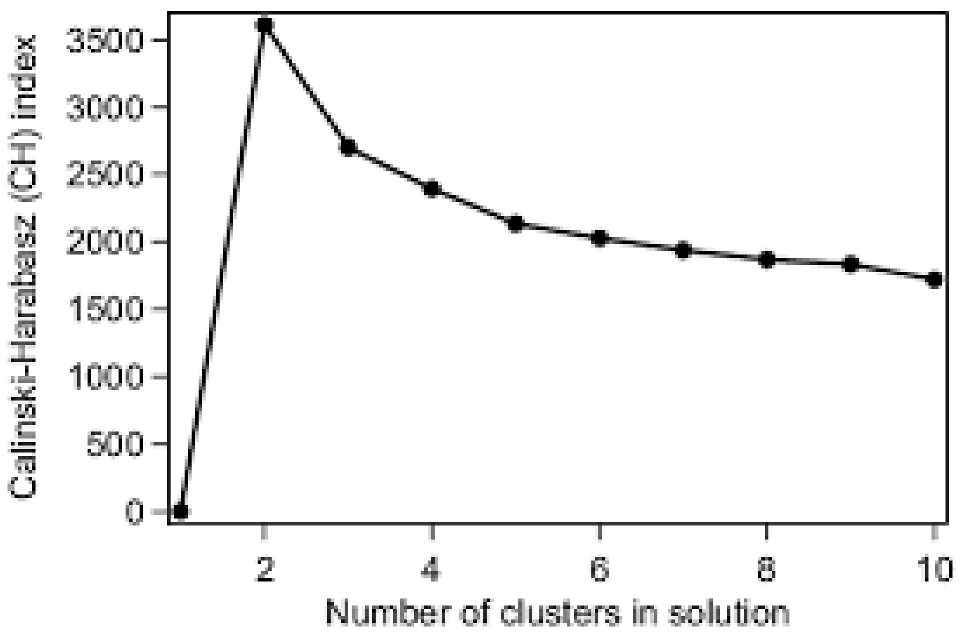
Feature Scaling

04

Feature Selection

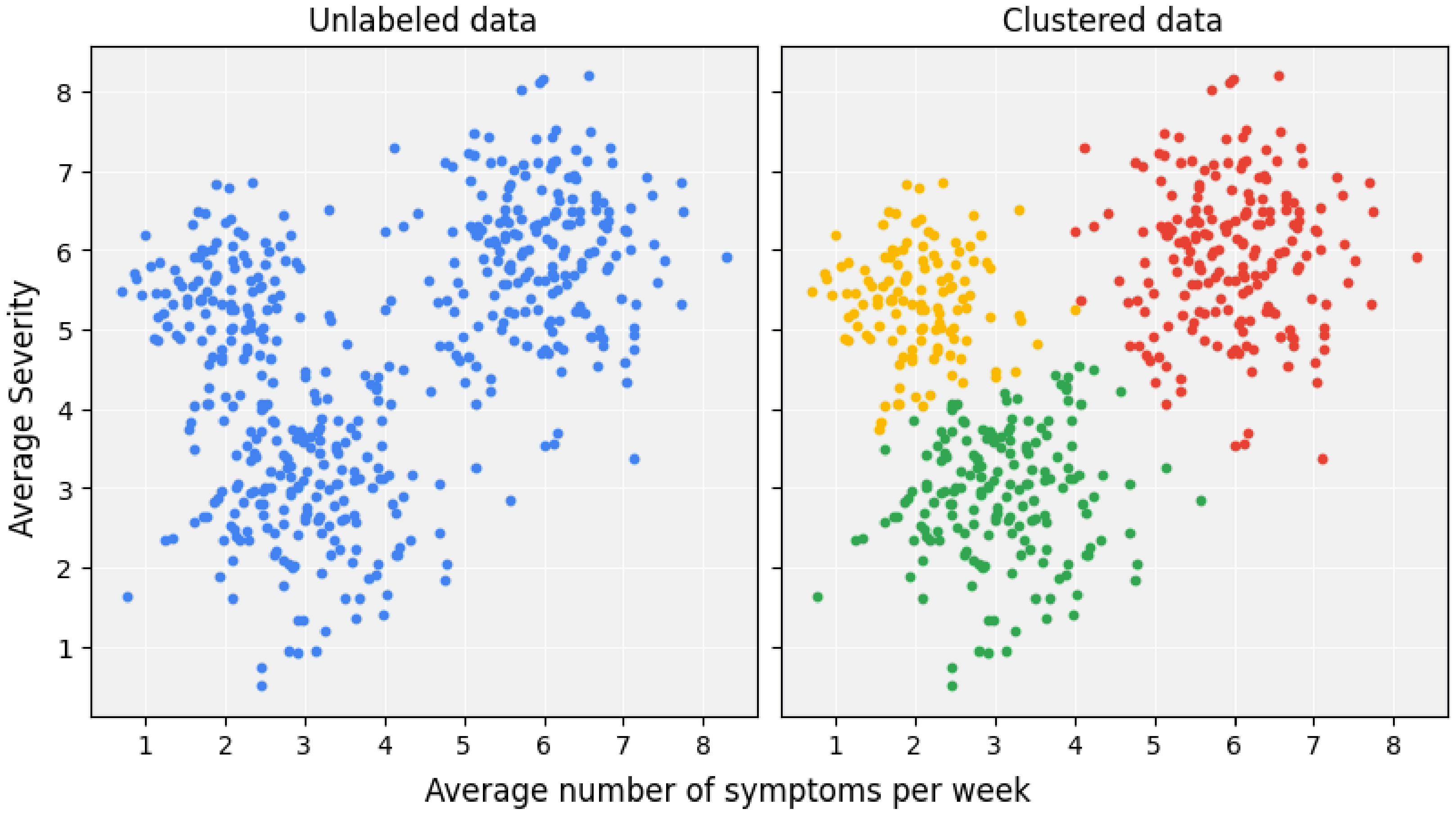
https://www.ibm.com/think/topics/feature-engineering
https://www.geeksforgeeks.org/machine-learning/what-is-feature-engineering/
https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d/

Choosing the Number of Clusters

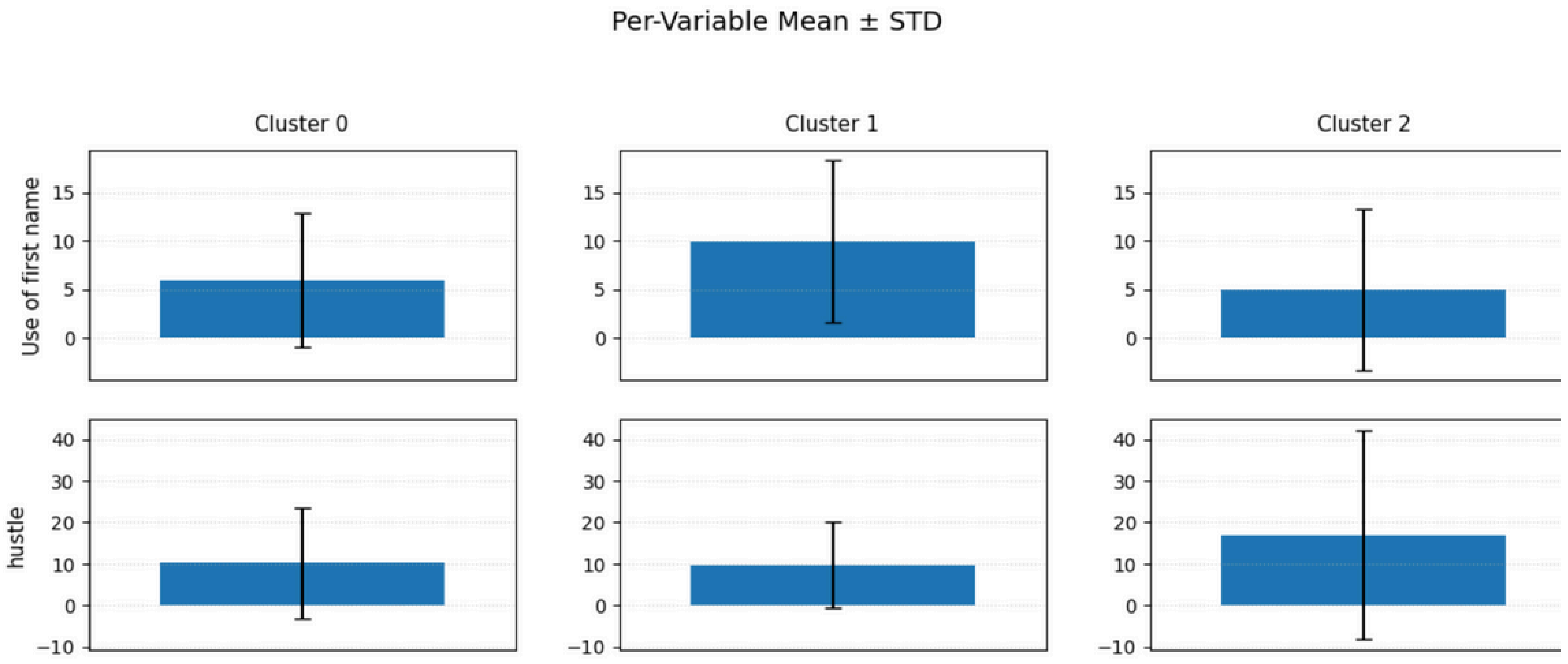
Elbow Method	Silhouette Score	Calinski–Harabasz (CH) Index
Plot the within-cluster variance for different K values	Measures how similar points are to their own cluster vs the nearest other cluster	Ratio of between-cluster variance to within-cluster variance
		

https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5/

Visualize Results



Scatter Plot



Mean and SD Plot

https://tinyurl.com/ml-workshop-1

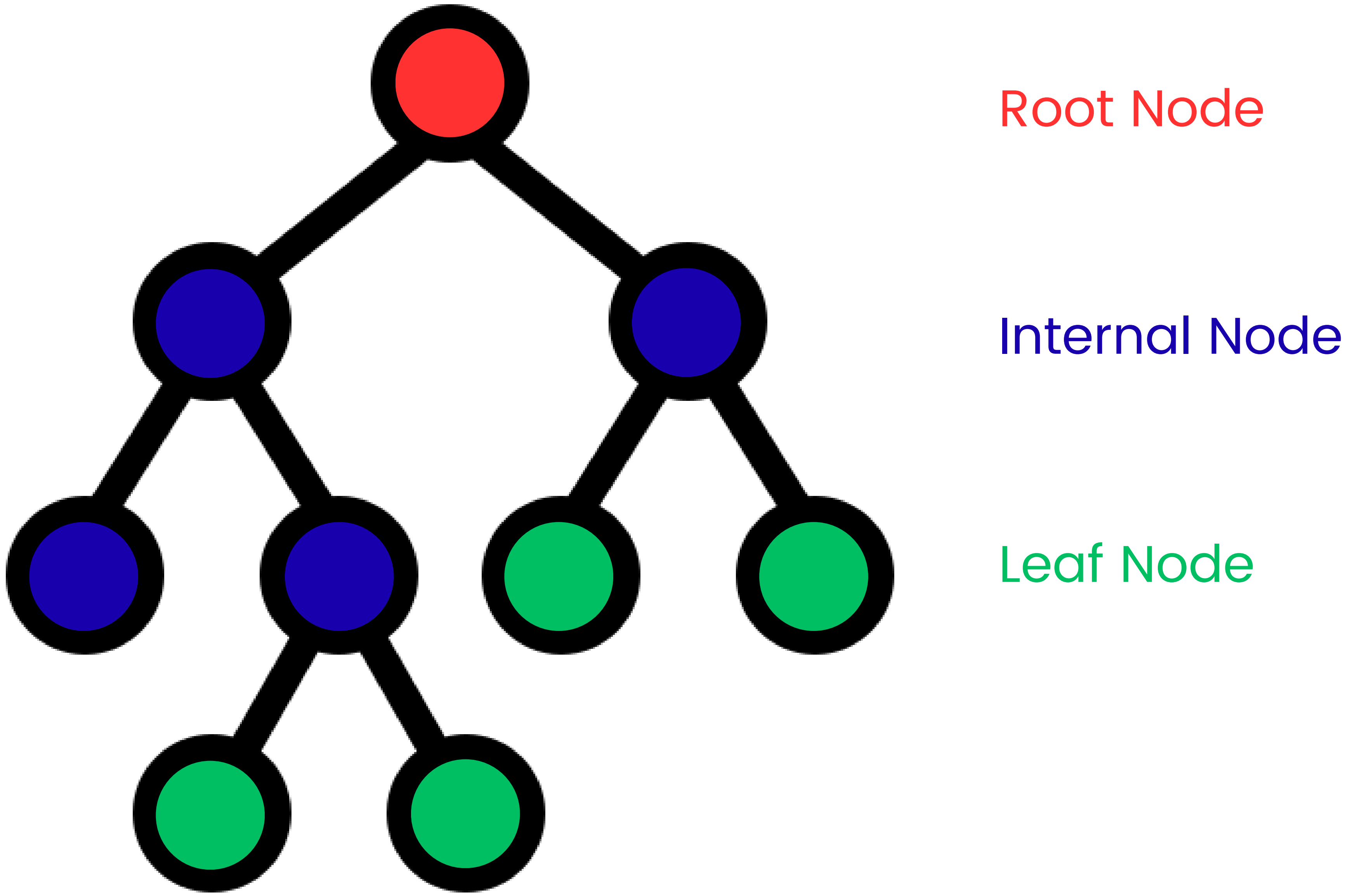
Code Walkthrough

K Means Clustering on Google Colab

Decision Trees

Classification & Regression

Walks through the possible outcomes and maps out decision pathways



Root Node

Internal Node

Leaf Node

Overview

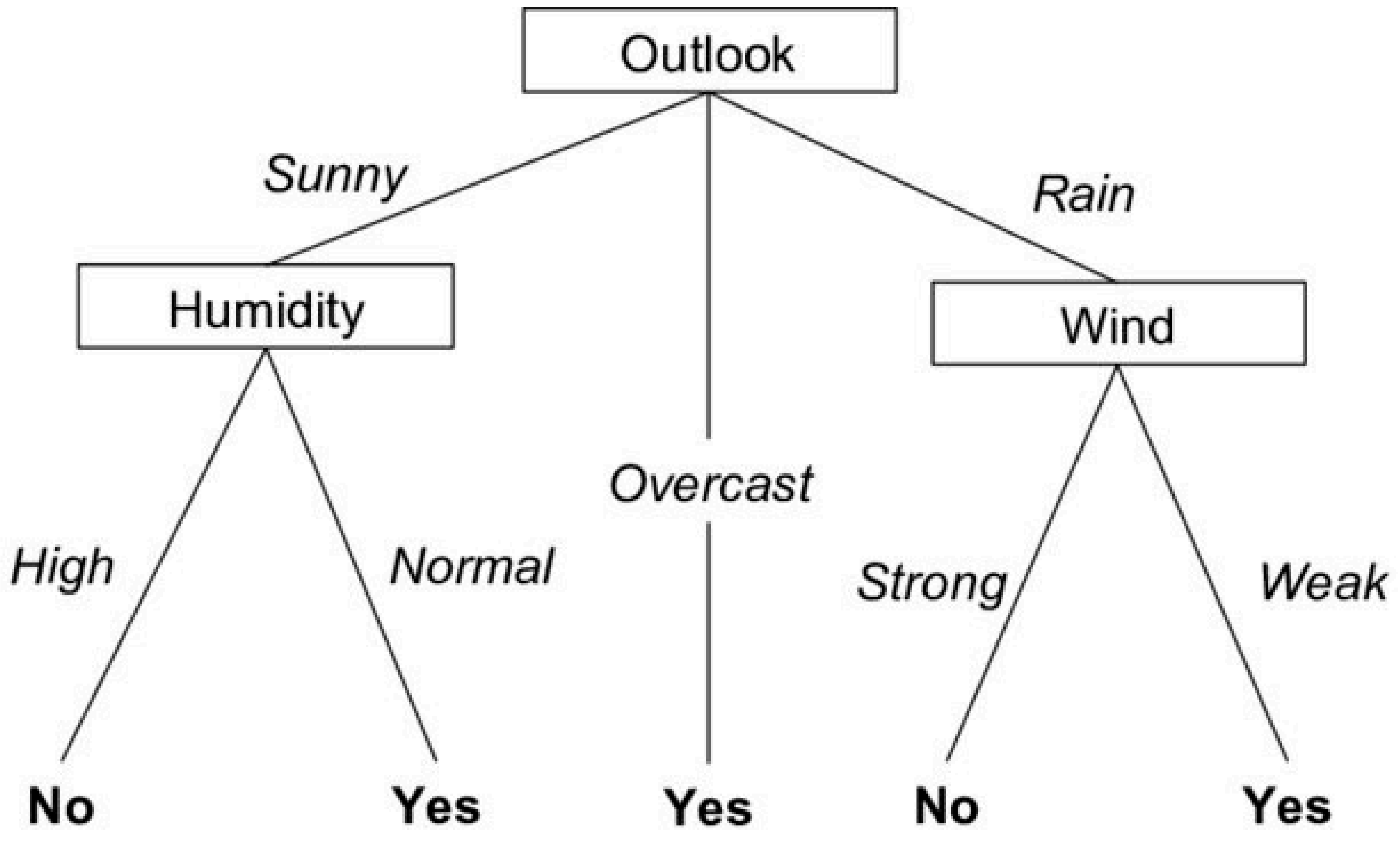
<table>
<tr><th>Items</th><th>Classification</th><th>Regression</th></tr>
<tr><td>Independent Variable</td><td>Numerical Data</td><td>Numerical Data</td></tr>
<tr><td>Dependent Variable</td><td>Categorical Data</td><td>Numerical Data</td></tr>
<tr><td>Advantages</td><td colspan="2">Easy to interpret
No need for feature scaling
Relatively quick to implement</td></tr>
<tr><td>Disadvantages</td><td colspan="2">Prone to overfitting
Unstable
Does not work as well with large dataset</td></tr>
</table>

https://www.ibm.com/think/topics/decision-trees
https://www.ibm.com/think/topics/overfitting

Rule Mining

A set of conditions used to determine decision making paths

“IF-THEN” framework



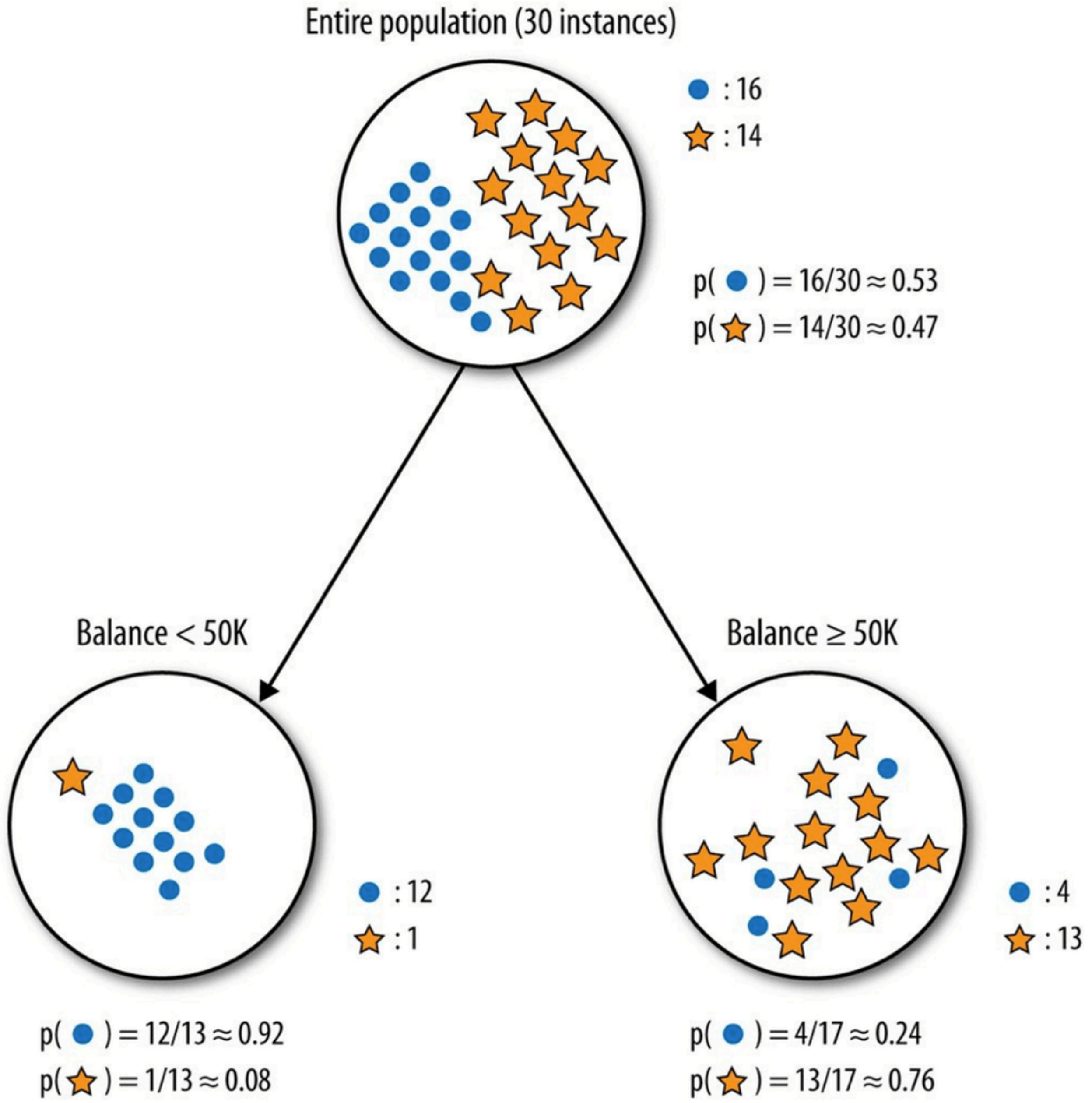
How Decision Trees Choose Splits?

Through a concept called....

“Impurity”

Gini Impurity - Criterion measure on how “impure” child nodes are in relation to parent node

Entropy - Measure on the amount of “disorderliness” or “uncertainty” in the data



https://www.geeksforgeeks.org/machine-learning/gini-impurity-and-entropy-in-decision-tree-ml/

Code Walkthrough

Decision Trees and Rule Mining

https://tinyurl.com/ml-workshop-2

Thank you for attending!

Feel free to send in any questions to:

Chaitanya [CHAITANY002@e.ntu.edu.sg]

Julian [NIE22.TQJ@e.ntu.edu.sg]