Intelligence System:

Unsupervised Learning in Machine Learning

សាកលវិទ្យាល័យនរតុន
NORTON UNIVERSITY

Intelligence System Development

2024 – 2025

Y4E1 – DCS – NU

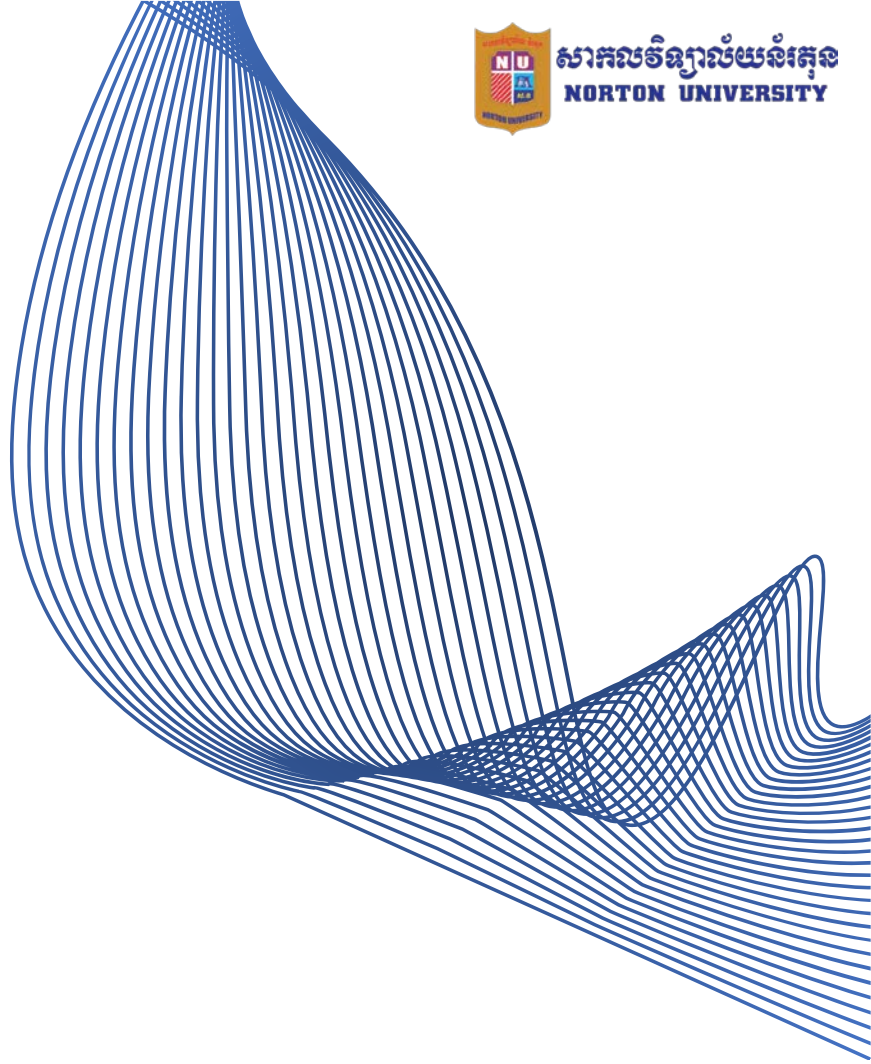
By: SEK SOCHEAT

Advisor to DCS and Lecturer

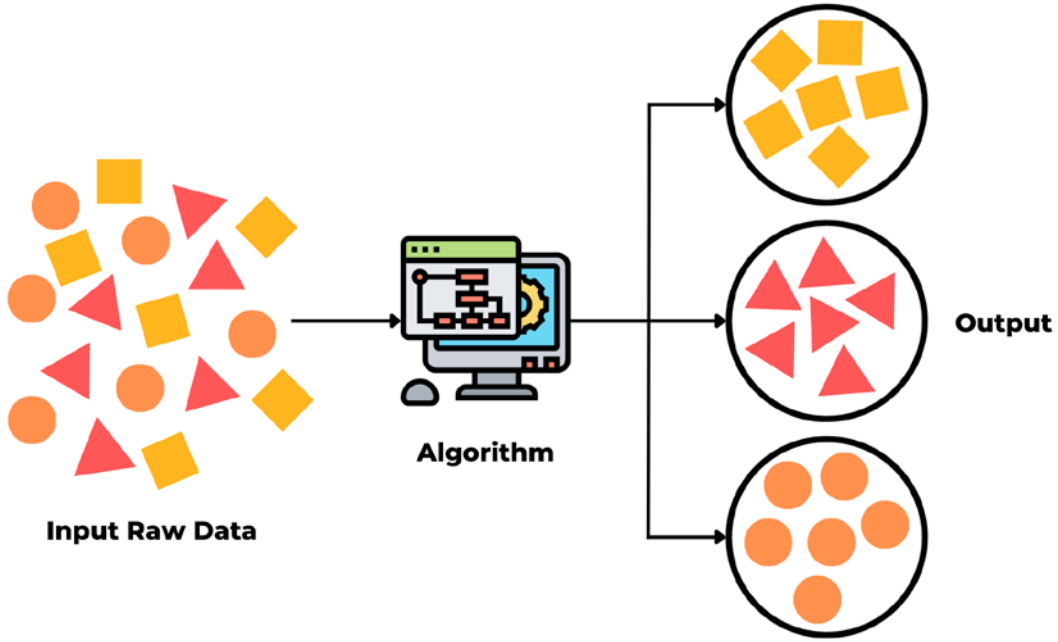
Mobile: 017 879 967

Email: socheat.sek@gmail.com

Table of Contents



Introduction to Unsupervised Learning



https://eastgate-software.com/wp-content/uploads/2023/10/Unsupervised-Learning-Clustering.png

Unsupervised learning is a type of machine learning where the model learns patterns and structures in data without labeled outputs. Unlike supervised learning, it works with datasets that lack target variables, aiming to identify hidden relationships, groupings, or structures in the input data.

Key Characteristics of Supervised Learning

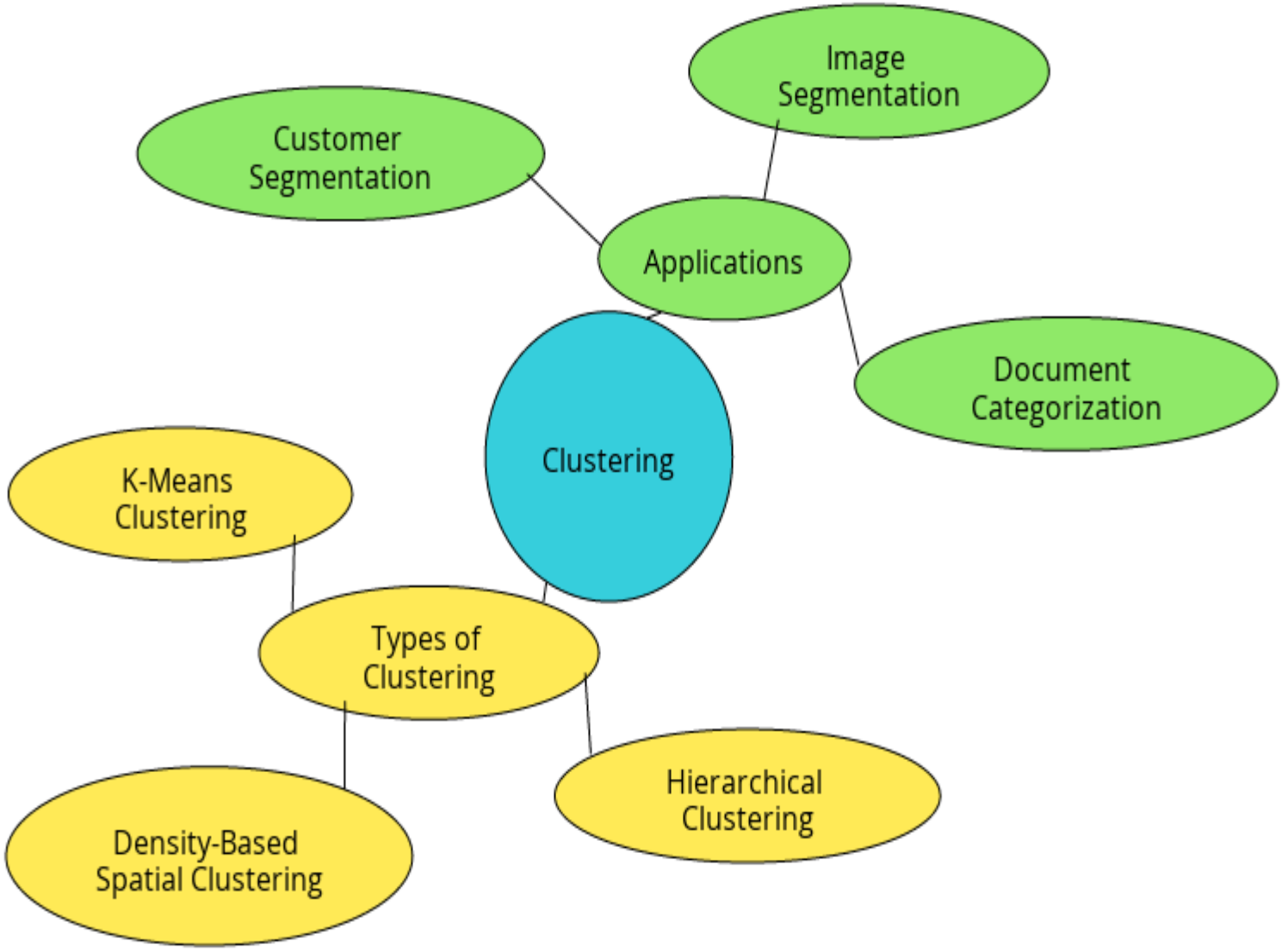
Key Concepts of Supervised Learning:

- **No Labels:** The data consists only of input features with no corresponding output labels.
- **Exploratory Analysis:** Focuses on discovering patterns or structures in the data.
- **Applications:** Often used in clustering, dimensionality reduction, and anomaly detection.

Types of Unsupervised Learning

What is Clustering?

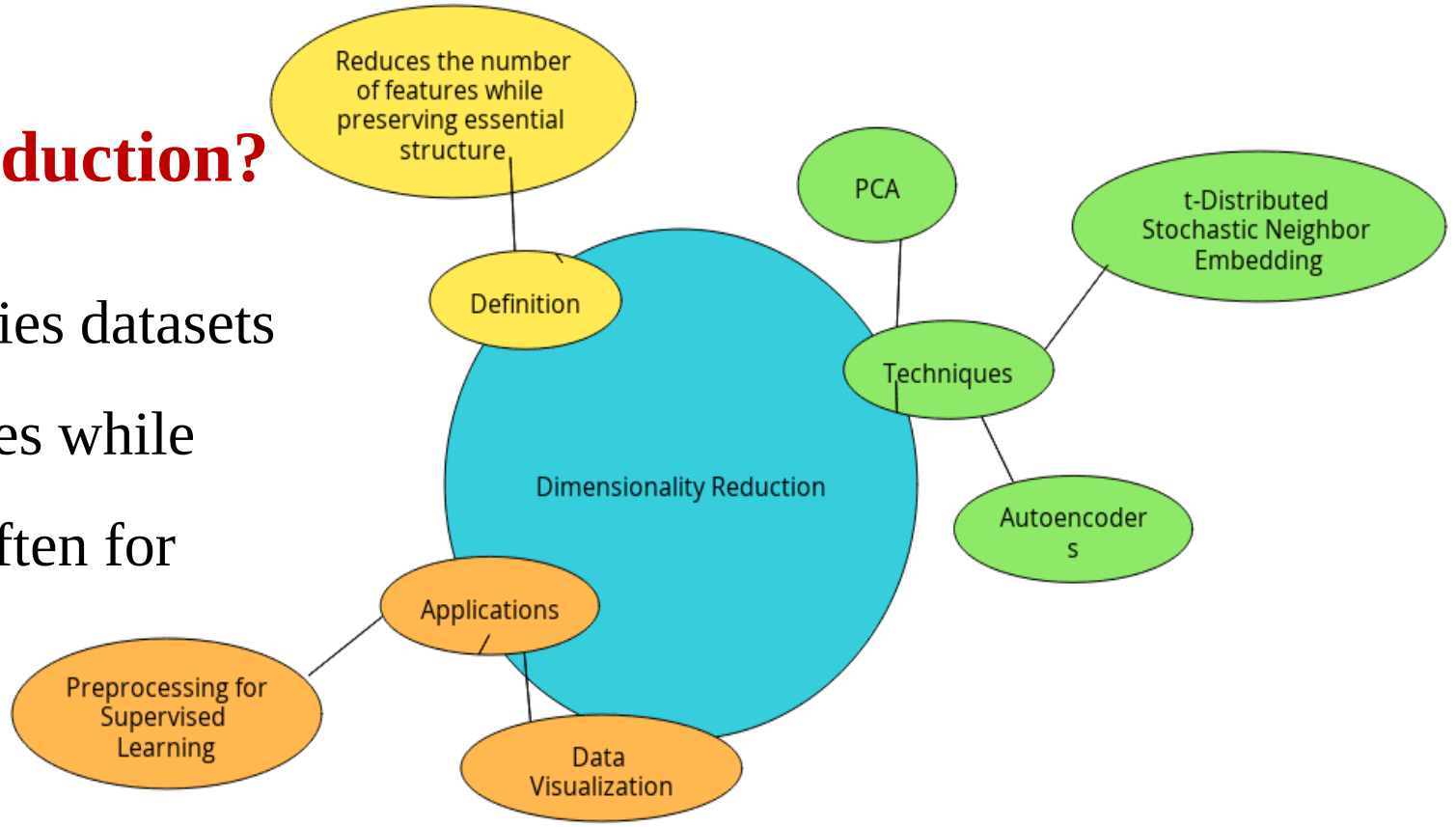
Clustering is an unsupervised learning technique that groups similar data points into clusters based on shared characteristics, helping identify patterns or structures in data.



Types of Unsupervised Learning

What is Dimensionality Reduction?

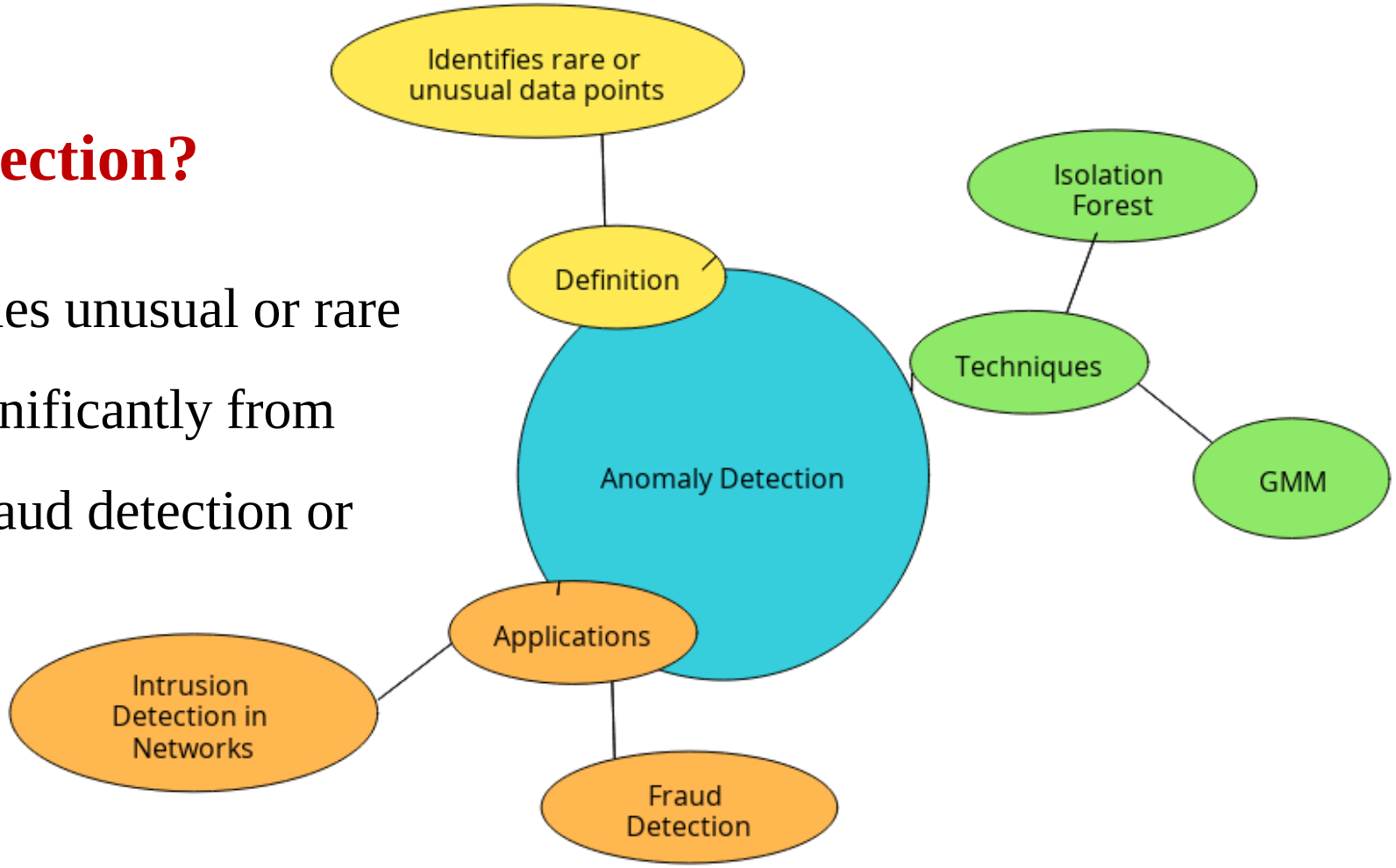
Dimensionality reduction simplifies datasets by reducing the number of features while retaining essential information, often for visualization or improved model performance.



Types of Unsupervised Learning

What is Anomaly Detection?

Anomaly detection identifies unusual or rare data points that deviate significantly from the norm, often used for fraud detection or error identification.



Popular Algorithms in Unsupervised Learning

Unsupervised learning uses algorithms like ***K-Means Clustering*** for grouping data, ***Hierarchical Clustering*** for tree-like group structures, ***DBSCAN*** for density-based clustering, ***PCA*** for reducing dimensions, and ***Autoencoders*** for efficient data compression and reconstruction.

1. K-Means Clustering: Divides data into k clusters by minimizing distances within clusters.

2. Hierarchical Clustering: Builds a tree-like structure (dendrogram) to show data groupings.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Groups data points by density, marking outliers as noise.

4. PCA (Principal Component Analysis): Reduces data dimensions for analysis and visualization.

5. Autoencoders: Neural networks designed to compress and reconstruct data efficiently.

Steps in Unsupervised Learning

Unsupervised learning involves preparing data, selecting a model, training it to find patterns, evaluating performance, and interpreting results for insights.

1. Data Preparation: Clean and preprocess the dataset (handle missing values, normalize features) and select relevant features for analysis.

2. Model Selection: Choose an appropriate algorithm (e.g., clustering or dimensionality reduction).

3. Training: Feed the data into the model to learn patterns or clusters.

4. Evaluation: Use metrics like silhouette score, Davies-Bouldin index, or visualizations to assess performance.

5. Interpretation: Analyze the results to gain insights or inform decisions.

Applications of Unsupervised Learning

Unsupervised learning is used in customer segmentation, recommendation systems, data compression, bioinformatics, and image processing to uncover patterns and insights.

1. Market Segmentation: Grouping customers based on purchasing behavior.

2. Recommendation Systems: Identifying patterns to suggest products or content.

3. Data Compression: Reducing dataset dimensions for storage efficiency.

4. Bioinformatics: Clustering genes or proteins with similar functions.

5. Image Processing: Identifying features in unlabeled images.

Challenges in Unsupervised Learning

Unsupervised learning faces challenges like lack of labeled data, result interpretability, scalability with large datasets, and sensitivity to algorithm parameters.

1. Lack of Ground Truth: Difficult to evaluate performance without labeled data.

2. Interpretability: Hard to interpret and explain model results.

3. Scalability: Some algorithms struggle with large datasets.

4. Parameter Sensitivity: Algorithms like K-Means require setting parameters (e.g., number of clusters) beforehand.

Example

The goal of this example is to automate the detection and identification of various fruits and vegetables in an image. In real-world scenarios, such tasks are essential for applications in:

- **Food Industry:** Sorting fruits and vegetables based on type and color.
- **Retail:** Automating checkout counters in grocery stores.
- **Agriculture:** Analyzing and classifying harvested crops.
- **Machine Learning:** Understanding how unsupervised learning techniques like K-Means clustering can be applied for image segmentation.

The challenge lies in:

- Identifying distinct fruit/vegetable types based on their dominant colors.
- Accurately mapping clusters to corresponding fruit or vegetable names.
- Processing diverse datasets where lighting, overlapping objects, or similar colors can make detection difficult.

```python
import cv2
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Load the image
image_path = "fruits_01.jpg"  # Update with the path to your uploaded image
image = cv2.imread(image_path)

if image is None:
    print("Error: Unable to read the image.")
    exit()

# Convert the image to RGB
image_rgb = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)

# Reshape the image into a 2D array of pixels and 3 color values (R, G, B)
pixels = image_rgb.reshape((-1, 3))

# Convert pixels to float for clustering
pixels = np.float32(pixels)

# Perform K-Means Clustering
k = 6  # Number of clusters
kmeans = KMeans(n_clusters=k, random_state=0, n_init=10)
labels = kmeans.fit_predict(pixels)
centers = kmeans.cluster_centers_

# Convert cluster centers to integers
centers = np.uint8(centers)

# Print cluster centers for debugging
print("Cluster Centers (RGB):", centers)

# Define known RGB colors for fruits/vegetables and their names
fruit_vegetable_colors = {
    (255, 165, 0): "Orange",
    (200, 50, 50): "Apple",
    (255, 255, 0): "Banana",
    (0, 255, 0): "Lime",
    (139, 69, 19): "Coconut",
    (255, 99, 71): "Tomato"
}
```

Example of Unsupervised Learning

Using K-Means Clustering for fruits detection

```
44
45 # Match cluster centers to known fruit/vegetable colors
46 detected_items = []
47 for center in centers:
48     color_tuple = tuple(center)
49     closest_match = None
50     min_distance = float('inf')
51     for known_color, name in fruit_vegetable_colors.items():
52         distance = np.linalg.norm(np.array(color_tuple) - np.array(known_color))
53         if distance < min_distance:
54             min_distance = distance
55             closest_match = name
56     detected_items.append(closest_match if closest_match else "Unknown")
57
58 # Display the original and segmented images
59 plt.figure(figsize=(12, 6))
60
61 plt.subplot(1, 2, 1)
62 plt.title("Original Image")
63 plt.imshow(image_rgb)
64 plt.axis("off")
65
66 plt.subplot(1, 2, 2)
67 plt.title(f"Segmented Image with {k} Clusters")
68 plt.imshow(centers[labels.flatten()].reshape(image_rgb.shape))
69 plt.axis("off")
70
71 plt.tight_layout()
72 plt.show()
73
74 # Print detected fruits and vegetables
75 print("Detected Fruits and Vegetables:")
76 for item in set(detected_items):  # Use `set` to avoid duplicates
77     print(item)
78
```

Example of Unsupervised Learning

Using K-Means Clustering for fruits detection

How Its Work:

1. Image Loading: Loads the uploaded image for clustering.

2. K-Means Clustering: Clusters the image into k-clusters based on dominant colors.

3. Color Mapping: Maps cluster centers to predefined fruit colors to identify the fruits.

4. Output:

- Displays the segmented image with clustered fruits.
- Prints the detected fruit names.

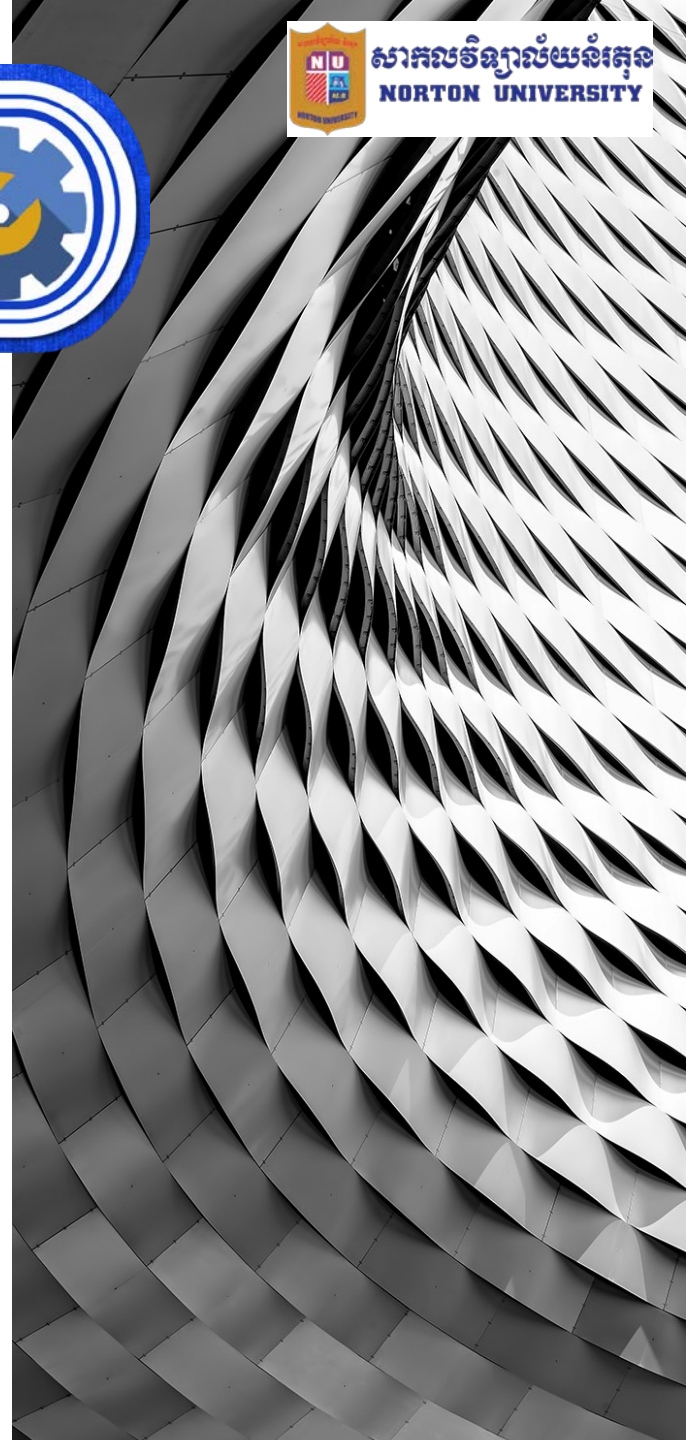
Homework:

Answer Questions below:

1. What is the primary goal of unsupervised learning?

2. What are some common techniques used in unsupervised learning?

3. What are the main challenges of unsupervised learning?



សាកលវិទ្យាល័យនរតុន
NORTON UNIVERSITY

Thank you