

2025

4Geeks Academy: data science cohort 12

# DAY 25: UNSUPERVISED LEARNING

# TODO

## UNSUPERVISED LEARNING

Types & applications: clustering, dimensionality reduction

## K-NEAREST NEIGHBORS PROJECT

Submit K-nearest neighbors Project Tutorial (K-nearest neighbors module), if you haven't done so already

## CLUSTERING PROJECT

Work on K-means Project Tutorial (Unsupervised Learning Module), plan to finish before class Friday.

# TOPICS

01 CLUSTERING: K-MEANS

02 DIMENSIONALITY REDUCTION: PCA

# CLUSTERING: K-MEANS

**WHAT** Unsupervised learning technique to cluster (group) data by similarity

**WHY** No need for labels - data is grouped based on structure of feature space  
Useful for identifying subpopulations in data

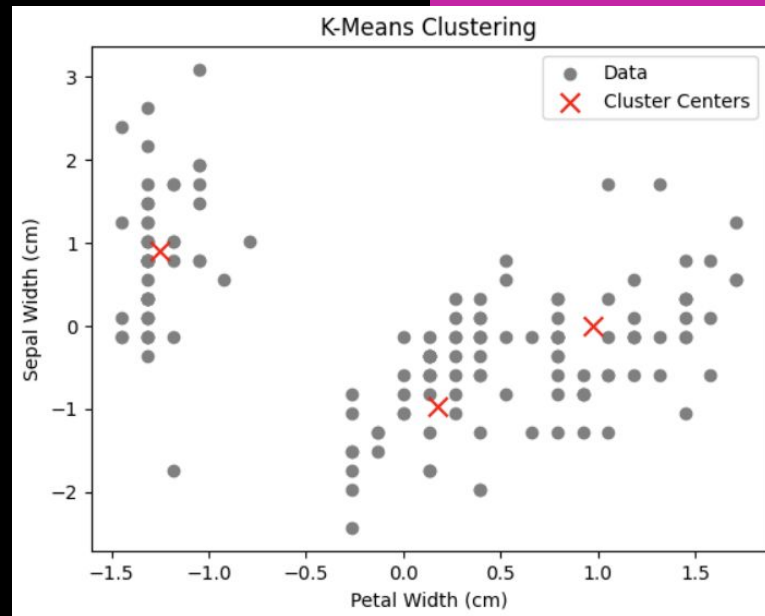
## HOW

1. Initialize k cluster center points in feature space
2. Assign each observation to nearest cluster
3. Update cluster center point location
4. Repeat until cluster centers stop moving

**Scikit-learn** implementation: `KMeans()`

### Hyperparameters:

1. `n_clusters`: number of clusters to find
2. `init`: how to initialize cluster centers
3. `n_init`: how many different initializations to test,
4. `max_iter`: maximum cycles to run
5. `tol`: amount of center movement considered 'stable'



# DIMENSIONALITY REDUCTION: PCA

**WHAT** Unsupervised feature engineering technique to reduce the size of feature space

**WHY** Reduce computational complexity  
Reduce noise and mitigate overfitting  
Can allow visualization/interpretation of high dimensional data sets

## HOW

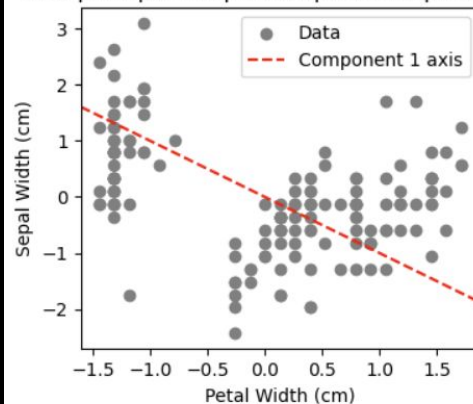
1. Data is transformed into a new set of features (components)
2. First component aligns with the direction of greatest variation in feature space
3. Second component aligns with the direction of greatest variation remaining, etc
4. Select how many components you want for analysis/models

**Scikit-learn** implementation: `PCA()`

**Hyperparameters:**

1. `n_components`: number of components to keep

First principal component: petal & sepal widths



Distribution of first principal component

