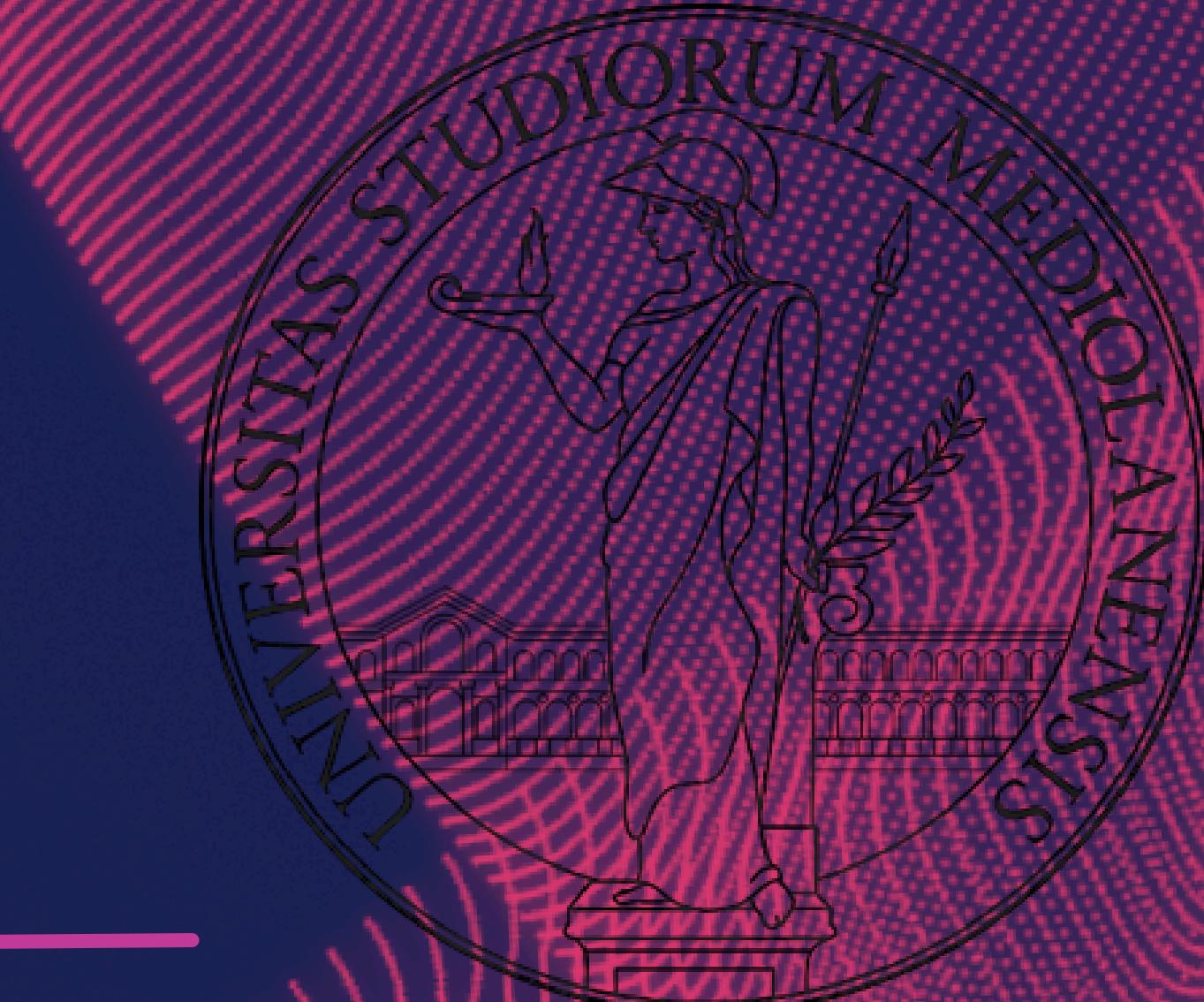
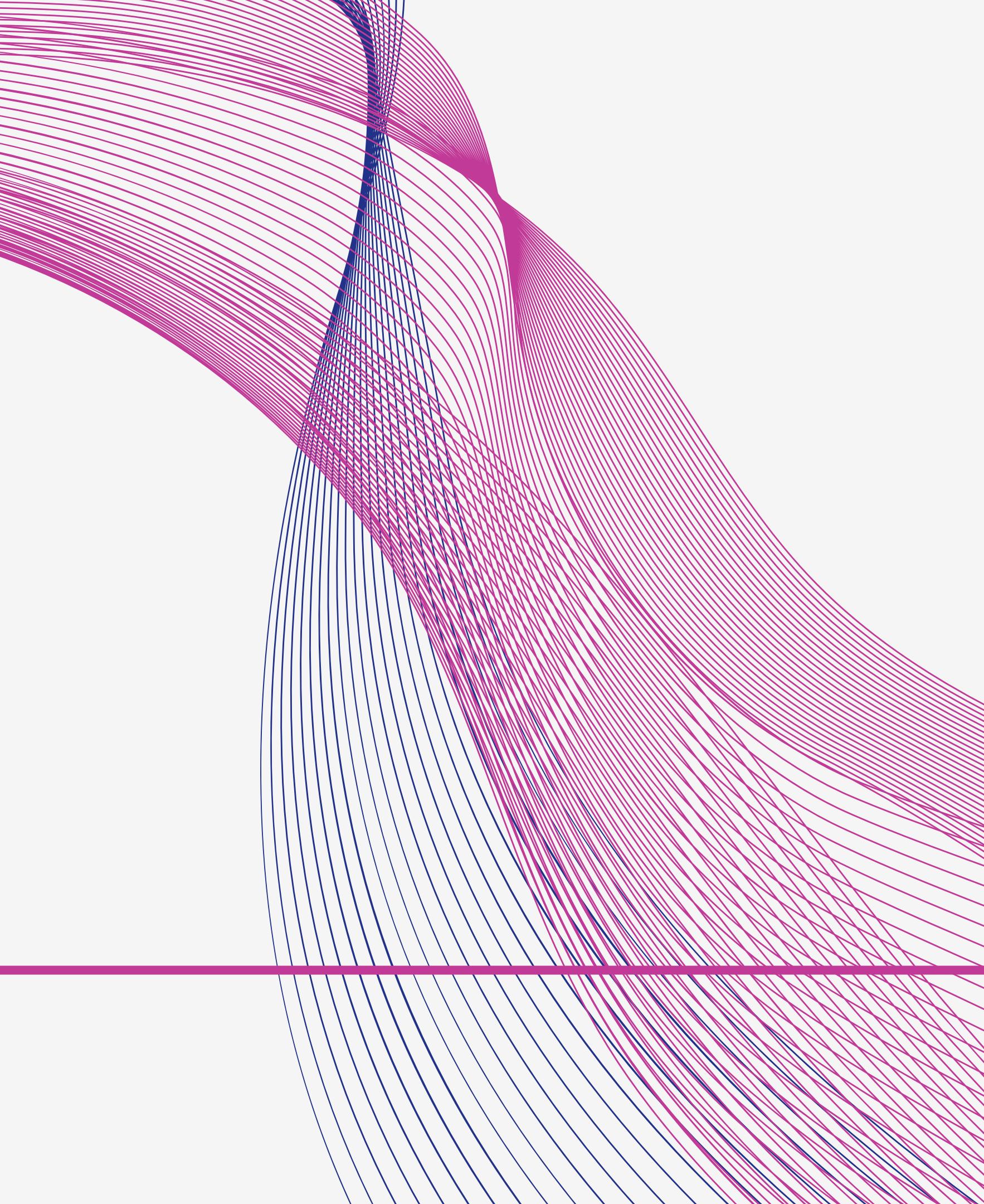


Coding for Data Science, Python Module

Unsupervised Learning P6: I mammiferi depongono uova?

Fatemeh Jandari
January 2025





CONTENT

- Introduction
 - Data
 - Methodology
 - Results
-

Introduction

- Goal: Cluster animals based on their features to identify patterns and relationships.
- Approach: Applying K-Means clustering both with and without dimensionality reduction techniques.
- Evaluation: Clustering effectiveness is measured using accuracy.

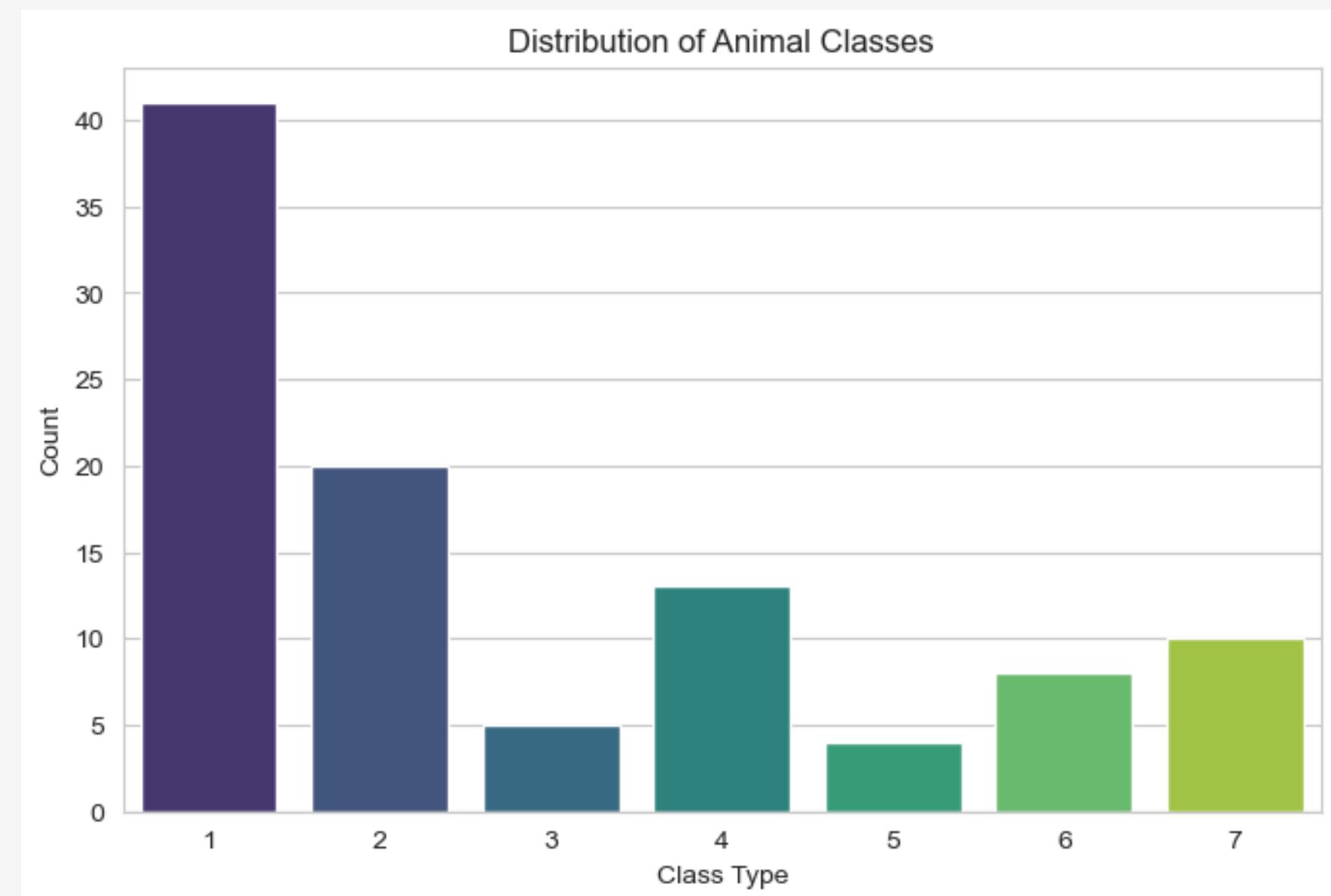
THE DATA

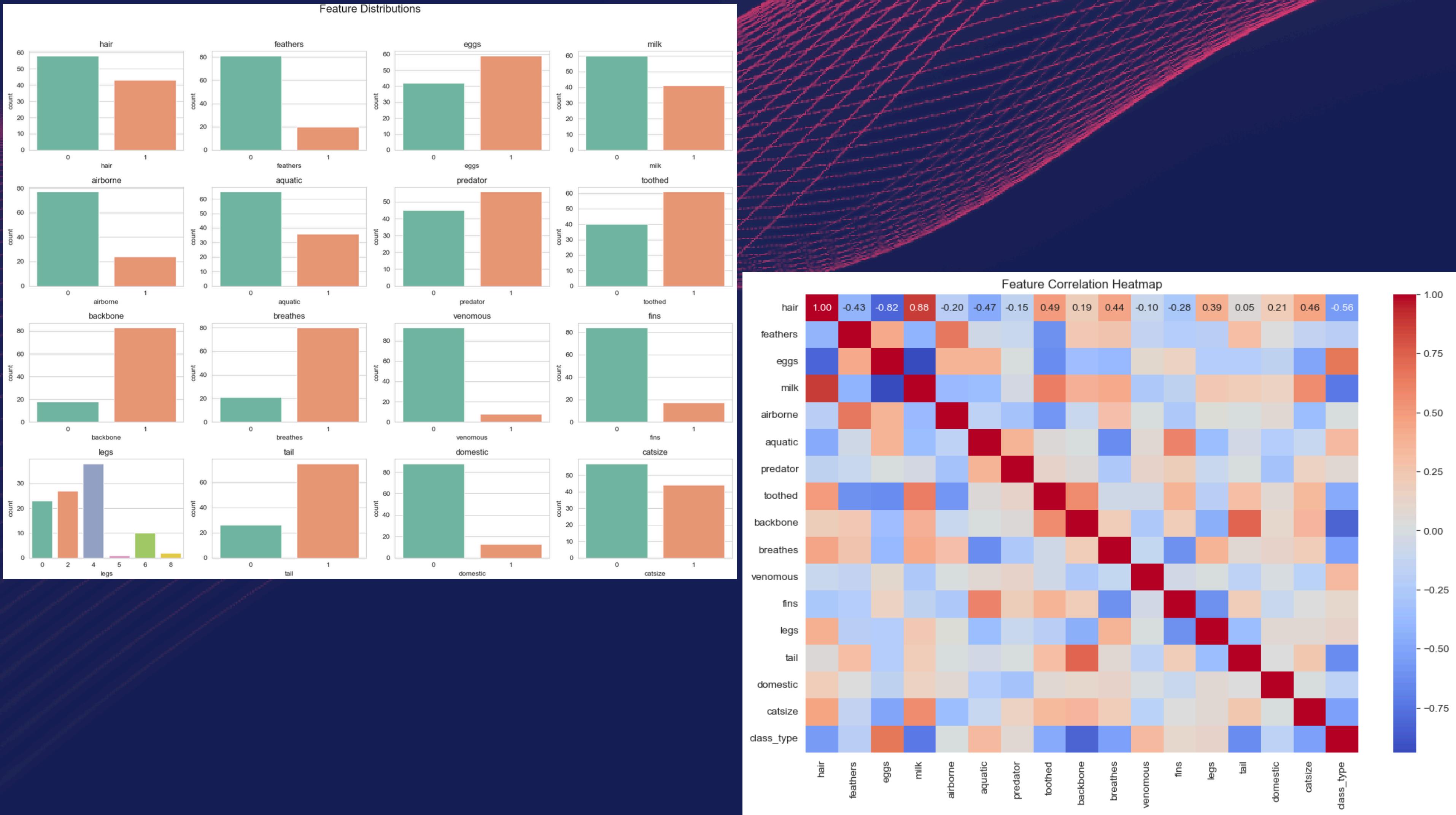


animal_name	hair	feathers	eggs	milk	airborne	aquatic	predator	toothed	backbone	breathes	venomous	fins	legs	tail	domestic	catsize	class_type
aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
antelope	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
bear	1	0	0	1	0	0	1	1	1	1	0	0	4	0	0	1	1
boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1	0	1	1
buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1	0	1	1
calf	1	0	0	1	0	0	0	1	1	1	0	0	4	1	1	1	1
carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	4
catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
cavy	1	0	0	1	0	0	0	1	1	1	0	0	4	0	1	0	1

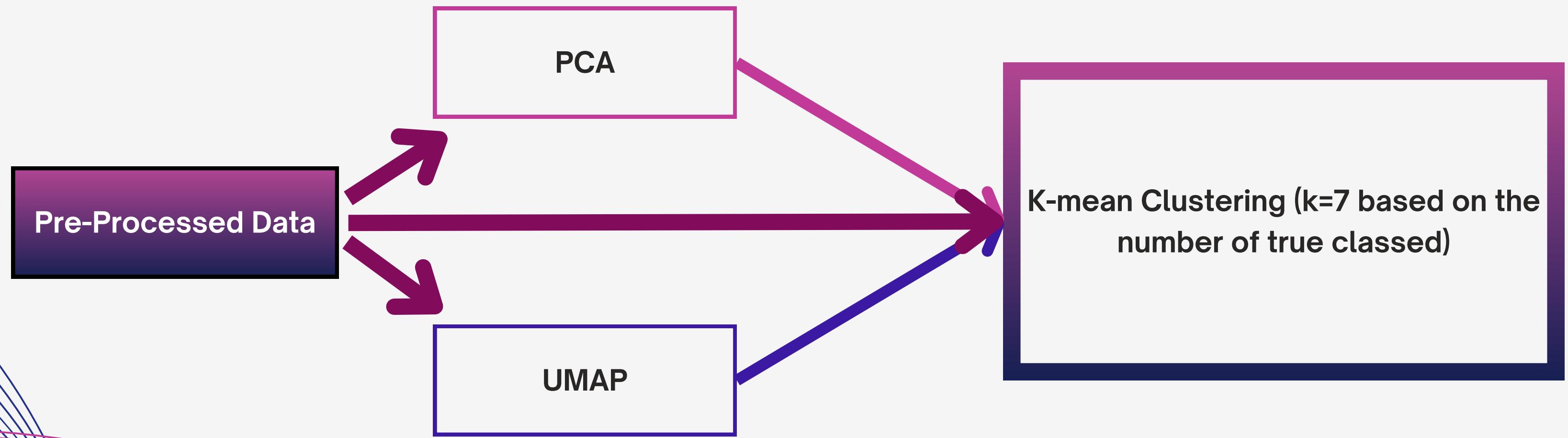
Exploratory Data Analysis (EDA)

- Inspecting its structure and missing values.
- Visualizing the class and features distribution and correlations



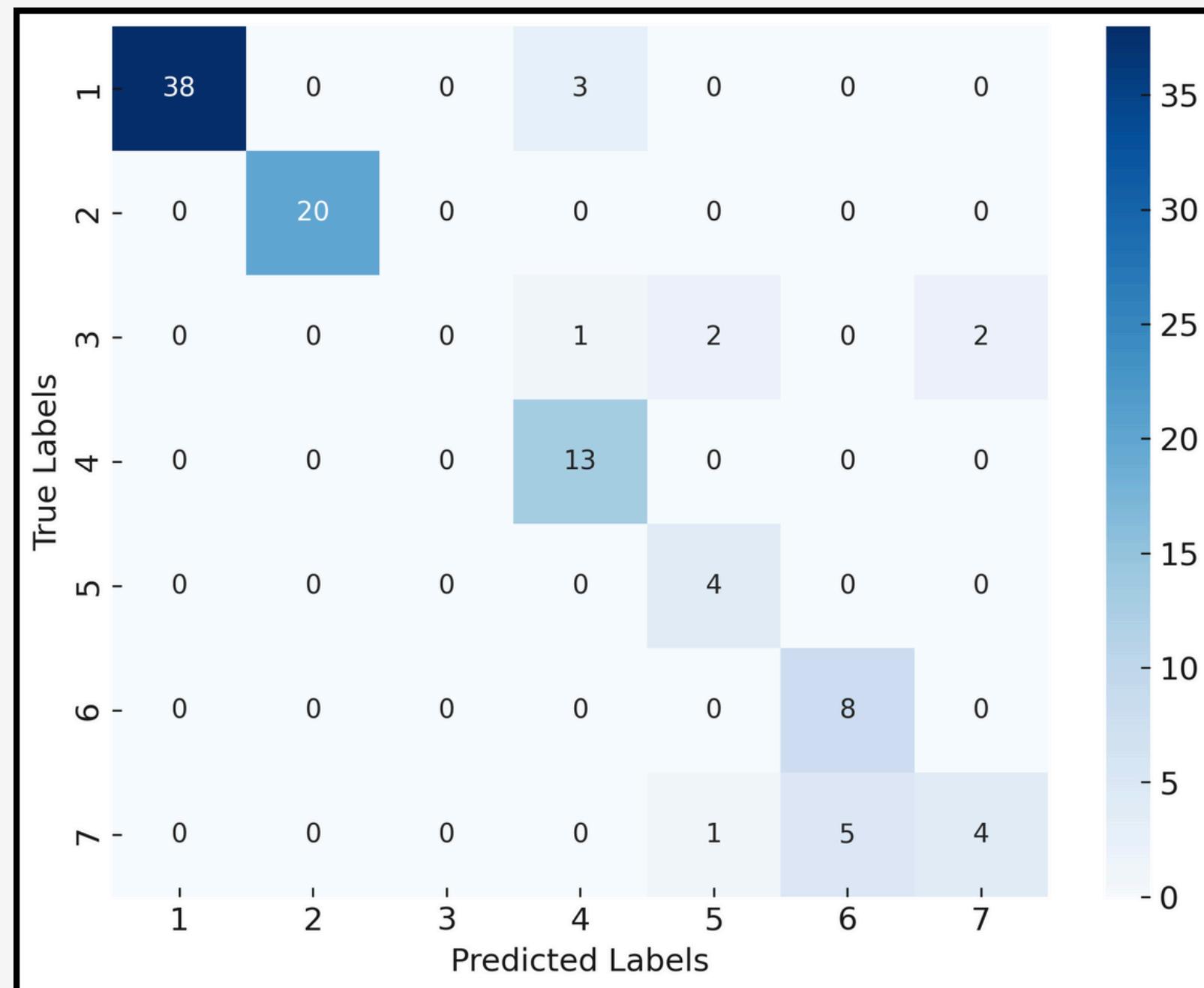


Clustering Methodology



Model 1: K-means Clustering without Dimensionality Reduction

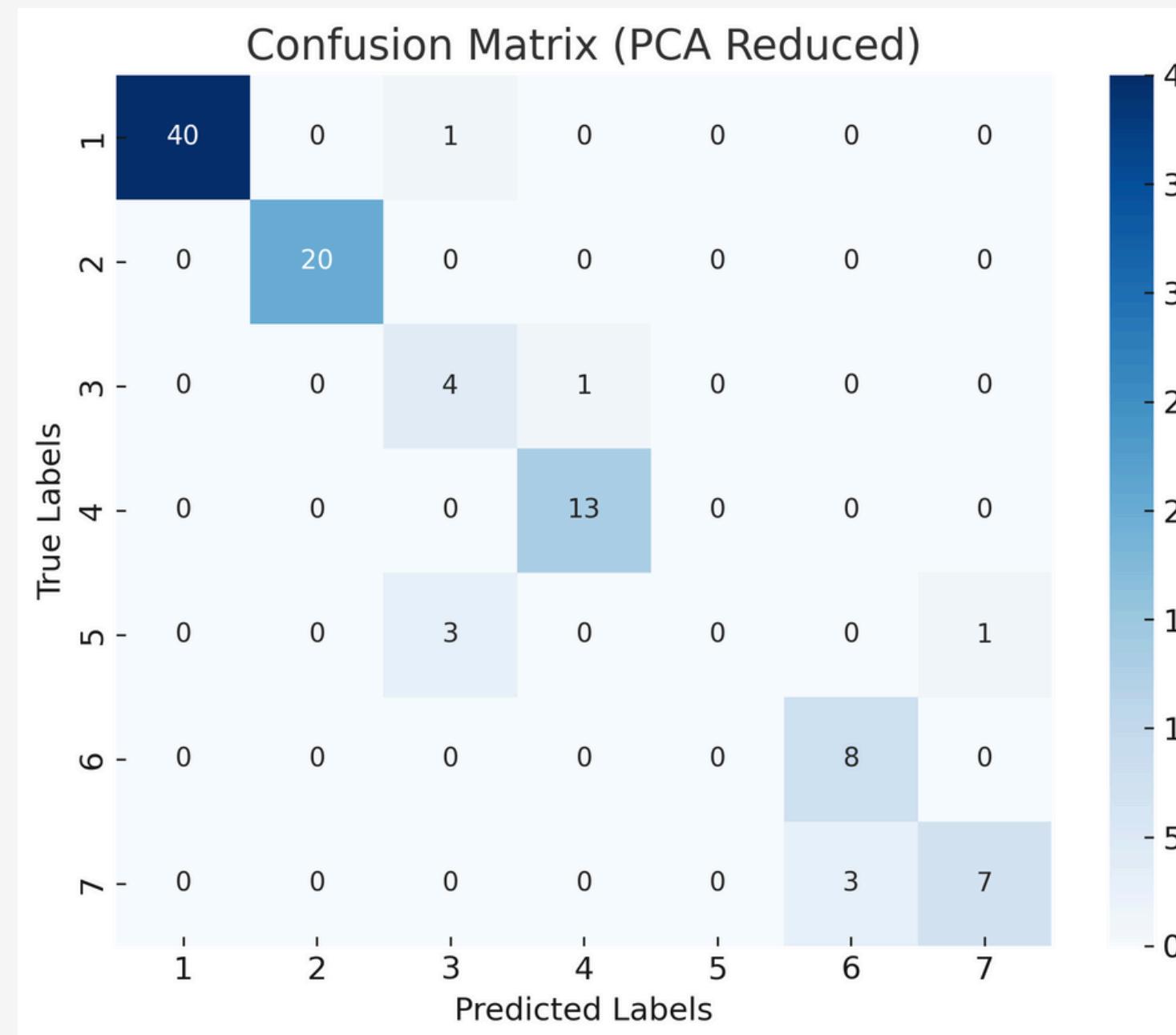
- The K-Means clustering algorithm is applied directly to the dataset.
- Cluster assignments are mapped to true labels using the most frequent class in each cluster.
- Performance is evaluated using a confusion matrix.



Accuracy
0.86

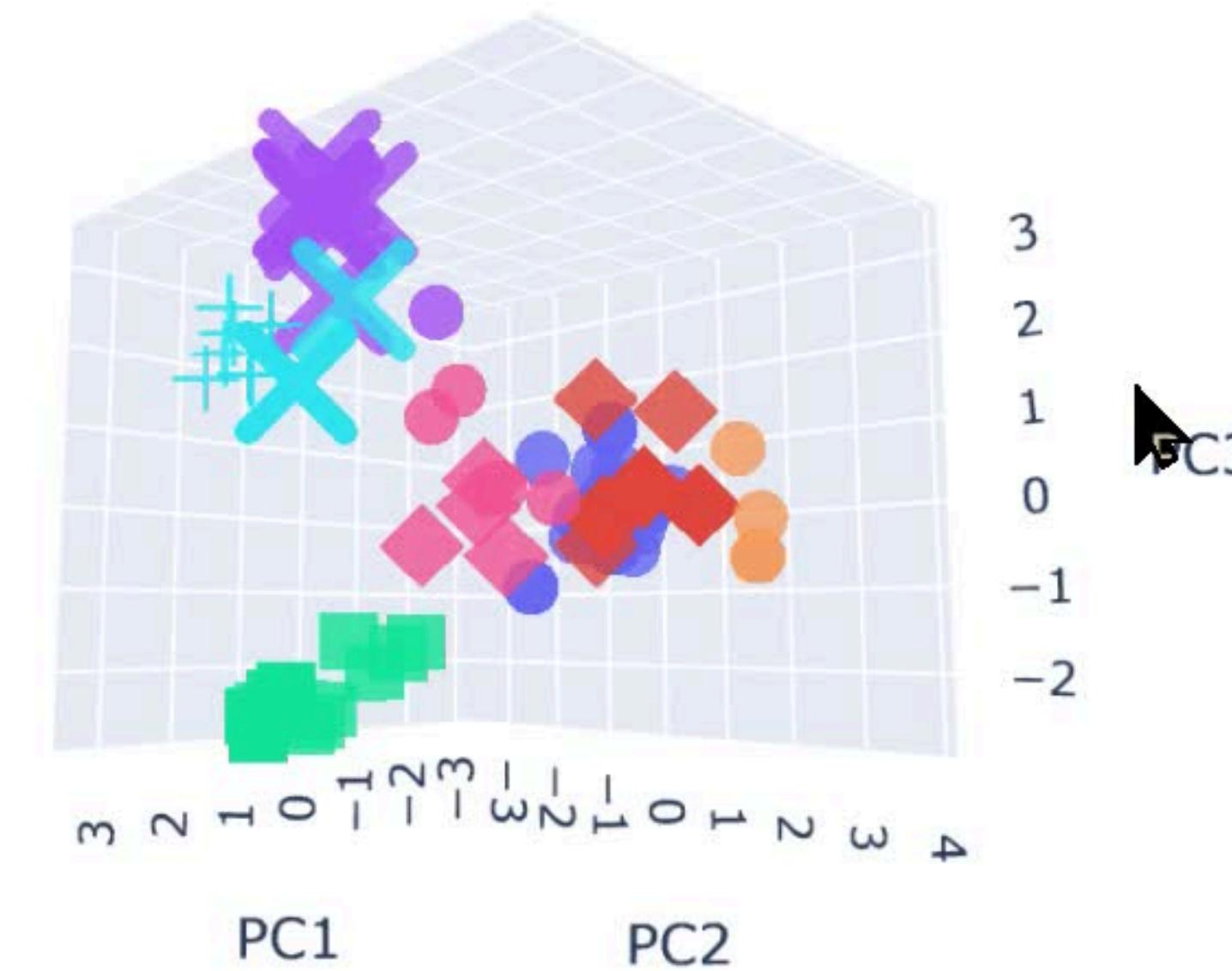
Model 2: K-means Clustering with PCA (Principal Component Analysis)

- PCA is applied to reduce the data to 3 dimensions for visualization.
- K-Means clustering is performed on the reduced dataset.
- Clusters are mapped to actual classes, and performance is evaluated.
- Results are visualized in an interactive 3D scatter plot using Plotly



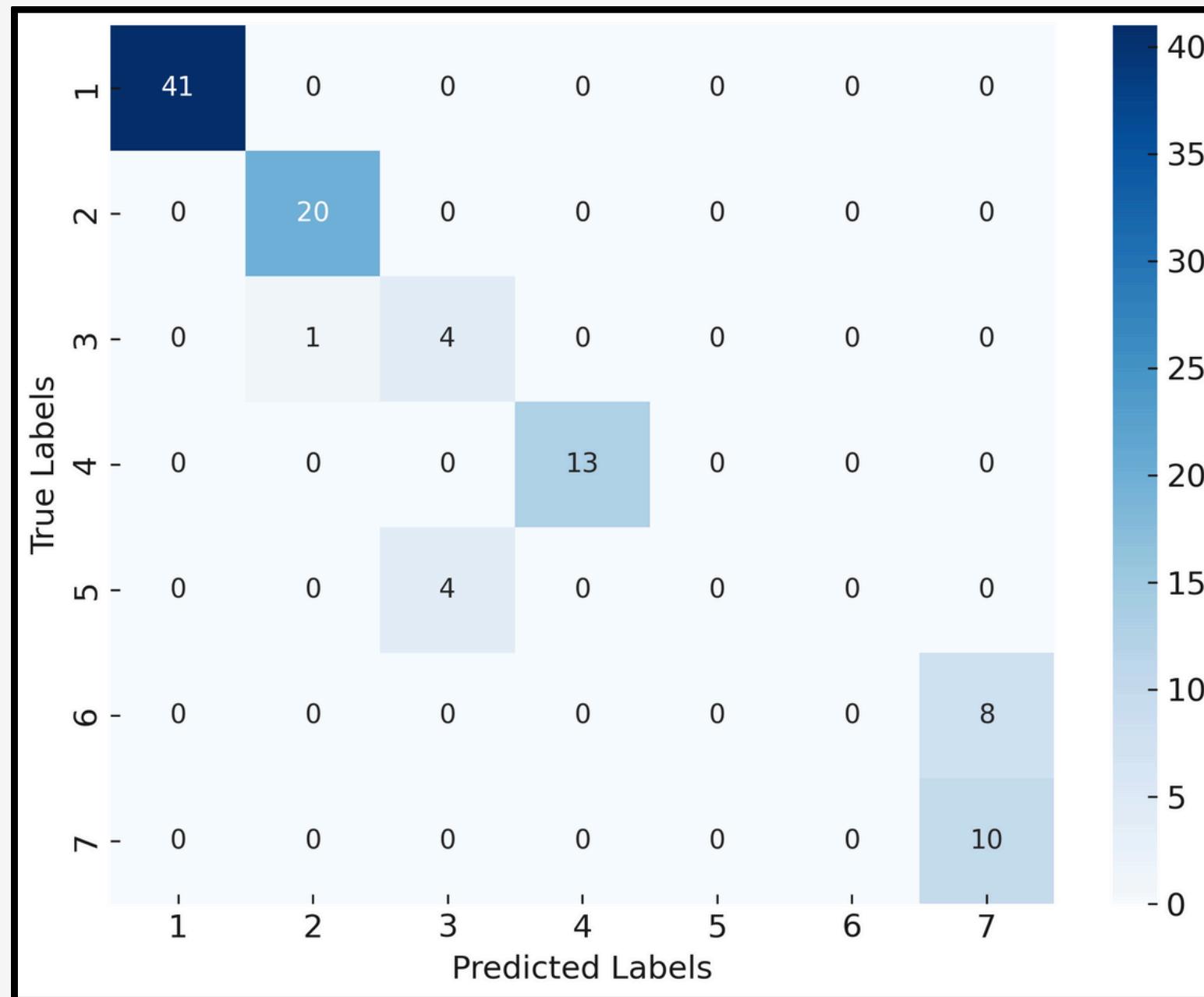
Accuracy
0.91

3D PCA Clustering of Animals



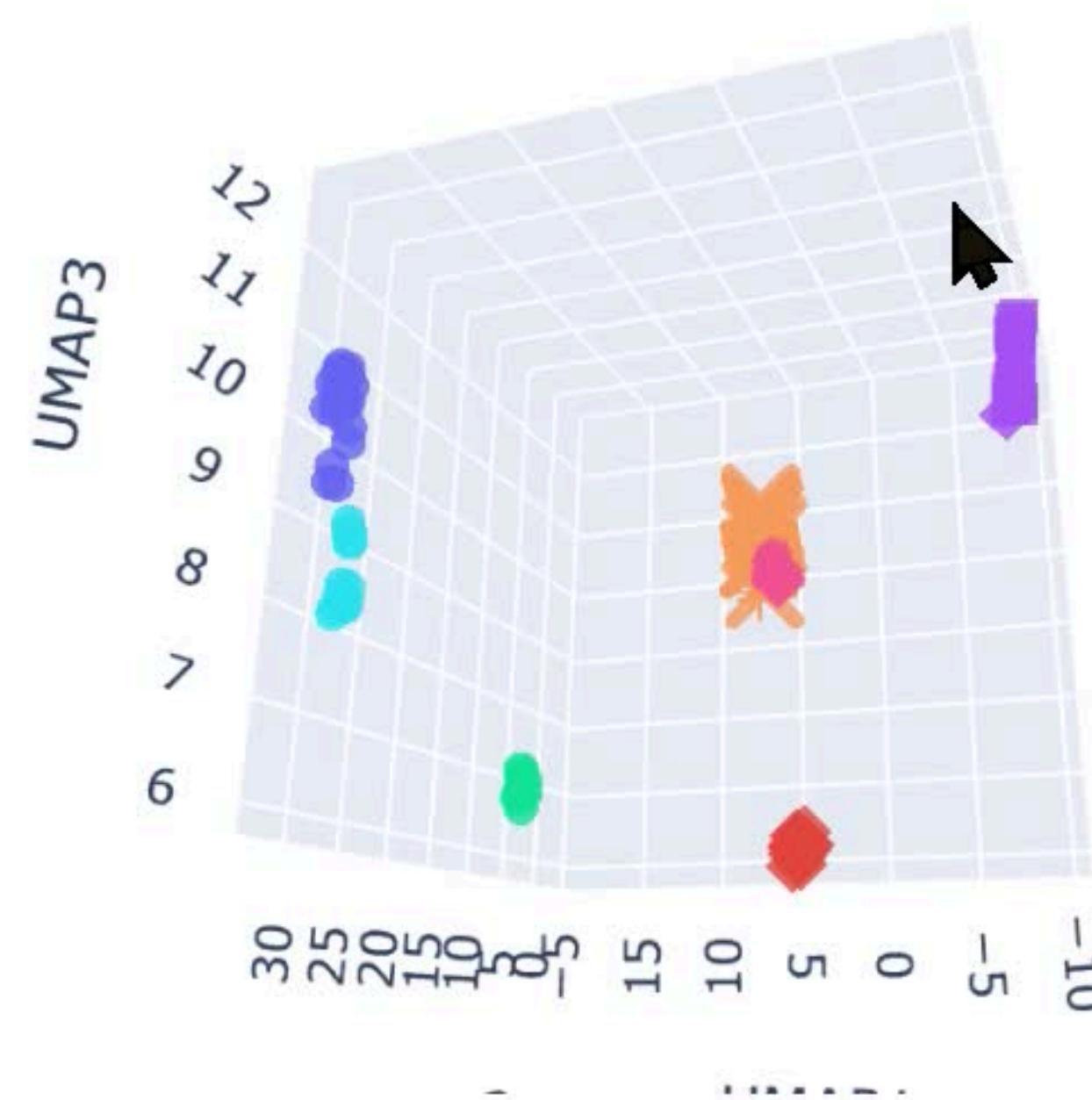
Model 3: K-means Clustering with UMAP (Uniform Manifold Approximation and Projection)

- UMAP is applied to reduce the data to 3 dimensions.
- K-Means clustering is performed on the reduced dataset.
- Cluster assignments are evaluated and visualized using an interactive 3D plot.



Accuracy
0.87

3D UMAP Clustering of Zoo Dataset



Comparison

Model	Accuracy
K-mean without Dimensionality Reduction	0.86
K-mean with PCA	0.91
K-mean with UMAP	0.87



THANK YOU FOR YOUR ATTENTION