SDS 2020: M1 Assignment 2

Unsupervised Learning with Pokémon

Description

This time you will work with Pokemon data. No data munging needed. Just old-school ML.

Data

The data is available through the URL:
https://sds-aau.github.io/SDS-master/00_data/pokemon.csv

It contains data on 800 Pokemon from the 1st to the 6th generation.

Tasks

1. Give a brief overview of data, what variables are there, how are the variables scaled and variation of the data columns.

2. Execute a PCA analysis on all **numerical** variables in the dataset. Hint: Don't forget to scale them first. Use 4 components. What is the cumulative explained variance ratio?
   Hint: I am not sure this terminology and code was introduced during class, but try and look into cumulative explained variance and sklearn(package) and see if you can figure out the code needed.

3. Use a different dimensionality reduction method (eg. UMAP/NMF) – do the findings differ?

4. Perform a cluster analysis (KMeans) on all numerical variables (scaled & before PCA). Pick a realistic number of clusters (up to you where the large clusters remain mostly stable).

5. Visualize the first 2 principal components and color the datapoints by cluster.

6. Inspect the distribution of the variable “Type1” across clusters. Does the algorithm separate the different types of pokemon?

7. Perform a cluster analysis on all numerical variables scaled and **AFTER** dimensionality reduction and visualize the first 2 principal components.

8. Again, inspect the distribution of the variable “Type 1” across clusters, does it differ from the distribution before dimensionality reduction?