SDS 2019: M1 Assignment 2

Description

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

Data

You will find the dataset for this assignment under:
https://github.com/SDS-AAU/M1-2019/raw/master/data/pokemon.csv

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

Tasks

1. **Unsupervised ML**

    a. Execute a PCA analysis on all **numerical variables** in the dataset. Hint: Don't forget to scale them before. Use 4 components. What is the cumulative explained variance ratio?
    b. Perform a cluster analysis (either k-means or hierarchical clustering algorithm) on all numerical variables (scaled & before PCA). Apply the elbow method to determine a “pragmatic” number of clusters.
    c. Visualize the first 2 principal components and color the datapoints by cluster.
    d. Inspect the distribution of the variable “Type1” across clusters. Does the algorithm separate the different types of pokemon?

2. **Supervised ML**
   Your task will be to predict the variable “legendary”, indicating if the pokemon is a legendary one or not.
    a. Perform necessary ML preprocessing of your data if deemed necessary.
    b. Split the data in a training (75%) and test (25%) dataset.
    c. Define a n-fold cross-validation workflow for your model testing.
    d. Fit three separate models on your training data, where you predict the “legendary” variable. Use a 1. Logistic regression, 2. Decision tree, and 3. another algorithm of choice to do so.
    e. Use the fitted models to predict the “legendary” variable in your test data.

f. Evaluate the performance of these 3 models by comparing the predicted and the true values of “legendary” in the test data. To do so, also create a confusion matrix.

Submission

18. September 12:00. Peergrade.io (link + submission details will be sent out on Monday)

Please submit a PDF version of your notebook with a link to the corresponding colab notebook included. Please make sure(eg. own test in “anonymous” setting in your browser) that others can acess it.