Università della Svizzera italiana

Facoltà di scienze informatiche

how to read the metadata??

how we should represent our data:
Dendrogram
maps for countries or places

QUESTIONS TO ANSWER:
what are different categories of food based on the manufacturing places?

what are the different categories of food based on their ingredients?

highest food based on macnumtirents?

Data Analytics

Academic Year 2022-23

Course Project N.02: Food

Prof. Fabio Crestani

For this assignment you will work individually to carry out simple tasks of data analysis given a specific dataset. The goal of this assignment is to use Python and complementary libraries on a given dataset in order to *explore* and *analyze* the given data and *draw conclusions*.

Description

The data lists various attributes of foods regarding their ingredients and categories. The dataset contains attributes such as creator, manufacturing places, categories, countries, additives, various ingredients, existing vitamins, etc.

This dataset contains a set of different features representing food facts. Your goal is to cluster these foods based on their various attributes. For example, what are different categories of food based on the manufacturing places? Or, what are the different categories of food based on their ingredients. Your tasks are to:

- Explore and describe the data (*i.e.*, standard descriptive statistics, visualize the variables with different graphs, draw distributions and histograms of variables, are there outliers? Any interesting observation? Any correlations? Etc.)
- Pre-process the data (*i.e.*, handle and fill unknowns if there are any, etc.)
- Use at least two clustering algorithm and compare them against one another. What is the most optimal number of clusters?
- Evaluate and compare the performances of the model.

Submission procedure and evaluation

You should produce a report of your work and its evaluation along with the source code. It will be a concise explanation of how you tackled the different tasks, the reasons of your choices, successive conclusions, plots you produced, results of the decisions and their accuracy, *etc*.

Use Jupyter Notebook to produce results of the commands in a single .ipynb file. For more information check: https://jupyter.org/documentation

The report (max 10 pages) and the code of the project need to be submitted via iCorsi.

Please, upload all the required items in a single file and name it following the structure: **noProject_proj.[zip|tar.gz|7z]**.

The dataset regarding this project can be downloaded from: http://ir.inf.usi.ch/da-datasets/

algorithms we should try using:
1. K-means clustering
2. BFR algorithm
3. CURE algorithm

FOR EACH ALGO DO MEASURE OF SIMILARRITY

https://www.kaggle.com/code/dhanyajothimani/basic-visualization-and-clustering-in-python

https://towardsdatascience.com/visualizing-clusters-with-pythons-matplolib-35ae03d87489

https://plotly.com/python/v3/ipython-notebooks/baltimore-vital-signs/

https://cseweb.ucsd.edu/~jmcauley/pml/code/chap9.html

https://towardsdatascience.com/best-practices-for-visualizing-your-cluster-results-20a3baac7426