Advanced business analytics - the power of predictive models

Course reference, mode of study: 226161-1380, SMMD and NMMS, type: laboratory

Task No. 3, deadline: May 19, 2023

Dataset:

https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

Task description:

1. [2 points] From the indicated set, select 3 variables that you will use in modeling - briefly justify your choice (based on EDA analysis - literature based choice may be additional work, but cannot be the basis for justifying the choice of variables). This section should include a graphical analysis, an analysis of relationships between variables and of variables distributions.
2. [1.5 points] Using the selected variables (adjust them if there's a need), build a segmentation model using the method K-means (according to the rules discussed in class). Briefly justify your choice of optimal number of clusters and the choice of the best starting points.
3. [1.5 points] Describe the groups selected on the basis of the model, interpret the statistics obtained for them. The created clusters should be visualised.

Guidelines:

Tasks can be solved and written to the output report using R Markdown, JupiterLab or SAS. Please ensure that the report format requirements (.html or .pdf) are met in each case. **All group members should be listed at the very beginning of the report.**

Scoring: max 5 points

Substantive side, programming, quality of report execution.

How to submit the task:

A dedicated folder for this task will be created in MS Teams to which one member of a project group should upload the task solution. The person from the group should attach the program code in the format (.sas, .py or .R) and the report (.html or .pdf) to the task in the MS Teams application joined in .zip or .rar file.

Attention! The archived file should be named according to following pattern: ***task_3_ab123456*.zip**, where "ab123456" is the student number of one of the team members.