Topic 9:

Clustering and Classification

Instructions:

The 2 students assigned to this topic can work separately on clustering and classification. Nevertheless, the video can be structured such, that it starts with a common definition of clustering and classification by highlighting the differences.

In the **Theory** part, should give a comprehensible introduction to the different approaches of clustering (such as K-means, Gaussian Mixture Models and Agglomerative Clustering) and Classification (such as Decision Trees [e.g. Random Forest], Naïve Bayes and Logistic Regression).

In the **Application** part, the given dataset should be cleaned and prepared for the further analysis and first the clustering methods should be applied with clear instructions how to implement the methods in python, which packages to use und how to evaluate the methods. Furthermore, the model outputs should be described comprehensibly with highlighting the most important elements. Tables and graphs should always be explained as well.

Also define some research questions regarding the patterns you want to examine.

Suggested procedure for the **Application** part:

1. Perform data cleaning and tests to meet the requirements for the methods (e.g. test for normal distribution if required)
2. Perform the clustering and classification methods step-by-step (use the Notebook Template provided in Isis)
   2.1. Check requirements (even if violated please continue)
   2.2. Run the models on the data and interpret the model output
   2.3. Validate the model's performance and compare the model outputs with each other
3. You can get creative and experiment with the methods.

Some references:

https://machinelearningmastery.com/clustering-algorithms-with-python/

https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/

If you have any open questions, please send an email to vera.schmitt@tu-berlin.de

Deadline for submission Video + Report: 06.01.22 upload both in the respective folder in Isis