Fundamentals of Data Science - project

The aim of the project is to familiarize you with practical algorithms and problems of modern data analysis and tools used to implement selected task. The project is **individual**. It is also aimed at encouraging you to use some of the modern data science tools.

The report along with any attachments (source code) should be submitted in electronic form using Microsoft Teams. A special assignment section will be prepared there.

The report should not contain too much theoretical information - only the most important practical aspects of the problem under consideration. Particularly appreciated will be the ability to analyze and formulate valuable technical conclusions critically.

The deadline for submitting the report is **December 16th 2024**. After this date, the final evaluation will take into account the delay in the implementation of the task. On that day we will also take exam. Information about the exam will be published on MS Teams.

Project meetings are not obligatory, but rather consultations for those interested.

The task:

Please download *fashion-mnist* dataset.

Please perform the following steps

1. Do the summary of the data and discuss it.
2. Reduce data dimensionality
3. Visualize the reduced dataset
4. Cluster the dataset (and evaluate clustering results with classification labels). Do we have some outliers here?
5. Split the dataset into training and testing.
6. Perform classification and evaluate its result.

Please ask ChatGPT how to do these things and compare how this information differs from the one you initially prepared. The result of this test should be discussed in the report. In your report, apart from that please present your approach and obtained results, code should be sent as an attachment or a link to a repository. The report should be concise and do not exceed 10-15 pages.

Please try to think outside the box and test methods other than the one used for the labs. The task is to be completed without using deep learning for classification. You can try however to complete steps 3-6 with some visual embeddings as an alternative to typical dimensionality reduction (e.g. VGG-16 embeddings).

Note: If you fall into a problems with computational complexity please resample the dataset to reduce its size.