1. *Distance Functions.* Given the dataset "distance-function-dataset.csv" (available in TeachCenter), select or develop a suitable distance function to compare instances (row), based on the values of the features (columns).

   (a) On what observations from the dataset do you base your decisions? (bullet points)

   (b) Would you conduct any additional feature engineering steps?

   (c) What distance function do you choose? (in case of a custom one, please provide a description/pseudocode/...)

   (d) Would the distance function depend on the succeeding processing, e.g., different function for PCA, DBSCAN, or SVM?

2. *Dimensionality Reduction.* Consider a dataset of 100 dimensions/features (real numbers), and the goal is to derive a 2D visualisation of the dataset.

   (a) What would be a suitable approach if the dependencies in the data are all linear?

   (b) What would be a suitable approach if there are density-based local structures in the data?

   (c) What would be a suitable approach if most of the features are Gaussain noise?

   (d) What types of noise are there and how do they affect the dimensionality reduction?

3. *Clustering.* Given the dataset “clustering-dataset.csv” (available in TeachCenter), which consists of observations of 5 dimensions, the goal is to find the groups of rows that form clusters.

   (a) Which methods did you apply to find the clusters, and why? (bullet points)

   - Describe pre-processing steps (if conducted)
   - Describe what distance measures you have chosen
   - Describe how you determine the number of clusters

   (b) What clusters did you find and how would you describe the distribution of the points within each cluster?

   - Describe each found cluster, including shape, and amount of points within the cluster.

4. *Classification.* Select 3 classification algorithms of your choice that should be diverse (i.e., not based on the same underlying principles).

   (a) For each algorithm list the main assumptions (e.g., on the data characteristics, types of dependencies). (bullet points)

   (b) For each algorithm list 1-2 application scenarios, where these assumptions are being met.