Assignment IV — Submission Sheet

This document contains questions to help you reflect about the operations applied to the data during this assignment. You have to fill it in and submit it on Studium.

Task 0: Warm-up

- **Number of instances:** 1200
- **Number of attributes:** 2

Number of instances in each class:

Class	Instances
A	960
N	240

Task 1: KNN-Based Anomaly Detection

Performance Results:

Metric	Value
AUC-ROC	0.8235
Average Precision	0.4754

Visualization Results: See the attatched image

Task 2: Parameter Sensitivity

Which "n_neighbors" and "method" corresponds to the best performance? Write the number of neighbors and the performance score for the specified metric:

Metric (Best)	n_neighbors	Method	Score
AUC-ROC	8	mean	0.94
Average Precision	6	mean	0.81

Task 3: Local Outlier Factor (LOF)

Which n_neighbors corresponds to the best performance? Write the number of neighbors and the performance score for the specified metric:

Metric (Best)	n_neighbors	Score
AUC-ROC	4	0.52
Average Precision	4	0.26

According to the results from trying different `n_neighbors`, which algorithm (KNN or LOF) is more sensitive to hyperparameters? Please explain your findings below.

LOF is worse for this dataset BUT is way less sensitive to hyperparameters as the ROC-AUC and Average Precision vary way less as we adjust the hyperparameter

---

Task 4: Real-world Anomaly Detection

- **Which algorithms and metrics have you chosen for this task? Report and explain the results from your analysis.**

Algorithms	AUC-ROC	Average Precision
LOF	0.5899	0.1953
IForest	0.9326	0.5549
OCSVM	0.6092	0.2013

In this specific case the isolation forest greatly outperformed both the local outlier factor and one-class SVM. This might be due to the overall density in the dataset being low with little variance, meaning densisty/distance related functions might perform worse.

- **Have you noticed any difference between the metrics in use and the chosen algorithm?** While LOF and OSCVM perform more similarly to each other the AUC-ROC metric is way higher for the Isolation forest algorithm. This might be due to it not relying on distance for classification.
- **Which algorithm seems more appropriate for this task?**
  Given its splendid perofmance, the isolation forest algorithm seems most appropriate for the cardio data set. Since the isolation forest is also used in other health-care related datasets it is the best candidate.