ANOMALY DETECTION IN HYPOTHYROIDISM DATASET

Unsupervised Learning

Andrea Palmieri - 921785 - a.palmieri13@campus.unimib.it
Andrea Yachaya - 913721 - a.yachaya@campus.unimib.it

AGENDA

- Dataset, preprocessing and metrics
- Hierarchical Clustering
  - Average and Complete Linkage
- K-Means and K-Medoids
- K-Prototypes
- DBSCAN
- Local Outlier Factor
- Adjusted Rand Index
- Results

INTRODUCTION

Focus of Study:

- Application of the following unsupervised learning methods for anomaly detection on unlabelled hypothyroidism dataset.
  - Distance-Based clustering
  - Density-Based clustering

Dataset Overview:

- 7,200 observations
- **Mixed data types:**
  - 15 binary
  - 6 continuous features.
- No labeled data available.

DATASET PREPROCESSING

Continuous Features Standardization:

- To prevent any single feature from disproportionately influencing the results, they were scaled using *Z-Score*

$$Z = \frac{x - \mu}{\sigma}$$

Binary Features Handling:

- As these features already indicated the presence or absence of a feature, no one-hot encoding was necessary

GOWER DISTANCE

Euclidean distance not suitable due to mixed data types nature of dataset

Therefore **Gower Distance was used**

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \cdot \delta_k(x_{ik}, x_{jk})}{\sum_{k=1}^{p} w_k}$$

$$\delta_k(x_{ik}, x_{jk}) = \frac{|x_{ik} - x_{jk}|}{\max(x_k) - \min(x_k)}$$ Numerical Features

$$\delta_k(x_{ik}, x_{jk}) = \begin{cases} 0 & \text{if } x_{ik} = x_{jk} \\ 1 & \text{if } x_{ik} \neq x_{jk} \end{cases}$$ Categorical Features

All weights equal = 1 due to missing specific domain knowledge

DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA):

- PCA transforms data into a set of orthogonal uncorrelated components, ordered by the amount of variance they explain in the data.
- PCA is not suitable for mixed data types.
    - E*xcluded from this study.*

Factor Analysis of Mixed Data (FAMD):

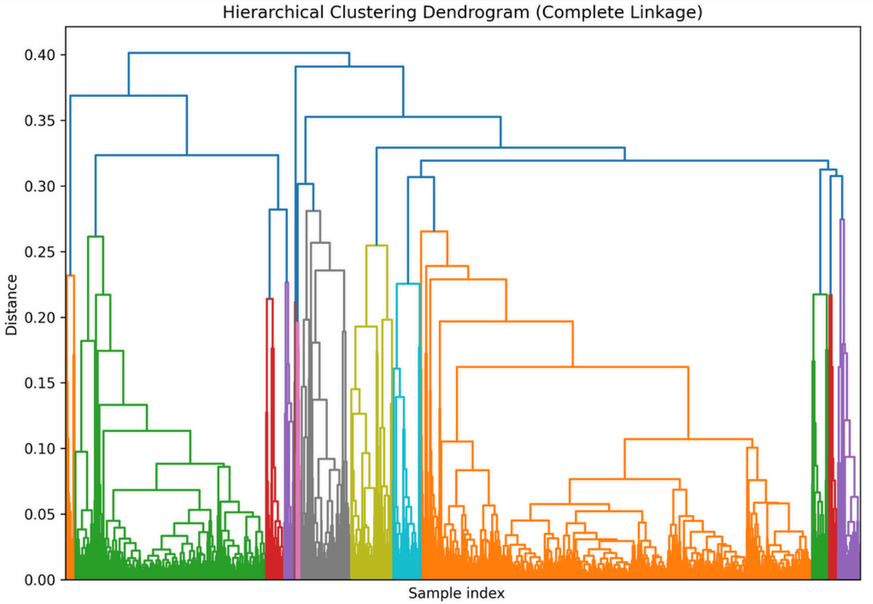
- FAMD is specifically designed to handle mixed data types datasets
- FAMD was used only for visualization
    - It’s not suited for anomaly detection due to the difficulty in reconstructing the original features from the FAMD coordinates

HIERARCHICAL CLUSTERING: DENDROGRAMS

Hierarchical clustering produces nested clusters, visualized as a dendrogram. There are several methods to compute inter-cluster proximity: single, complete, average, centroids-based and Ward's.
We opted for complete and average linkage for their compatibility with mixed data types datasets.

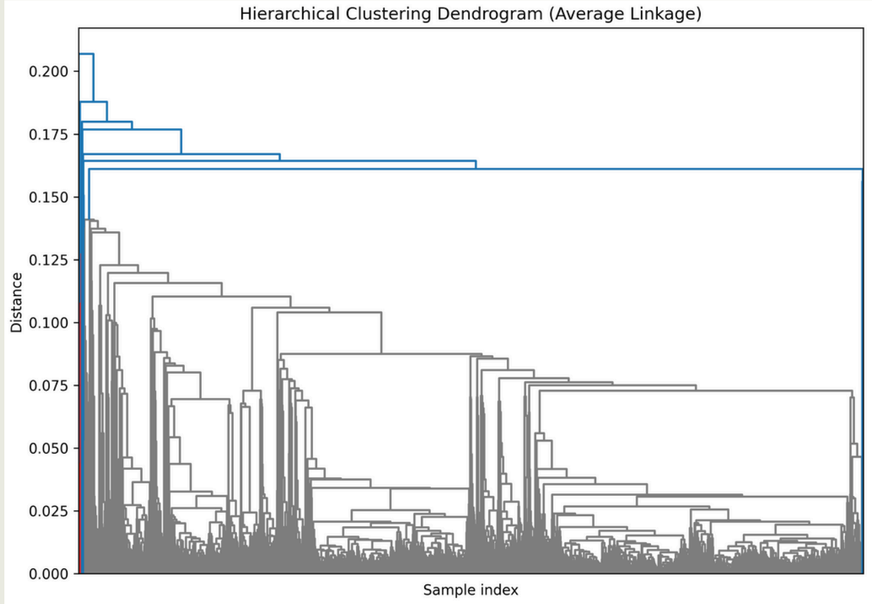
COMPLETE LINKAGE

Highly branched structure with many **small and tight clusters.**



AVERAGE LINKAGE

A more balanced hierarchical structure with fewer branches and **larger clusters**, and a **gradual increase in distance** between clusters as they merge.



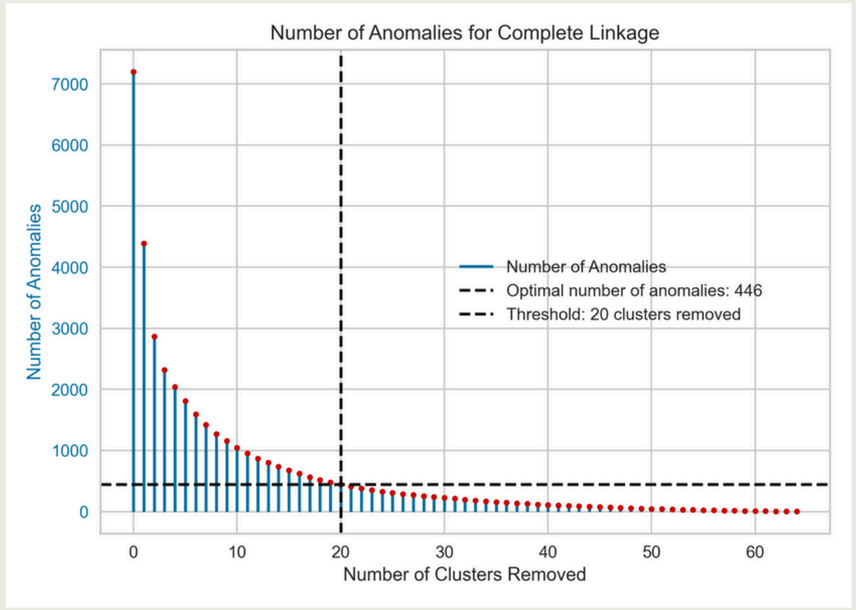
HIERARCHICAL CLUSTERING: DETECTING ANOMALIES

The optimal distance metric was the one with the highest Silhouette score.
Clusters were ranked by size, and larger clusters were excluded to identify anomalies. The exclusion threshold was set where anomaly numbers stopped changing significantly.

COMPLETE LINKAGE

Optimal distance: 0.147, Silhouette score: 0.519.

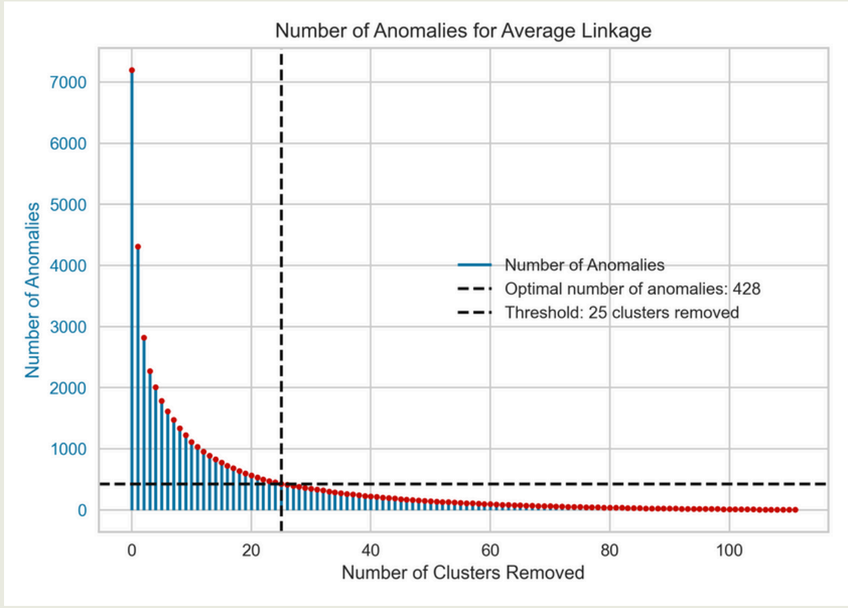
At 20 clusters removed, 446 anomalies found.



AVERAGE LINKAGE

Optimal distance: 0.065, Silhouette score: 0.649.

At 25 clusters removed, 428 anomalies found.

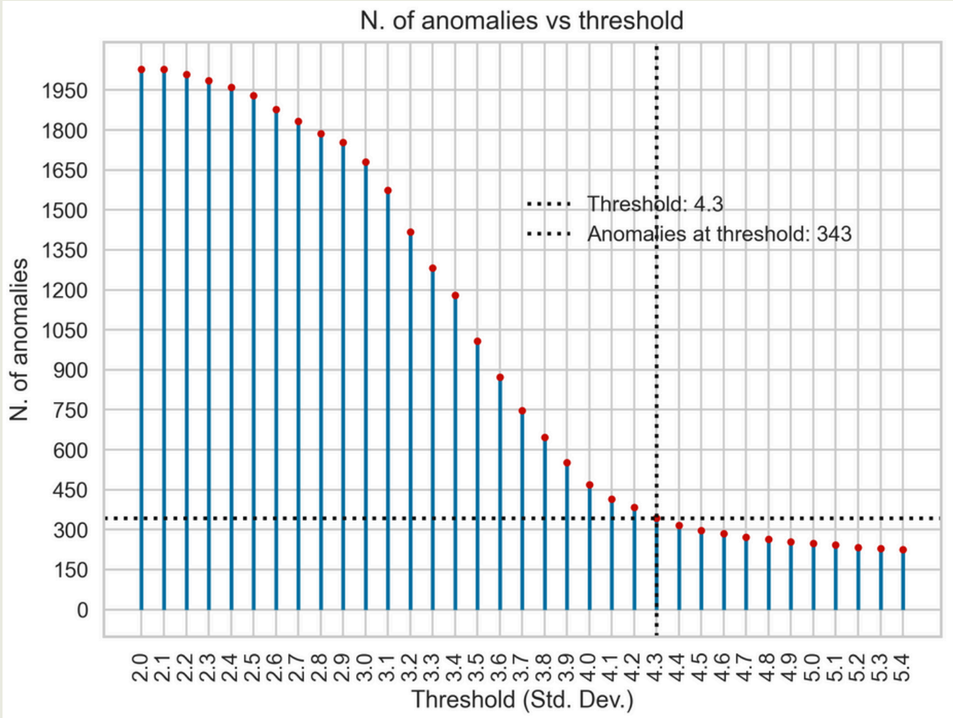


K-MEANS AND K-MEDOIDS

K-Means was not used due to the Euclidean distance incompatibility with mixed data types datasets.

K-Medoids

- can use the precomputed Gower Distance matrix
  - suitable for our dataset
- uses **medoids** - actual data points - as centers of the clusters.
- **Optimal number of clusters**
  - determined using the elbow method: **5 clusters.**
- **Anomalies** were **identified** by elements deviating more than a specified number of standard deviations from the mean distance to their nearest medoid.



343 anomalies found with a threshold of 4.3 standard deviations from the mean.

K-PROTOTYPES

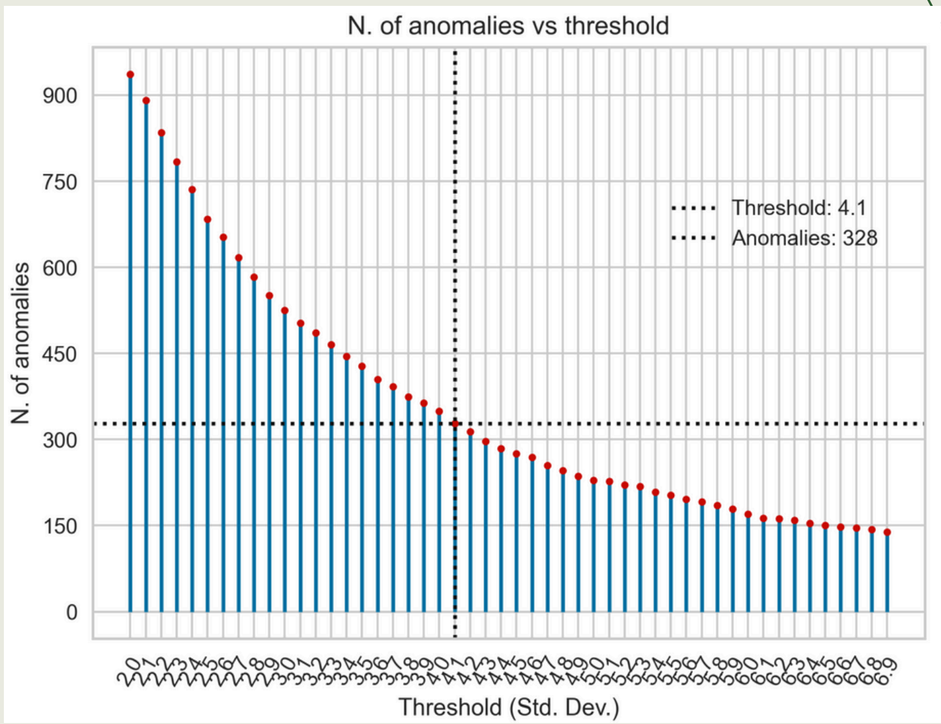
Uses **prototypes** - synthetic points - as centers of the clusters, computed as:

- for **numerical attributes:** the **mean** of the cluster
- for **categorical attributes:** the **mode** of the cluster

Optimal number of clusters

- determined using the elbow method: **4 clusters.**

Anomalies identified by elements deviating more than a specified threshold from their prototype



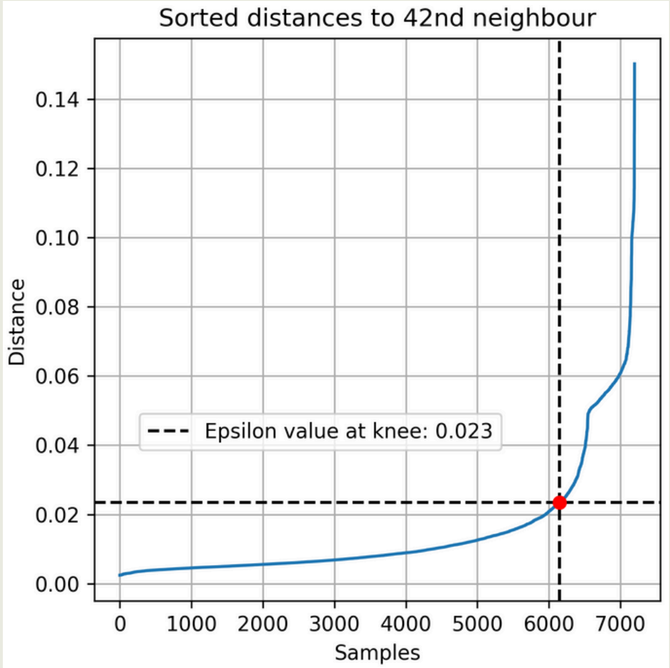
328 anomalies found with a threshold of 4.1 standard deviations from the mean

DBSCAN

Density-based method that groups data points based on the number of neighbours in a radius.

In requires **two key parameters:**

- **ε (epsilon)**: radius around a point to search for neighbors
  - A **starting value** was determined by computing the distance from the farthest 42nd data point, then by considering the knee of the resulting curve



- **MinPts:** minimum number of points within ε-radius to define a core point
  - **Determined heuristically as twice the number of features***
    - MinPts = 42

We used the precomputed Gower Distance matrix.

* [Sander et al. 1998] J¨org Sander, Martin Ester, HansPeter Kriegel, and Xiaowei Xu. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery, 2:169–194, 1998.

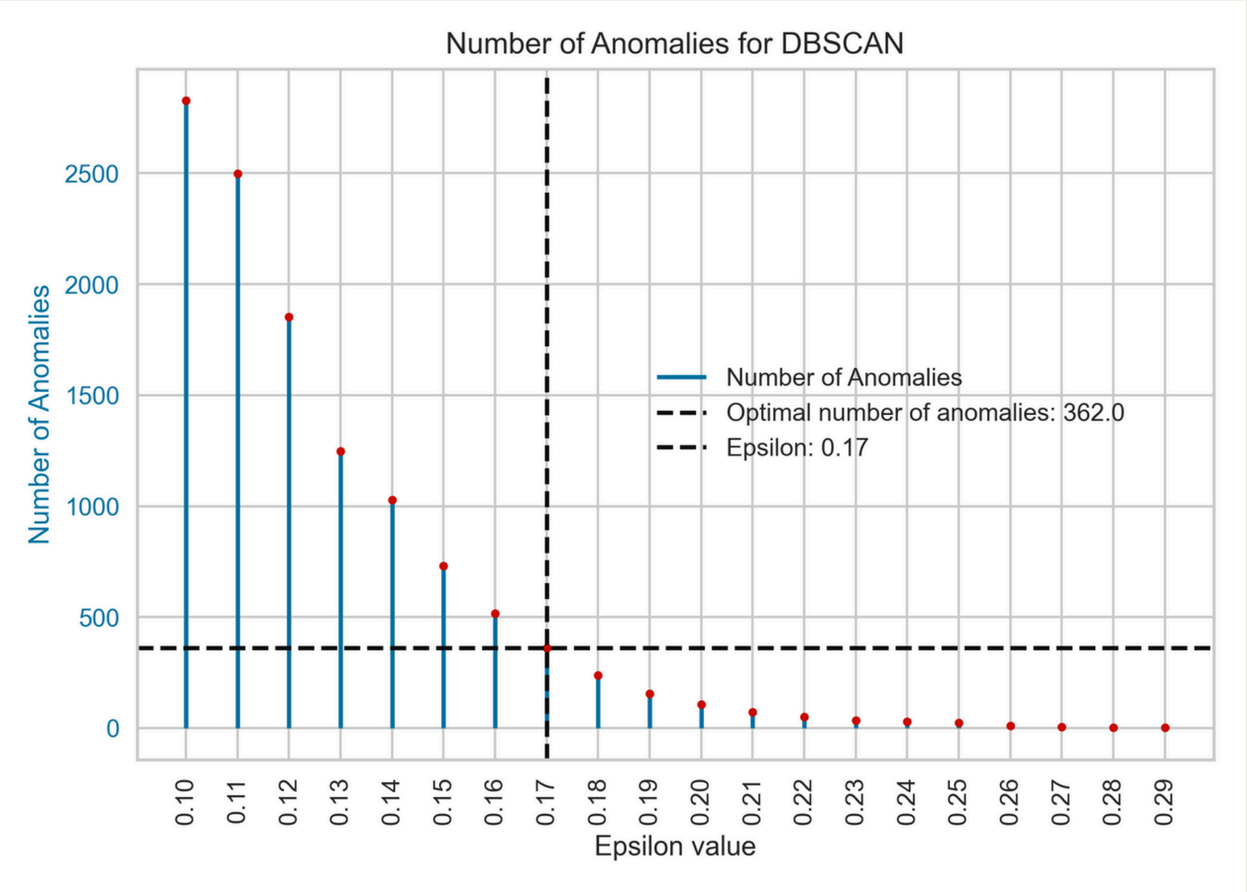
DBSCAN

Anomalies were **identified** as directly proportional to the average distance to neighbouring points and inversely proportional to the density (number of points in the ε-radius).

The **optimal Epsilon** was chosen visually as the point where anomaly numbers stopped changing significantly.

Sparser points will be considered as more anomalous.

362 anomalies found with a Epsilon of 0.17



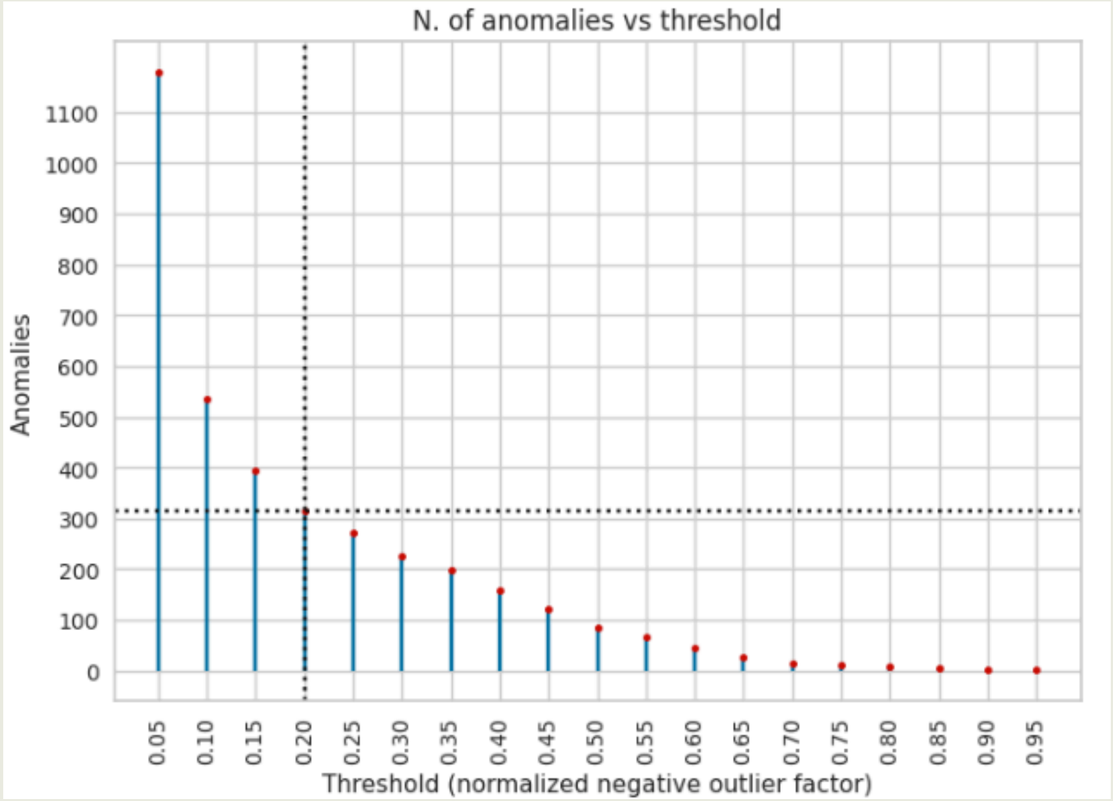
LOCAL OUTLIER FACTOR (LOF)

Computed using the precomputed Gower Distance matrix.

For each data point, **LOF returns its negative outlier factor** of being anomalous, then normalized to [0 - 1]

- The threshold was chosen visually

316 anomalies found with a threshold of 0.20



ADJUSTED RAND INDEX

To evaluate the differences in the results returned by different anomalies detection techniques, we employed the use of the **Adjusted Rand Index (ARI)**, which is a **measure used to evaluate the similarity between two data clusterings,** accounting for the possibility of random agreement.

ARI	HC Average	HC Complete	DBSCAN	K-Medoids	K-Prototypes	LOF
HC Average	1	0.70	0.74	0.54	0.05	0.67
HC Complete	0.70	1	0.54	0.41	0.04	0.58
DBSCAN	0.74	0.54	1	0.57	0.07	0.47
K-Medoids	0.54	0.41	0.57	1	0.12	0.32
K-Prototypes	0.03	0.01	0.04	0.09	1	0.30
LOF	0.67	0.58	0.47	0.32	0.05	1

COMPUTING THE FINAL PROBABILITIES

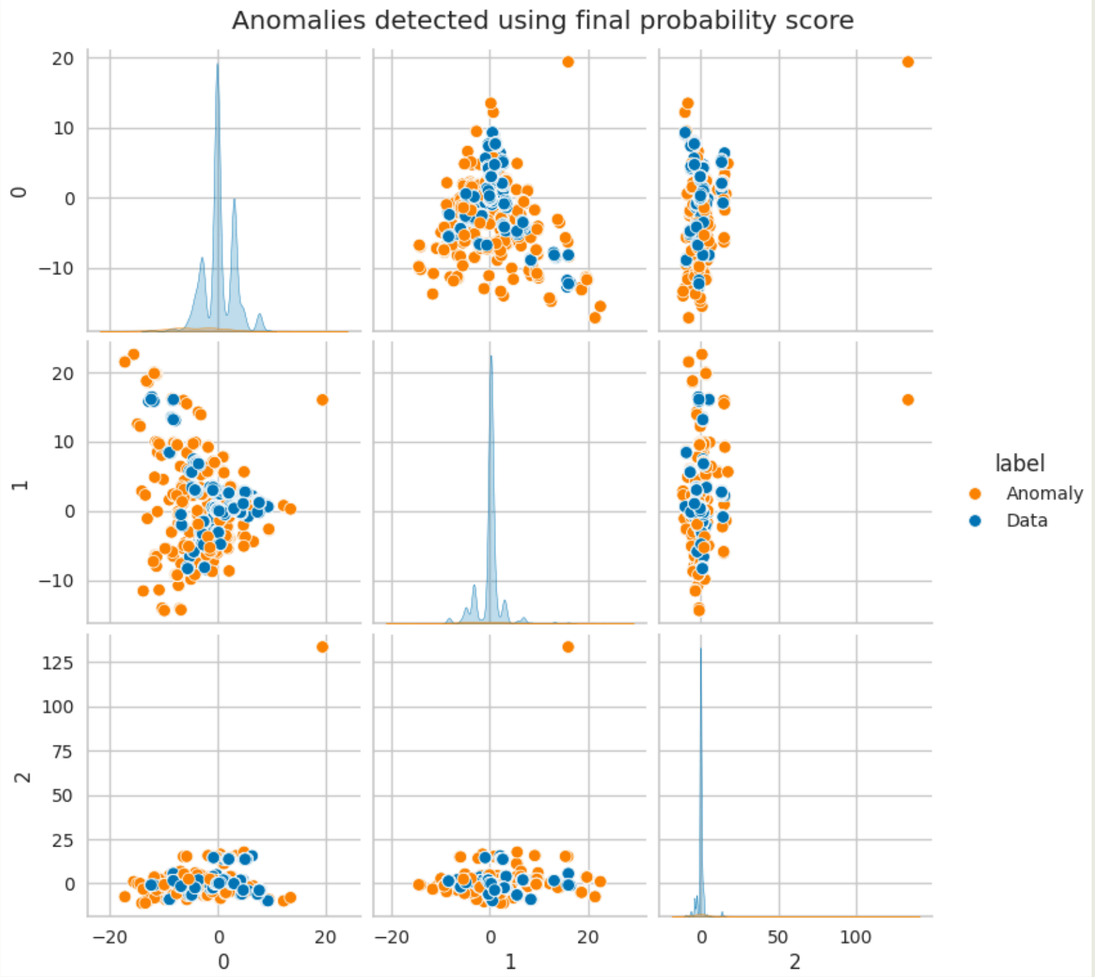
The **final probabilities are the mean probability** of

- HC Average
- HC Complete
- DBSCAN

probabilities.

A total of **402 anomalies** were found.

ARI	HC Average	HC Complete	DBSCAN
HC Average	1	0.70	0.74
HC Complete	0.70	1	0.54
DBSCAN	0.74	0.54	1



Thank you.

Andrea Palmieri - 921785 - a.palmieri13@campus.unimib.it
Andrea Yachaya - 913721 - a.yachaya@campus.unimib.it