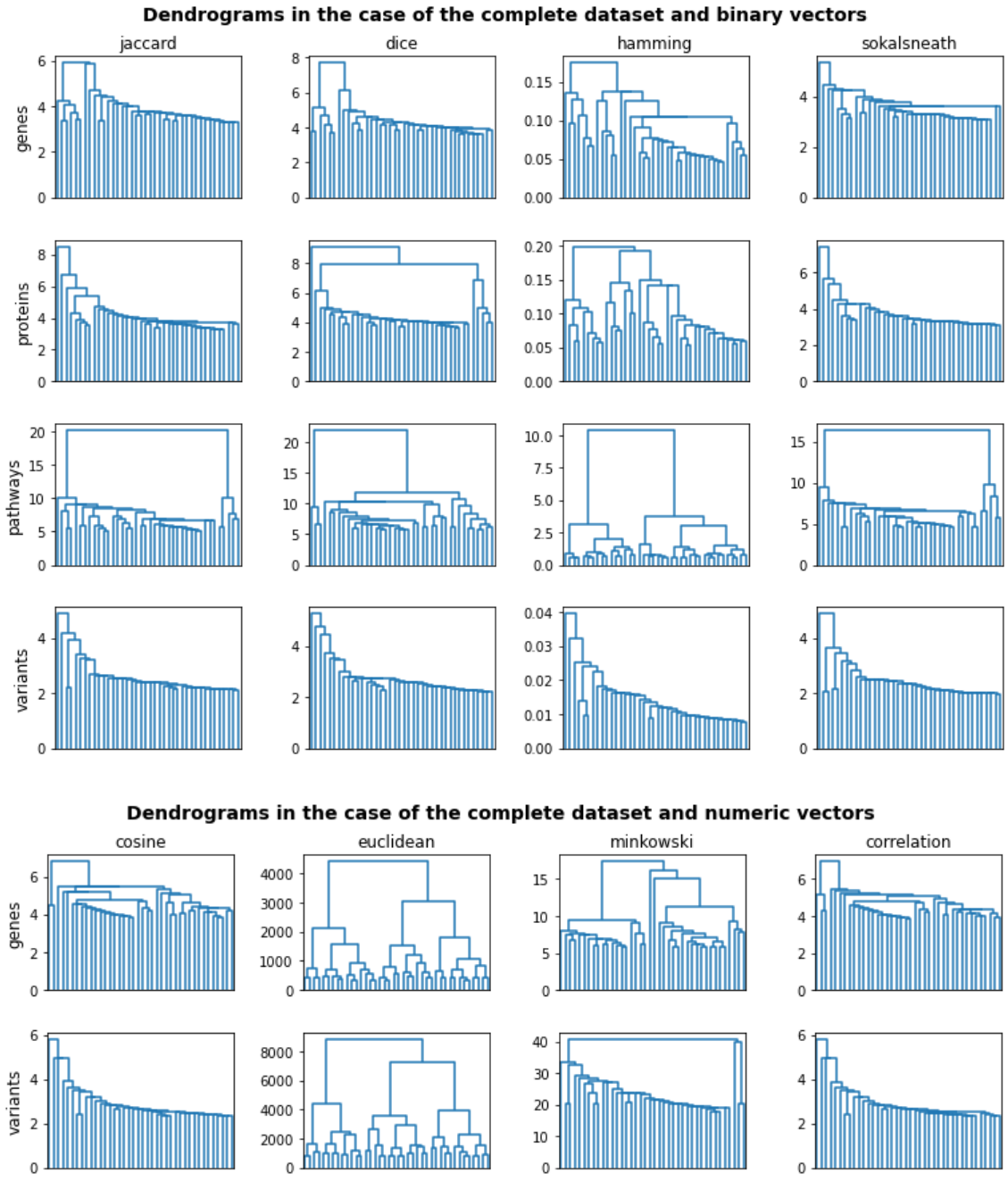
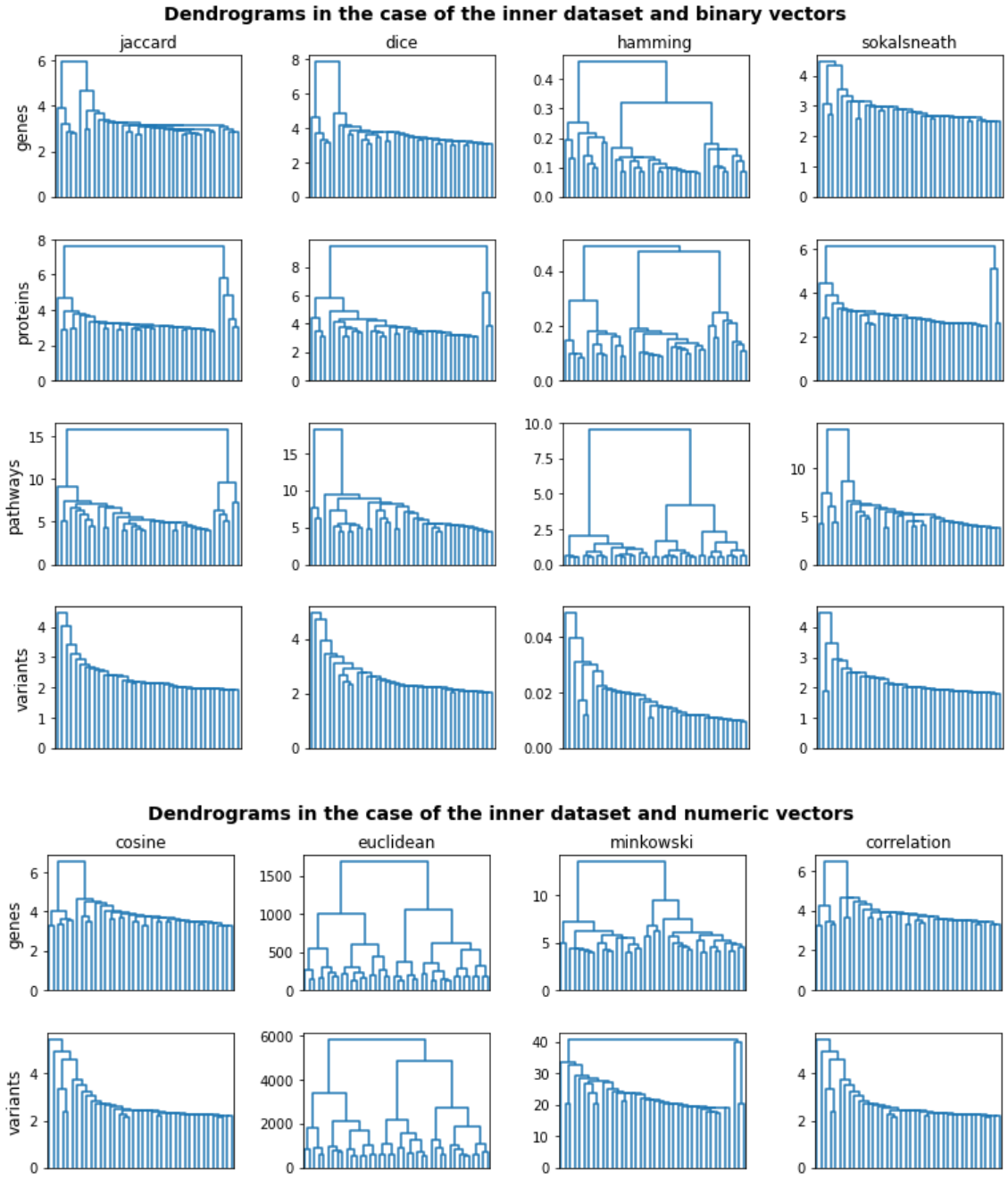
# Supplementary Information – Publication I

## S1. Hierarchical clustering dendrograms

From the different distance matrices (obtained from the distinct datasets, type of vectors and features) and using linkage method “ward”, the following dendrograms have been generated.




***Supplementary Figure S1. Representation of the dendrograms generated from the complete sets of diseases by applying hierarchical clustering.*** *We have included both the dendrograms obtained by computing distance matrices from binary and numeric vectors (and considering each metric and feature).*




***Supplementary figure S2. Representation of the dendrograms generated from the inner set of diseases by applying hierarchical clustering.*** *We have included both the dendrograms obtained by computing distance matrices from binary and numeric vectors (and considering each metric and feature).*

# S2. Formal distance metrics definitions

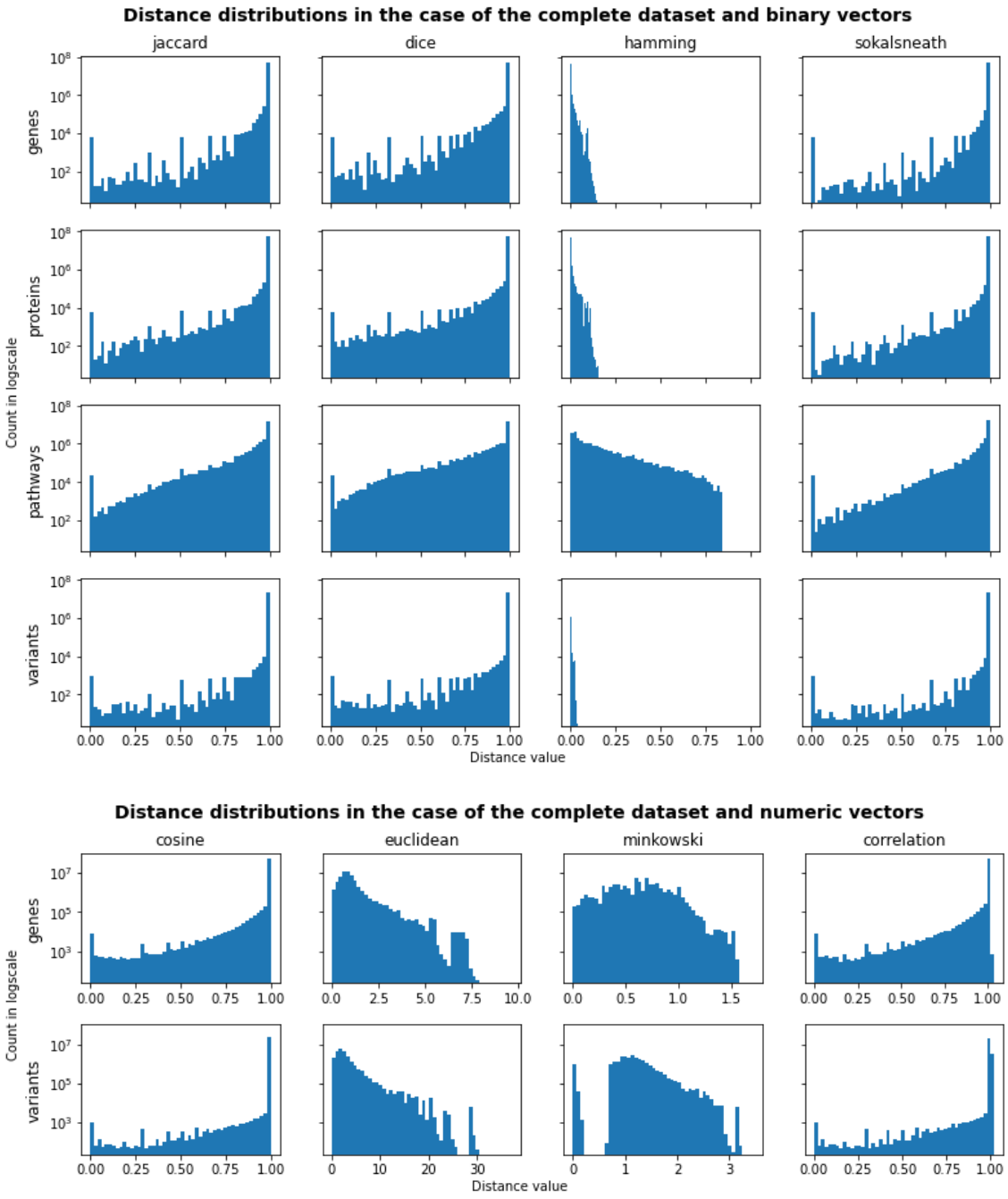
The distance between two diseases based on the feature vectors representing them was computed based on the metrics that are here defined. Distances were calculated between pair of diseases, represented by A and B, which are the feature vectors associated to each of the diseases. The vectors were formed of *i* features, from 1 to *N*.

- **Dice** $d_{Dice}(A,B) = 1 - \frac{2\sum_{i=1}^{N}(A_i \times B_i)}{\sum_{i=1}^{N} {A_i}^2 + \sum_{i=1}^{N} {B_i}^2}$
- **Hamming** $d_{Hamming}(A,B) = \sum_{i=1}^{N} |(A_i - B_i)|$
- **Jaccard** $d_{Jaccard}(A,B) = 1 - \frac{\sum_{i=1}^{N}(A_i \times B_i)}{\sum_{i=1}^{N} {A_i}^2 + \sum_{i=1}^{N} {B_i}^2 - \sum_{i=1}^{N}(A_i \times B_i)}$
- **Sokal-sneath** $d_{Sokal-Sneath}(A,B) = 1 - \frac{\sum_{i=1}^{N}(A_i \times B_i)}{\sum_{i=1}^{N} {A_i}^2 + 2\sum_{i=1}^{N} {B_i}^2 - 2\sum_{i=1}^{N}(A_i \times B_i)}$
- **Correlation** $d_{Correlation}(A,B) = 1 - \frac{\sum_{i=1}^{N}((A_i - \overline{A}) \times (B_i - \overline{B}))}{\sqrt{\sum_{i=1}^{N}(A_i - \overline{A})^2} \times \sqrt{\sum_{i=1}^{N}(B_i - \overline{B})^2}}$
- **Cosine** $d_{Cosine}(A,B) = 1 - \frac{\sum_{i=1}^{N}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{N} {A_i}^2} \times \sqrt{\sum_{i=1}^{N} {B_i}^2}}$
- **Euclidean** $d_{Euclidean}(A,B) = \sqrt{\sum_{i=1}^{N}(A_i - B_i)^2}$
- **Minkowski** $d_{Minkowski}(A,B) = \left(\sum_{i=1}^{N} |(A_i - B_i)^p|\right)^{1/p}$

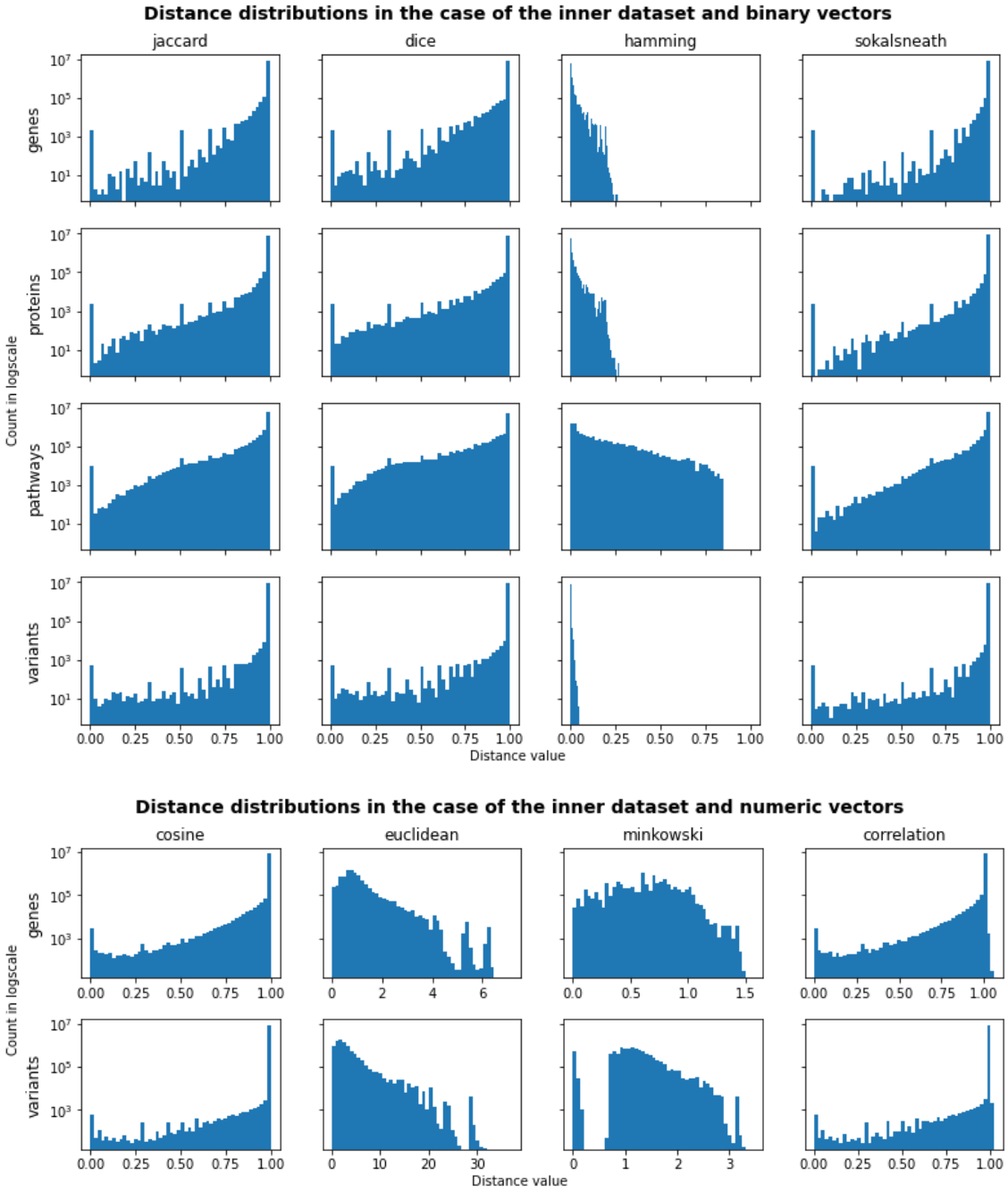
Minkowski's *p* parameter was set in this paper to 5.

# S3. Distributions of the distance matrices

The different distance distributions accordingly to the distinct dataset, type of vectors, features and metrics have been represented in the following plots.




***Supplementary figure S3. Representation of distance distributions between diseases in the complete datasets.*** *We have included both the distributions of distances obtained from binary and numeric vectors (and considering each metric and feature).*




***Supplementary figure S4. Representation of distance distributions between diseases in the inner dataset.*** *We have included both the distributions of distances obtained from binary and numeric vectors (and considering each metric and feature).*

# S4. Formal evaluation metrics definitions

To evaluate the clustering results obtained, different internal evaluation metrics were considered. They are described hereunder. They all refer to the computed global value for model (taking into account all the instances).

The notation used will be the same for the different metrics. For dataset, $X$, of $N$ diseases, supposing $X$ is partitioned into $K$ clusters, $C = \{c_1, c_2, ..., c_K\}$, and represented as vectors in an $F$-dimensional space: $X = \{X_1, X_2, \ldots, X_N\} \subseteq \Re^F$. The centroid of a cluster $c_K$ is its mean vector, $\overline{c_k} = 1 / |c_k| \sum_{x_i \in c_k} x_i$ and, likewise, the centroid of the dataset is the mean vector of the whole dataset, $\bar{X} = 1/N \sum_{x_i \in X} x_i$. The distance between two points $x_1$ and $x_2$, computed by whatever metric, is denoted as $d(x_1, x_2)$.

**Silhouette coefficient**

The Silhouette coefficient is defined for each sample and is composed of two scores, shown below. Higher values of this metric are related to a model with better defined clusters.

For each disease $x_i \in X$, $a(x_i)$ is calculated as the average distance between $o$ and all other objects in the cluster to which $o$ belongs. Similarly, $b(x_i)$ is the minimum average distance from $o$ to all clusters to which $o$ does not belong. Formally, $a(x_i, c_k) = \frac{\sum_{x_j \in c_l} d(x_i, x_j)}{|c_k| - 1}$, and $b(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ \frac{\sum_{x_j \in c_l} d(x_i, x_j)}{|c_l|} \right\}$.

That is, $a$ is the mean distance between a sample and all other points in the same class, measuring the closeness of points in the same cluster, whereas $b$ is the mean distance between a sample and all other points in the next nearest cluster, measuring the distance of points to different clusters.

The Silhouette coefficient of $x_i$ is then defined as $s(x_i, c_k) = \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k),\ b(x_i, c_k)\}}$, while the Silhouette coefficient of the whole model is computed as $S(C) = 1/N \sum_{c_k \in C} \sum_{x_i \in c_k} s(x_i, c_k)$.

**Calinski-Harabasz index**

CH index is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid. The separation is based on the distance from the centroids to the global centroid. This metric is computed as follows: $CH(C) = \frac{N - K}{K - 1} \frac{\sum_{c_k \in C} |c_k| d(\overline{c_k}, \bar{X})}{\sum_{c_k \in C} \sum_{x_i \in c_k} d(x_i, \overline{c_k})}$.

**Davies-Bouldin index**

DB index estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. It is defined as follows: $DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{d(\overline{c_k}, \overline{c_l})} \right\}$, where $S(c_k) = 1/|c_k| \sum_{x_i \in c_k} d(x_i, \overline{c_k})$.