Università degli studi di Milano-Bicocca

Study of the DBCV Statistical Metric for Internal Clustering in Unsupervised Machine Learning

Authors:
Sabino Giuseppe - 852287 - g.sabino@campus.unimib.it

September 12, 2024

Contents

1 Introduction

Evaluating clustering outcomes is one of the primary challenges in clustering analysis. Validation of clustering can be categorized into three types: external, internal, and relative methods [1]. External validation methods, such as the Adjusted Rand Index, assess the clustering results by comparing them to a known clustering solution. However, since clustering is inherently an unsupervised method without any predefined ground truth, external measures are often impractical.

In practical applications, internal and relative validation measures are more commonly used. Internal validation metrics assess the clustering quality based solely on the data at hand. Relative validation metrics, a subset of internal metrics, allow for the comparison of different clustering solutions to identify the best one.

In the context of density-based clustering, clusters are viewed as regions with high data density separated by areas of low density. Despite the extensive research on relative validation measures, little attention has been given to evaluating density-based clustering results.

Following the original paper [2], the aim of this project is to produce documentation that can serve as a guide to the metric DBCV (Density-Based Clustering Validation), a technique for assessing the quality of clustering results. It evaluates how well the identified clusters reflect the underlying data structure by measuring the density of points within clusters, both locally and globally. This approach helps determine if the formed clusters are meaningful and accurately represent the true structure of the data, rather than being artifacts of the clustering algorithm.

2 State of the art

To provide a comprehensive overview of the current advancements in density-based clustering evaluation, this section reviews recent studies and methodologies. The focus is on the application of the Density-Based Clustering Validation (DBCV) metric, a pivotal tool for assessing the quality of clustering results. The DBCV metric's ability to measure both local and global point density within clusters ensures a more accurate reflection of the underlying data structure. This review highlights the utility and superiority of DBCV in various domains, illustrating its effectiveness in improving clustering outcomes and addressing the limitations of traditional validation metrics.

Subtyping schizophrenia using polygenic scores (PGS) can enhance therapeutic decision-making by overcoming the limitations of traditional methods. Recent studies have identified PGS clusters associated with specific genetic characteristics and treatment patterns, demonstrating the effectiveness of advanced clustering approaches such as UMAP and DBSCAN, with the DBCV index used as the validation metric [4]. Recent research has developed a privacy-preserving density-based clustering protocol using DBSCAN with secure two-party computation. The quality of clustering is evaluated with measures such as the Adjusted Rand Index, the Silhouette Coefficient, and the Density-Based Clustering Validation (DBCV), the latter proving superior for evaluating density-based clustering because it also considers noise, providing a more accurate assessment compared to traditional metrics [5].

Density-based clustering identifies regions of high density separated by regions of low density. A new density-based clustering algorithm employs the concept of "reverse nearest neighbour" and a single parameter, which can be estimated using a clustering validity index. The Density-Based Clustering Validation (DBCV) has proven superior for assessing the quality of density-based clustering, surpassing traditional indices like the Silhouette Width and Dunn's Index, as it measures the relative density connection between entities [6]. Machine learning techniques are employed to automatically identify and evaluate subtypes of hospitalized patients using routinely collected data, such as the National Early Warning Score 2. An iterative hierarchical clustering process, including dimensionality reduction via UMAP and clustering with HDBSCAN, uses the Density-Based Clustering Validation metric to evaluate clustering quality, showing superior performance compared to other validation metrics [7].

A systematic review examined automated methods for white matter tract segmentation, analyz-

ing various approaches including voxel-based, streamline-based clustering, and atlas-based methods. Among the evaluation metrics, the Density-Based Clustering Validation emerged as a key measure, providing an accurate assessment of the quality of density-based clustering [8].

Unlike other validation metrics, DBCV accounts for both the internal density of clusters and the density between clusters, making it particularly suitable for algorithms like DBSCAN. Its ability to consider noise and density variability in the data makes it superior to metrics such as the Silhouette Index and the Davies-Bouldin Index [9]. A study utilized quality metrics and ensemble strategies to enhance optimization and reduce computational load. The effectiveness of the Density-Based Clustering Validation index was demonstrated, showing promising performance regardless of cluster shape and in managing noisy datasets. The approach leverages the integration of multiple partitioning solutions to achieve more robust and effective results [10].

Process mining is an emerging discipline aimed at extracting process-based knowledge from event logs collected by business systems. Literature analysis using text mining and machine learning techniques has identified and predicted trends in key research areas. Using BERTopic and validating the results with the DBCV score, 49 topics were derived with a DBCV score of 0.366. This methodology demonstrated that DBCV is an effective metric for validating density-based clustering, ensuring accurate and non-overlapping cluster representation [11].

An approach combining machine learning and psychology offers new perspectives on the conceptualization of psychological processes. Utilizing UMAP and HDBSCAN, evaluation is based on the number of clusters produced and the unclustered points. The decision to forego application-agnostic measures like the Density-Based Cluster Validity Index DBCV, in favor of criteria closely tied to the research question, yielded more relevant results [12]. For extracting topics related to COVID-19, UMAP was applied to reduce the dimensionality of vector representations of hashtags, followed by clustering with HDBSCAN. A grid search identified the model with the highest relative validity score, using a fast approximation of DBCV to evaluate density-based and arbitrarily shaped clusters. The resulting clusters represent COVID-19-related topics, with topic vectors defined as the weighted mean of the hashtag vectors within the cluster [13].

3 DBCV

According to Hartigan's Density Contour Trees model [3], a good density-based clustering solution should have clusters where the lowest density within any cluster is still higher than the highest density in the regions between clusters.

The Density-Based Clustering Validation (DBCV) method takes into account both the density and shape characteristics of clusters. The concept of all-points-core-distance (aptscoredist) is introduced, which represents the inverse of the density of each point with respect to all other points in its cluster. Using aptscoredist, a symmetric reachability distance is defined, which is then used to construct a Minimum Spanning Tree (MST) within each cluster. The MST effectively captures both the shape and density of the cluster, as it is built on the transformed space of symmetric reachability distances. Using these MSTs, DBCV identifies the lowest density area within each cluster and the highest density region between clusters.

Consider a dataset $O = \{o_1, \ldots, o_n\}$ containing n objects in the $\mathbb{R}^d$ feature space. Let Dist be an $n \times n$ matrix of pairwise distances $d(o_p, o_q)$, where $o_p, o_q \in O$, for a given distance metric $d(\cdot, \cdot)$. Let $\text{KNN}(o, i)$ represent the distance between object o and its i-th nearest neighbor. Let $C = \{C_i\}$, $1 \leq i \leq l$, be a clustering solution containing l clusters and a (possibly empty) set of noise objects N, with n_i being the size of the i-th cluster and n_N the number of noise points.

To estimate the density of an object within its cluster, a traditional approach is to use the inverse of the distance required to find K objects within that distance. However, this density estimate relies on the distance to a single point. A more robust estimate, which considers all points within the cluster, gives greater weight to closer objects than to those further away. This approach, similar to Gaussian kernel density estimation, avoids dependence on a single point. Additionally, when defining mutual

reachability distance, the core distance should be comparable to the distances between objects within the cluster and should approximate the distance to the K-th nearest neighbor for a suitably small K.

It is essential to keep in mind the following key points.

- **Core Distance of an Object** The core distance of an object o in a cluster Ci, relative to all other ni - 1 objects in the same cluster, is defined as the inverse of density. This measure is based on the distances to the nearest neighbors and reflects the density of objects. Properties of Core Distance:

  1. The measure gives more weight to closer objects because it uses the inverse of KNN distances raised to the power of dimensionality.
  2. The core distance of an object lies between the second and the last nearest neighbor distances. This ensures that the measure is neither less than the second nearest neighbor distance nor more than the last one.
  3. In a uniform distribution of objects on a high-dimensional hypersphere, the core distance of an object at the center is approximately proportional to the logarithm of the total number of objects divided by the dimensionality of the space.

- **The Mutual Reachability Distance** between two entities o_i and o_j in the set O is defined as

$$d_{\text{reach}}(o_i, o_j) = \max\left\{\text{core\_distance}(o_i), \text{core\_distance}(o_j), d(o_i, o_j)\right\}. \tag{1}$$

  The Mutual Reachability Distance between two entities measures the maximum of three values: the core distances of each entity and the direct distance between them. It captures the greatest extent of separation or density influence between the two entities

- **Mutual Reachability Distance Graph** is a fully connected graph where the vertices represent entities in O, and the weight of each edge corresponds to the reciprocal reachability distance between the connected entities.

- **Mutual Reachability Distance MST** Let O represent a collection of entities and G denote the reciprocal reachability distance graph. The minimum spanning tree (MST) of G is referred to as MST_{RRD}.

- **Density Sparseness of a Cluster(DSC)** The Density Sparseness of a Cluster of a cluster C_i is defined as the highest edge weight among the internal edges in the Minimum Spanning Tree of Reciprocal Reachability Distance (MST RRD) for cluster C_i. Here, the MST_{RRD} is the minimum spanning tree constructed using the core distances of objects within C_i.

- **Density Separation(DSPC)** The Density Separation between two clusters C_i and C_j ($1 \leq i, j \leq l$, with $i \neq j$) is defined as the smallest reachability distance between the internal nodes of the Minimum Spanning Tree of Reciprocal Reachability Distance (MST_{RRD}) for clusters C_i and C_j.

  if a cluster has better density compactness than density separation we obtain positive values of the validity index

- **Validity Index of a Cluster** The validity of a cluster C_i ($1 \leq i \leq l$) is defined as:

$$V_C(C_i) = \frac{\min_{1 \leq j \leq l, j \neq i} \text{DSPC}(C_i, C_j) - \text{DSC}(C_i)}{\max\left(\min_{1 \leq j \leq l, j \neq i} \text{DSPC}(C_i, C_j), \text{DSC}(C_i)\right)} \tag{2}$$

- **Validity Index of a Clustering**

  The quality metric for the clustering solution $C = \{C_i\}$, where $1 \leq i \leq l$, is calculated as the weighted mean of the quality metrics for each cluster in C. This metric produces values between -1 and +1, with higher values reflecting a more effective density-based clustering solution.

$$\text{DBCV}(C) = \sum_{i=1}^{l} \frac{|C_i|}{|O|} \cdot V_C(C_i) \tag{3}$$

  Noise is indirectly addressed through the weighted average that considers both the cluster size ($|C_i|$) and the total number of objects under evaluation denoted as $|O|$

  So, considering the previous point, the overview of the DBCV Methodology is:

  To evaluate a clustering solution using DBCV, we begin by focusing on an individual cluster C_i and its constituent elements. We first determine the core distances for each object within C_i. From these core distances, we calculate the Mutual Reachability Distances (MRDs) between all pairs of objects within the cluster. Using these MRDs, we construct a Minimum Spanning Tree (MST_{MRD}) for the cluster. This process is repeated for every cluster in the dataset, resulting in l separate minimum spanning trees, each corresponding to a cluster.

  The validity of the clustering is then assessed using a density-based index derived from these MSTs. This index incorporates two key measures: density sparseness and density separation. The density sparseness of a cluster is quantified by the maximum edge weight in its MST_{MRD}, representing the region of lowest density within the cluster. In contrast, the density separation between two clusters is measured by the smallest MRD between their objects, indicating the region of highest density separating the clusters. Combining these two metrics, we obtain the DBCV validity index, which assesses the overall clustering quality.

  Briefly, the DBCV metric assesses the mean ratio of the distances from data points to their cluster centers to the distances from data points to the nearest points in different clusters. The concept is that a strong clustering solution should feature tight and distinct clusters, so this ratio should be significant.

4 Controlled experiment

In this section, it is applied the DBCV metric to artificial datasets using the DBSCAN algorithm. The goal is to analyze scenarios where clustering performs well (clearly separated clusters) and where it fails (overlapping clusters). This analysis will allow to evaluate the effectiveness of the DBCV metric in distinguishing between successful and unsuccessful clustering outcomes.

The first step involves generating the artificial datasets. Two different examples will be tested: the first with a dataset containing 300 points, and the second with 10,000 points. This approach allows to evaluate both the reliability and computational performance of the clustering algorithm under different conditions.

The datasets are created using the ‘make_moons‘ function, which produces two interlocking half-circle shapes. Each dataset differs in the amount of noise added, ranging from no noise to significant noise. Here's a description of each dataset:

1. X1: A perfectly shaped "two-moon" dataset with no noise, resulting in two clearly separated and well-defined clusters.

2. X2: The "two-moon" dataset with a small amount of noise (0.056), slightly blurring the boundaries between the clusters but still maintaining distinct groups.

3. X3: With moderate noise (0.111), the dataset shows more overlap and less distinction between the clusters, though the general shape is still recognizable.

4. X4: Further increased noise (0.167) leads to more overlapping points, making the clusters less distinct and more challenging to separate.

5. X5: As the noise level reaches 0.222, the two clusters begin to merge more significantly, with points increasingly spread out and overlapping.

6. X6: At a noise level of 0.278, the clusters are even more intertwined, making it difficult to clearly distinguish between the two groups.

7. X7: With noise at 0.333, the two clusters start to lose their original structure, with substantial overlap between the points.

8. X8: The noise level of 0.389 results in clusters that are highly mixed, with the original "two-moon" shape becoming barely recognizable.

9. X9: At a noise level of 0.444, the two clusters are almost entirely merged, with little distinction between them.

10. X10: The highest noise level (0.5) results in a dataset where the two clusters are almost completely indistinguishable, with the points distributed in a way that no longer resembles the original "two-moon" shape.

Each dataset progressively demonstrates the impact of increasing noise on the separation and clarity of the clusters, making them suitable for evaluating the performance of clustering algorithms.

The second step is to determine the optimal parameters for the DBSCAN model. We consider different values for ''min_samples'' (1, 5, 10, 50) and ''cluster_selection_epsilon'' (0.01, 0.1, 0.2, 0.5). The best results are achieved with 'min_samples = 10 and 'cluster_selection_epsilon = 0.01. With all the necessary components in place, the DBCV value is calculated using four different methods. For each method, the results will be presented alongside graphical representations to illustrate the outcomes. In the presence of a single image, note that the experiment with 10,000 points was not conducted due to the computational time required for the experiment with 300 points.

4.1 DBCV

Considering the metric proposed in [14], which provides a framework for implementing Density-Based Clustering Validation in Python.

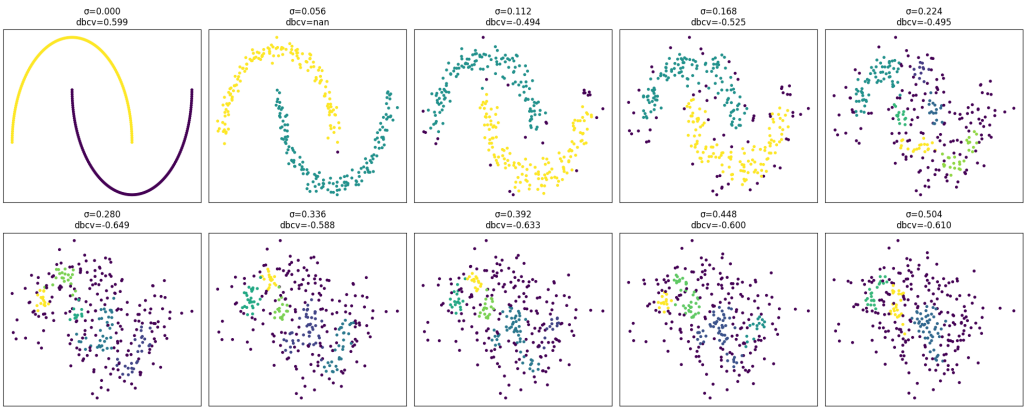

Figure 1: DBCV with 300 points. Execution time 93.43 seconds

4.2 Fast Density-Based Clustering Validation

Considering the metric proposed in [15], which provides an implementation for Python, with support for parallel and dynamically-adjustable high precision computation.

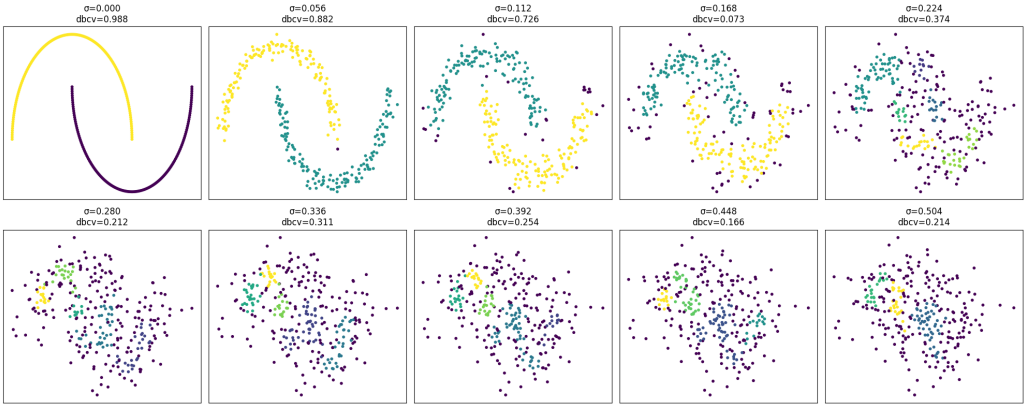

Figure 2: Fast DBCV with 300 points. Execution time 1.83 seconds

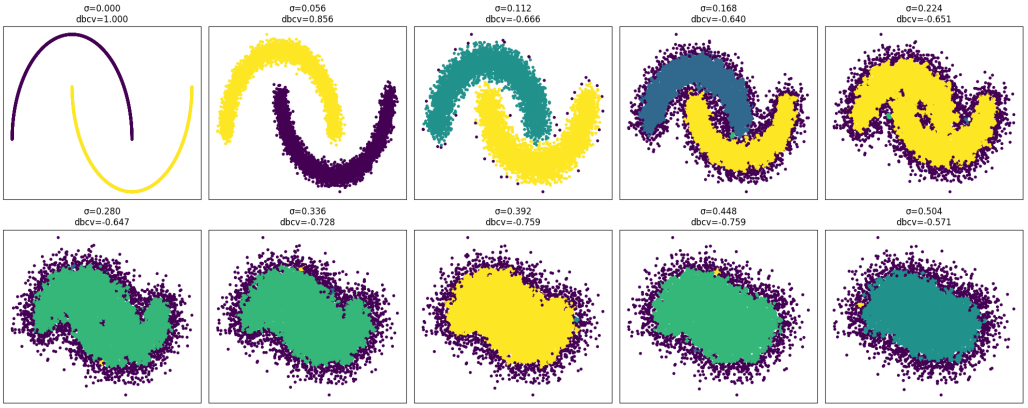

Figure 3: DBCV with 10000 points. Execution time 445.65 seconds

4.3 hdbscan validity_index

This relative validity is computed using the mutual-reachability minimum spanning tree rather than the all-points minimum spanning tree. As a result, the score may not serve as an absolute measure of clustering quality, but rather as a relative metric.[16]

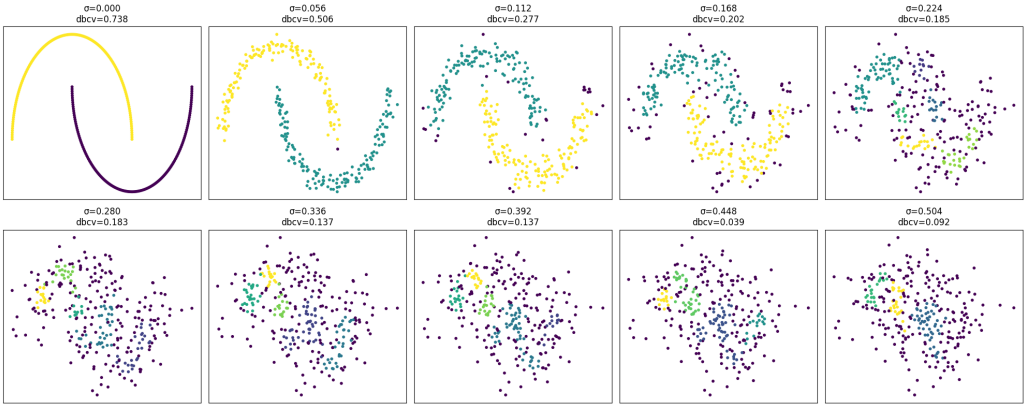

Figure 4: hdbscan validity_index with 300 points. Execution time 2.13 seconds

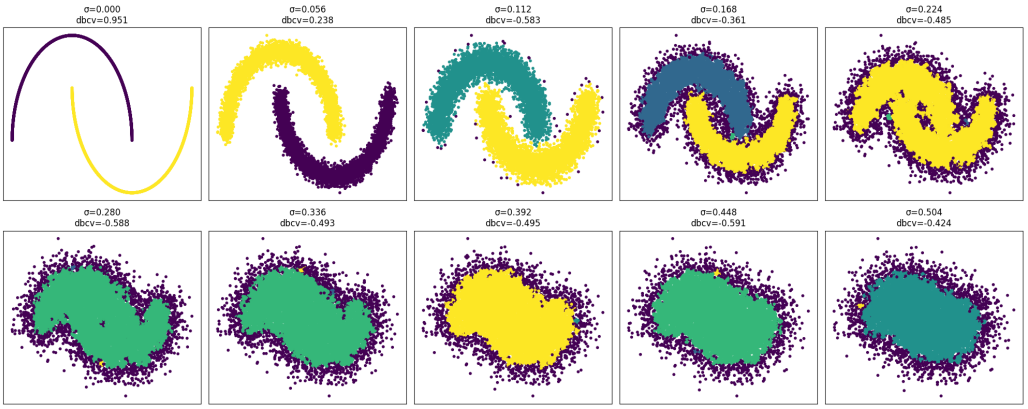

Figure 5: hdbscan validity_index with 10000 points. Execution time 176.28 seconds

4.4 Mathematical formula

Using the following formula:

procedure CALCULATE_DBCV_MATH(X, labels)
  $n \leftarrow$ length of X
  $unique\_labels \leftarrow$ unique elements in $labels$
  if length of $unique\_labels < 2$ **then**
    return -1
  end if
  $distances \leftarrow$ matrix of Euclidean distances between all points in X
  $dbcv\_sum \leftarrow 0$
  for each point i from 1 to n **do**
    $internal\_sum \leftarrow 0$
    for each point j from 1 to n **do**
      if $i \neq j$ **then**
        $max\_dik \leftarrow$ maximum distance between point i and all other points except j
        $internal\_sum \leftarrow internal\_sum + \frac{\text{distance between } i \text{ and } j}{max\_dik}$
      end if
    end for
    $dbcv\_sum \leftarrow dbcv\_sum + \frac{internal\_sum}{n}$
  end for
  $dbcv \leftarrow \frac{dbcv\_sum}{n}$
  return $dbcv$
end procedure

Figure 6: Mathematical formula with 300 points. Execution time 81.23 seconds

4.5 Compare metrics

Based on previous experiments, Fast DBCV has proven to be the most effective package in terms of performance, speed, and alignment with the experimental objectives. For this reason, it will be employed, using the previously selected parameters (min_samples: 10, cluster selection epsilon = 0.01), to perform a comparison with other metrics commonly used in the field of clustering, such as silhouette, Dunn, Davies-Bouldin, and Calinski-Harabasz

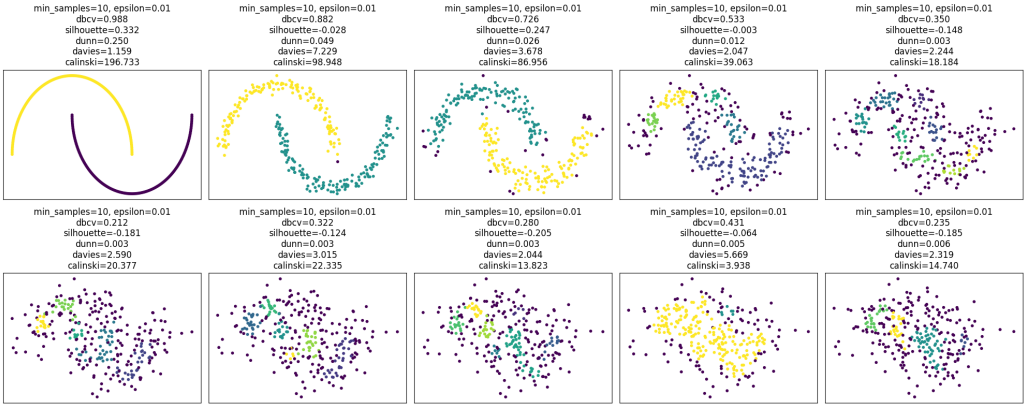

Figure 7: compare_metrics with the dataset with 300 points

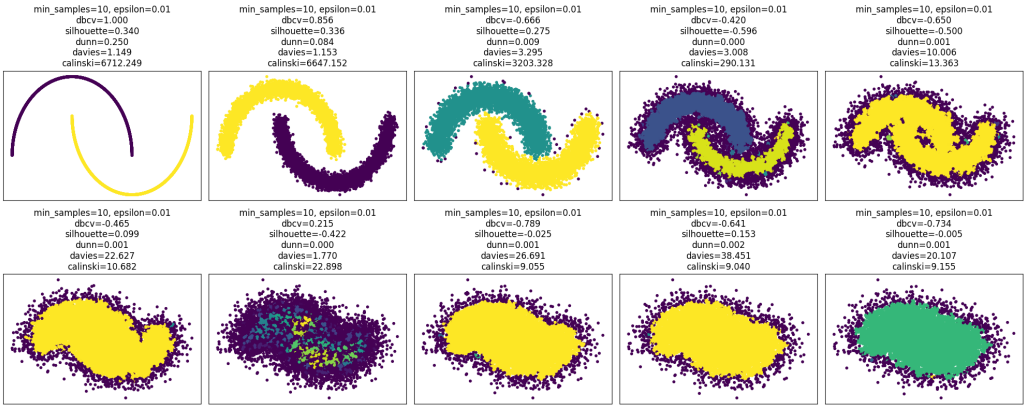

Figure 8: compare_metrics with the dataset with 10000 points

The evaluation of the two dataset is very similar. In panel 1, DBCV assigns a very high value, indicating strong cluster separation and cohesion. The Dunn score, which measures the ratio between minimum inter-cluster distance and maximum intra-cluster distance, is also relatively high. The Calinski-Harabasz score, assessing the ratio between inter-cluster and intra-cluster variance, is very high as well, suggesting well-separated clusters. However, the silhouette and Davies-Bouldin metrics indicate moderate clustering quality, possibly due to complex cluster shapes or variable densities.

In panel 4, DBCV and Calinski-Harabasz reflect moderate clustering quality, with DBCV taking into account the internal density of clusters. The silhouette and Dunn scores are extremely low, suggesting that the clusters are very close or overlapping. Despite their complex shapes or distributions, the low Davies-Bouldin score indicates that the clusters are reasonably distinct.

In panel 6, all metrics point to poor clustering quality. DBCV identifies mediocre density and separation, while silhouette and Dunn suggest significant overlap or poorly defined clusters. The low Calinski-Harabasz and high Davies-Bouldin scores further confirm poor separation between clusters.

In panel 11, all metrics concur on the low clustering quality. DBCV, silhouette, and Dunn clearly indicate poorly defined and likely overlapping clusters. The Davies-Bouldin and Calinski-Harabasz scores also suggest poor separation and significant internal variance.

The DBCV metric proves to be sensitive in evaluating clustering quality, especially when there are differences in cluster density and separation. However, it tends to provide more optimistic values compared to silhouette and Dunn, which are much stricter in the presence of overlapping or poorly defined clusters.THis because DBCV seems to focus more on density and internal cohesion, which may explain its higher scores in scenarios where other metrics indicate poor separation or overlapping clusters. Davies-Bouldin and Calinski-Harabasz offer a complementary view, assessing the separation

and intra-cluster variance, respectively. Overall, DBCV is particularly useful when cluster density is a critical factor.

5 Real experiment

Based on the work done so far, the objective is to test this metric on real-world data, specifically using 5 datasets derived from electronic health records (EHRs). For each dataset, a brief exploratory analysis of the main data will be conducted, describing the columns that make up the dataset. Each dataset undergoes PCA (Principal Component Analysis) to reduce its dimensionality to two dimensions. Subsequently, parameter validation is performed to select the settings that yield the highest DBCV (Density-Based Cluster Validity) score using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) as the clustering method. The parameters validated include: `min_cluster_size_range = [3, 5, 10, 15, 20, 30]` and `cluster_selection_epsilon_range = [0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9]`. Using the optimal parameters identified and the PCA-reduced dataset, clustering is performed with HDBSCAN, and the results are then plotted.

5.1 Neuroblastoma

The provided dataset appears to be a retrospective study of patients with neuroblastoma, a type of childhood cancer. Below is an explanation of the dataset columns and what each represents:

- **age**: Indicates the patient's age at the time of diagnosis.
- **sex**: Represents the patient's gender.
- **site**: Likely denotes the primary anatomical site of the tumor, encoded numerically for different locations.
- **stage**: The stage of neuroblastoma at diagnosis. The stage of a tumor reflects the severity and extent of the disease.
- **risk**: This column indicates the risk level associated with the patient's disease.
- **time_months**: Represents the duration of follow-up or the time elapsed until an event, such as relapse or death.
- **autologous_stem_cell_transplantation**: Indicates whether the patient received an autologous stem cell transplant, a form of treatment.
- **radiation**: This column indicates whether the patient received radiation therapy as part of their treatment.
- **degree_of_differentiation**: Represents the tumor's degree of differentiation, indicating how closely the cancer cells resemble normal cells. Lower values suggest a less differentiated and more aggressive tumor.
- **UH_or_FH**: Refers to the status of "Unfavorable Histology" (UH) or "Favorable Histology" (FH), which are categories used to classify tumor tissue type based on prognosis.
- **MYCN_status**: MYCN is an oncogene, and when amplified, it is associated with a worse prognosis in neuroblastoma.
- **surgical_methods**: Indicates the type or extent of surgery performed.
- **outcome**: This column represents the patient's outcome.

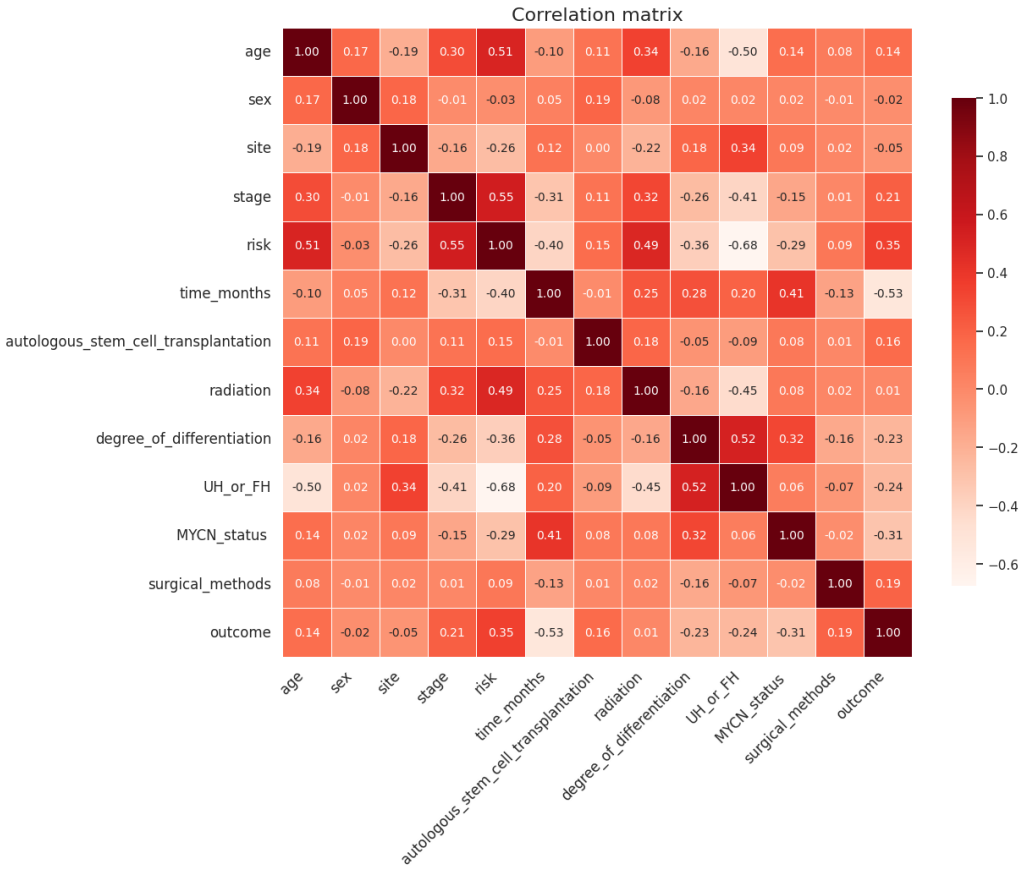

Figure 9: Correlation Matrix of Neuroblastoma's dataset's columns

The correlation matrix shows that there are significant correlations between some variables. PCA is useful for combining these correlated variables into new principal components, reducing dimensionality while preserving as much variance as possible. This step is helpful before applying any clustering model, as it reduces the risk of high correlations between variables distorting the clustering results. The correlation matrix is useful for PCA because this dimensionality reduction method works by identifying linear combinations of variables that capture most of the variability in the data. If there are strong correlations between certain variables, they might contribute significantly to the first principal components.

High correlations between variables such as risk and UH_or_FH (-0.68), risk and radiation (0.49), or outcome and time_months (-0.53) suggest that there is some informational redundancy. PCA will reduce dimensionality by combining these correlated variables into principal components that retain most of the original information, but in a reduced space.

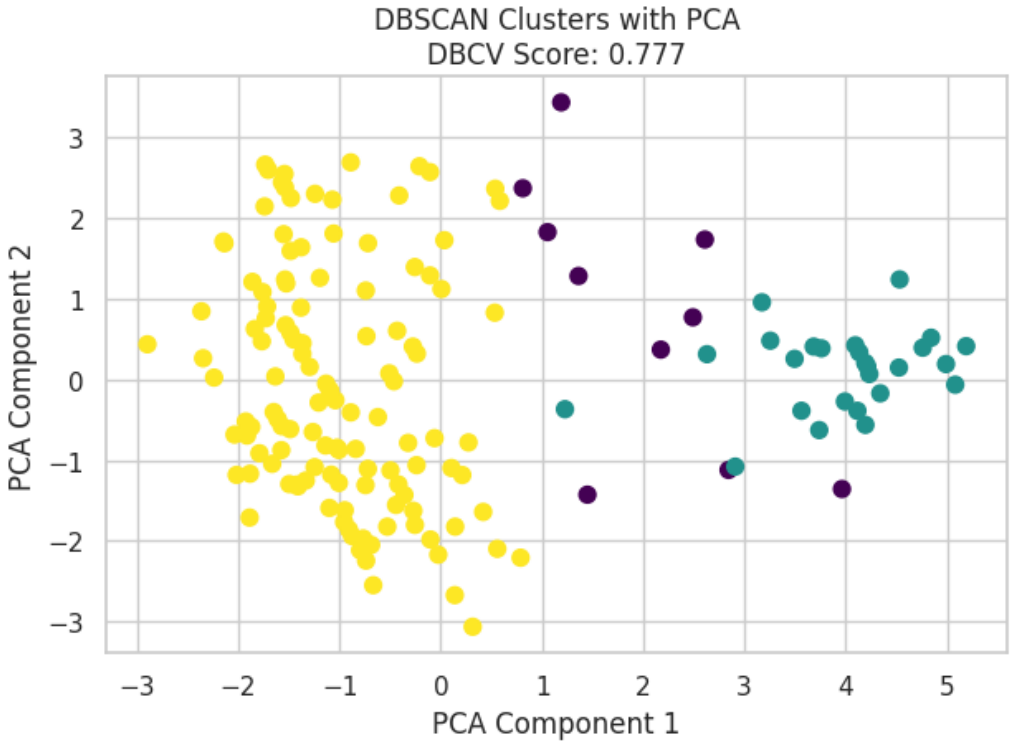

Figure 10: Clustering Plot of Neuroblastoma

A positive DBCV value, such as in this case (0.777), indicates that the identified clusters are coherent and well-separated from the noise present in the data. Three main clusters can be distinguished (yellow, green, purple). The clusters are fairly distinct from each other, with significant gaps between them, suggesting good separation.

The points colored in purple represent noise points. This aligns with the idea that the DBSCAN/HDBSCAN algorithm, in addition to identifying clusters, can distinguish between clusters and noise.

The yellow and green clusters show good density, meaning that the points within each cluster are close to each other, reflecting high internal cohesion. This is consistent with the relatively high DBCV, which rewards cluster compactness. The dispersion of points in the purple cluster contributes to a less than perfect DBCV (0.777 instead of a value closer to 1), as it reduces the average density and increases ambiguity in the clustering.

5.2 Diabetes

The dataset appears to contain data related to patients with type 1 diabetes. Below is an explanation of the columns:

- **age**: The patient's age in years.
- **duration.of.diabetes**: The duration of diabetes in years, i.e., the time since the diagnosis of type 1 diabetes.
- **body_mass_index**: The patient's body mass index (BMI).
- **TDD**: The total daily dose of insulin.
- **basal**: The amount of basal insulin, which is the long-acting insulin dose that the patient takes to maintain stable glucose levels between meals and overnight.

- **bolus**: The amount of bolus insulin, which is the short-acting insulin dose taken to cover carbohydrate intake during meals or to correct high blood glucose levels.
- **HbA1c**: The level of glycated hemoglobin, an indicator of long-term blood glucose control, expressed as a percentage.
- **eGFR**: The estimated glomerular filtration rate, an indicator of kidney function, expressed in ml/min/1.73 m^2.
- **perc.body.fat**: The patient's body fat percentage, representing the proportion of fat relative to total body weight.
- **adiponectin**: The level of adiponectin, a hormone produced by adipose tissue involved in glucose regulation and fatty acid breakdown.
- **free.testosterone**: The level of free testosterone in the blood, which can be important for assessing endocrine function, especially in patients with diabetes.
- **SMI**: The skeletal muscle index, which measures the amount of muscle mass relative to height or weight.
- **grip.strength**: The patient's hand grip strength, an indicator of overall muscle strength.
- **knee.extension.strength**: The strength of knee extension, another measure of muscle strength, particularly in the legs.
- **gait.speed**: The walking speed, likely expressed in meters per second, which can be an indicator of mobility and overall physical function.
- **ucOC**: The level of undercarboxylated osteocalcin, a bone protein that can be an indicator of bone turnover.
- **OC**: The total level of osteocalcin, a marker of bone metabolism.
- **weight_kg**: The patient's weight in kilograms.
- **insulin_regimen_binary**: A binary variable indicating the type of insulin regimen used by the patient.
- **sex_0man_1woman**: A binary variable indicating the patient's sex, where 0 represents a man and 1 represents a woman.

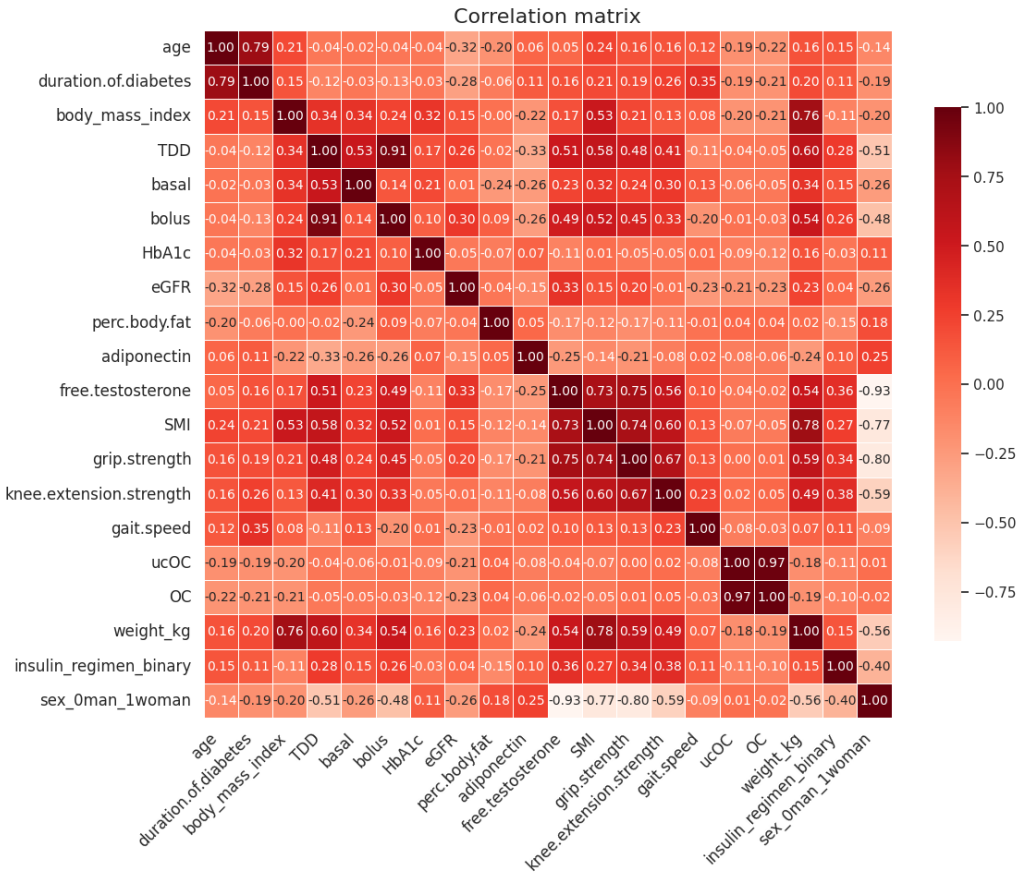

Figure 11: Correlation Matrix of Diabetes's dataset's columns

For this dataset, Basal and TDD have a strongly positive correlation (r = 0.91): These two variables are almost perfectly correlated, suggesting that they represent very similar information. During PCA, it is likely that these two variables will heavily contribute to a single principal component. PCA might reduce these two dimensions into just one, given that they carry redundant information. SMI (Skeletal Muscle Index) and Grip Strength (r = 0.75) are also strongly correlated, reflecting the close relationship between muscle mass index and grip strength. Knee Extension Strength and Sex (r = -0.59) show a strong negative correlation, indicating significant differences between sexes in this measure. During PCA, this could result in a principal component that strongly differentiates between men and women based on physical strength. SMI and Sex (r = -0.77) also show a strong negative correlation. This indicates that sex might play an important role in defining variability patterns in the dataset, with PCA likely capturing this difference in one of the first principal components. Strongly correlated variables (such as Basal and TDD, or SMI and Grip Strength) will likely be represented by the same principal component, significantly reducing dimensionality. The high correlation between certain pairs of variables indicates information redundancy that PCA can exploit to reduce the number of dimensions without losing significant information. For instance, Basal and TDD could be summarized into a single axis in PCA, reducing model complexity without sacrificing accuracy.

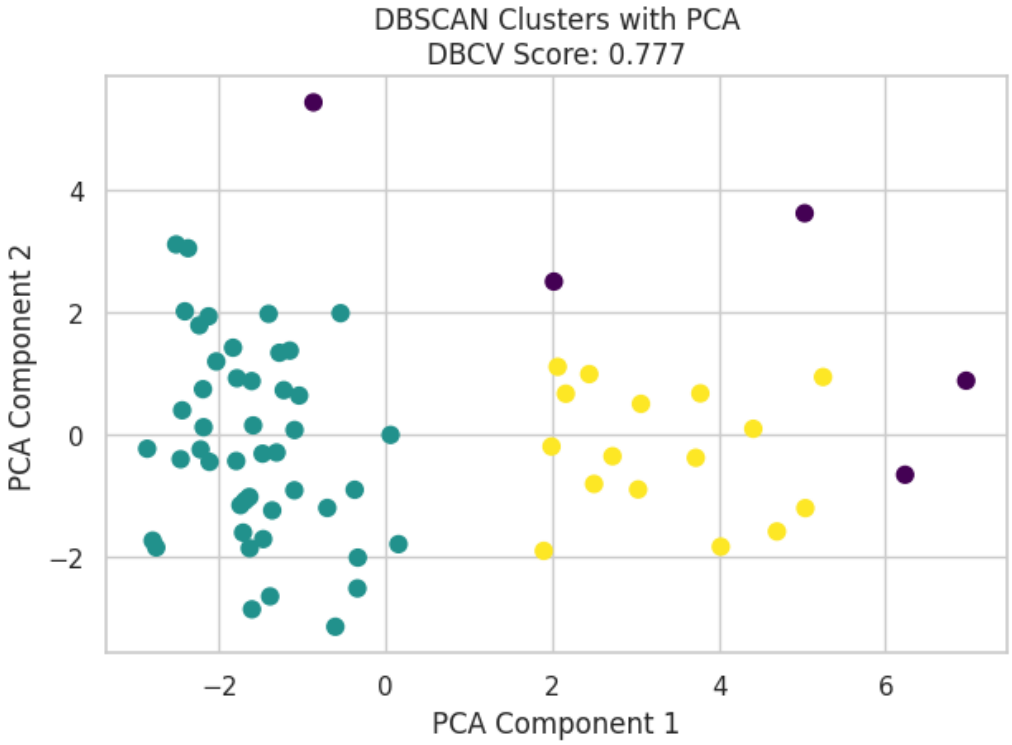

Figure 12: Clustering Plot of Diabetes

The DBCV value of 0.777 indicates that the clusters are well-defined and separated, with most data points tightly grouped within dense clusters. A value of **0.777** suggests a good cluster structure, indicating that the HDBSCAN algorithm has effectively identified the dense regions of data that form well-defined clusters, despite the presence of noise. The identified clusters have relatively good quality, and the separation between clusters is clear, with the algorithm handling the noise effectively.

5.3 Sepsis

The dataset appears to pertain to patients with sepsis or SIRS (Systemic Inflammatory Response Syndrome). Below is an explanation of the columns:

- **Age**: The patient's age in years.
- **sex_woman**: A binary variable indicating the patient's sex, where 0 represents a man and 1 represents a woman.
- **diagnosis_0EC_1M_2_AC**: A coded variable for the patient's diagnosis, where:
  - 0 indicates "EC",
  - 1 indicates "M",
  - 2 indicates "AC".
- **APACHE II**: The APACHE II score (Acute Physiology and Chronic Health Evaluation II), a severity-of-disease classification system used to assess the severity of critical illness and predict mortality.
- **SOFA**: The SOFA score (Sequential Organ Failure Assessment), an indicator of organ dysfunction and mortality risk in critically ill patients.
- **CRP**: The C-Reactive Protein (CRP) level in the blood, a marker of inflammation.

- **WBCC**: The White Blood Cell Count (WBCC), measuring the total number of white blood cells per microliter of blood.
- **NeuC**: The Neutrophil Count, a subtype of white blood cells responsible for the primary immune response.
- **LymC**: The Lymphocyte Count, a type of white blood cell important for adaptive immune response.
- **EOC**: The Eosinophil Count, a type of white blood cell involved in allergic reactions and parasitic infections.
- **NLCR**: The Neutrophil-to-Lymphocyte Count Ratio, an inflammatory index used to assess the inflammatory response and mortality risk.
- **PLTC**: The Platelet Count, an indicator of the blood's ability to clot.
- **MPV**: The Mean Platelet Volume, an index of the average size of platelets, which may indicate platelet activation.
- **Group**: A variable that may indicate the study group or category to which the patient belongs.
- **LOS-ICU**: The Length of Stay in the Intensive Care Unit (ICU), expressed in days.
- **Mortality**: A binary variable indicating the mortality outcome during the study period, where 0 represents survival and 1 represents death.

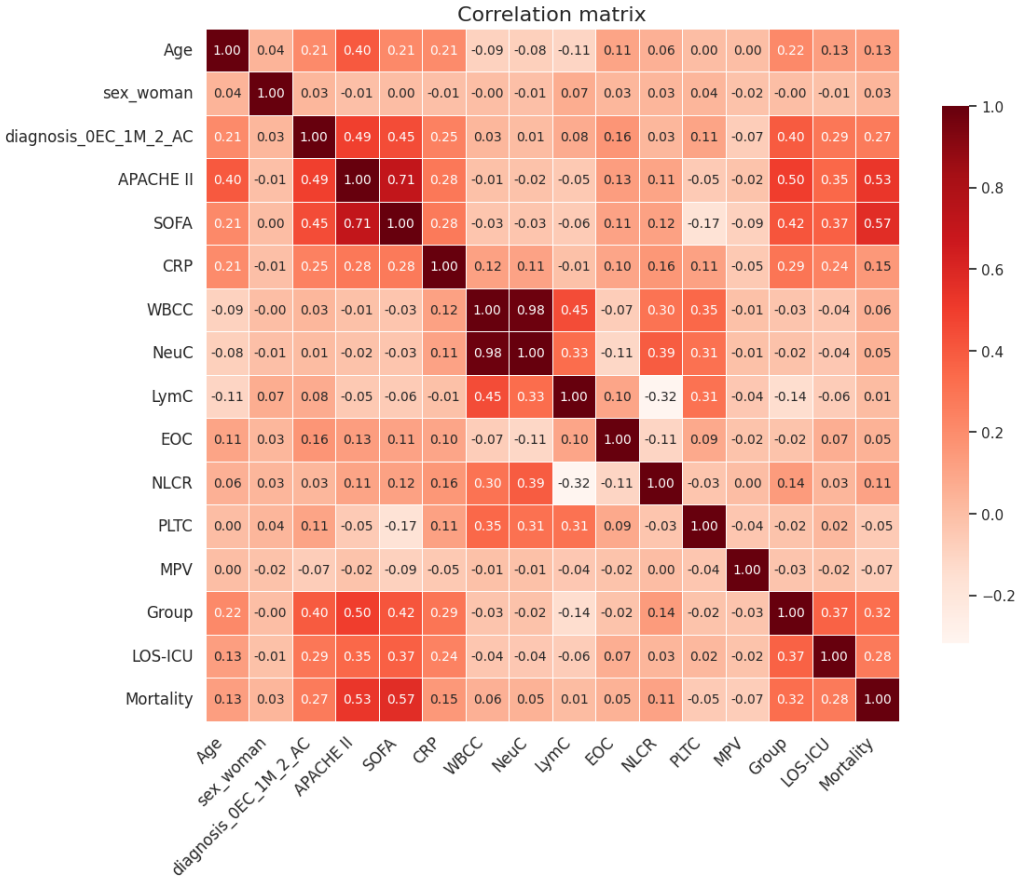

Figure 13: Correlation Matrix of Sepsis's dataset's columns

For this dataset, WBCC (White Blood Cell Count) and NeuC (Neutrophil Count) are almost perfectly correlated (r = 0.98), suggesting that the total white blood cell count and the neutrophil count are closely linked. In PCA, these variables are likely to be combined into a single principal component, reducing the dataset's dimensionality without significant loss of information.

The APACHE II score and the SOFA score show a strong correlation, reflecting the fact that both are used to assess the severity of conditions in intensive care patients. During PCA, it is likely that these two scores will heavily contribute to the same principal component.

The correlation between group and the APACHE II score is significant (0.5). PCA might capture this relationship in a principal component that represents an axis of clinical severity or risk associated with patient groups.

There is a significant correlation (0.57) between the SOFA score and mortality, suggesting that patients with higher SOFA scores tend to have higher mortality rates. This relationship is likely to be captured in one of the first principal components in PCA.

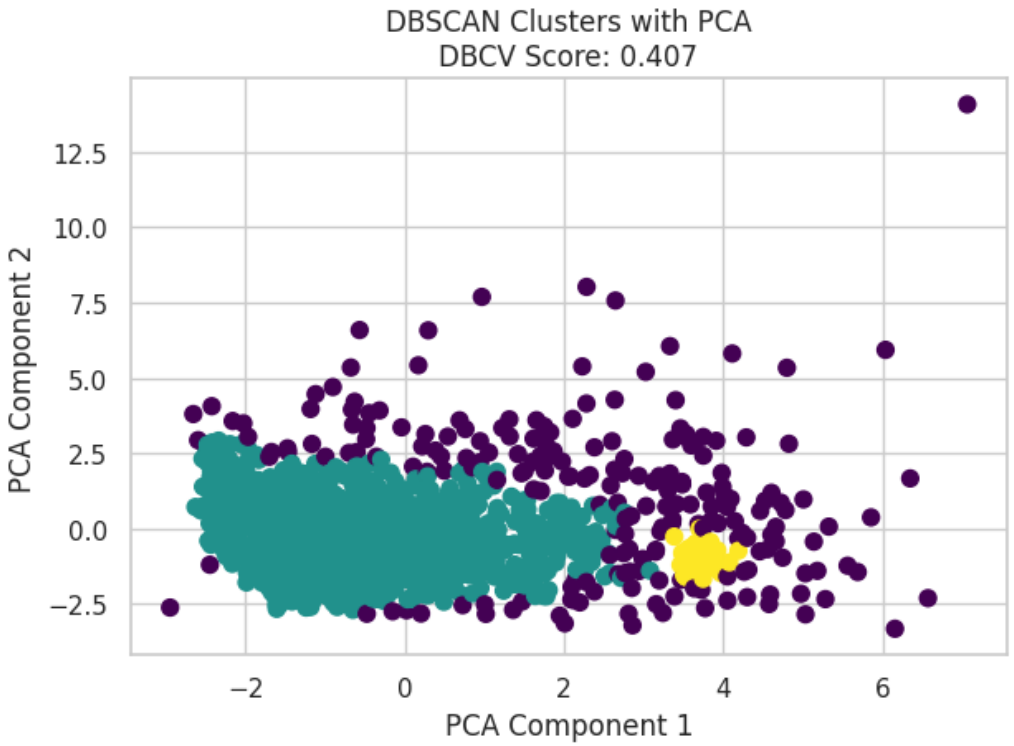

Figure 14: Clustering Plot of Sepsi

The DBCV value of 0.407 indicates a less well-defined and potentially more problematic cluster structure. A DBCV of 0.407 suggests that while clusters are present, they are not as well-defined or separated as in other scenarios. There is greater uncertainty in the separation of clusters, with a higher proportion of points that may be ambiguously classified.

The significant number of noise points might indicate that a substantial portion of the dataset has not been well captured by the identified clusters.

Compared to the previous graph with a DBCV of 0.777, the clustering quality here is lower. The identified clusters are not as distinct. The proximity of points belonging to different clusters and the presence of a large amount of noise suggest that the data may be more complex to separate.

5.4 Heart Failure

The dataset appears to pertain to patients with heart failure and depression. Below is an explanation of the columns:

- **id**: A unique identifier for each patient.
- **Age (years)**: The patient's age in years.
- **Male (1=Yes, 0=No)**: A binary variable indicating the patient's sex, where 1 represents a male and 0 represents a female.
- **PHQ-9**: The Patient Health Questionnaire-9 score, a screening tool used to measure the severity of depression. Higher values indicate more severe depression.
- **Systolic BP (mm Hg)**: The patient's systolic blood pressure, measured in millimeters of mercury (mm Hg). This represents the pressure in the arteries when the heart contracts.
- **Estimated glomerular filtration rate**: The estimated glomerular filtration rate (eGFR), an indicator of kidney function.
- **Ejection fraction (%)**: The ejection fraction, measuring the percentage of blood the left ventricle pumps out of the heart with each contraction. It is an indicator of cardiac function.
- **Serum sodium (mmol/l)**: The level of sodium in the blood, measured in millimoles per liter (mmol/l).
- **Blood urea nitrogen (mg/dl)**: The level of blood urea nitrogen, measured in milligrams per deciliter (mg/dl).
- **Etiology HF (1=Yes, 0=No)**: A binary variable indicating whether the patient's heart failure has a specific cause (1=Yes) or not (0=No).
- **Prior diabetes mellitus**: A binary variable indicating whether the patient has a history of diabetes mellitus (1=Yes) or not (0=No).
- **Elevated level of BNP/NT-BNP (1=Yes, 0=No)**: A binary variable indicating whether the levels of BNP (Brain Natriuretic Peptide) or NT-proBNP (N-Terminal pro B-type Natriuretic Peptide) are elevated.
- **Time from HF to Death (days)**: The time in days from the diagnosis of heart failure to the patient's death.
- **Death (1=Yes, 0=No)**: A binary variable indicating whether the patient died (1=Yes) or survived (0=No) during the study period.
- **Time from HF to hospitalization (days)**: The time in days from the diagnosis of heart failure to subsequent hospitalization.
- **Hospitalized (1=Yes, 0=No)**: A binary variable indicating whether the patient was hospitalized (1=Yes) or not (0=No) during the study period.

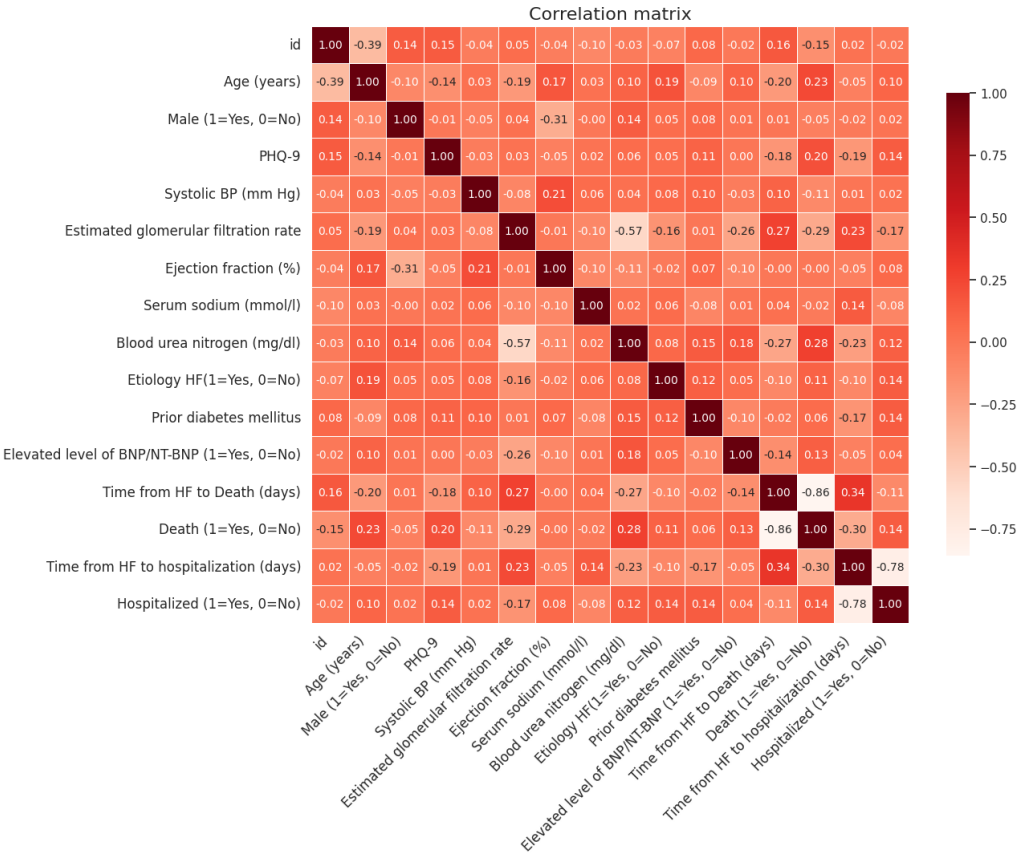

Figure 15: Correlation Matrix of Heart failures's dataset's columns

Death and Time from HF to Death show a strong negative correlation (r = -0.86). This indicates that as the time from heart failure to death increases, the likelihood of death decreases, which is intuitive.

Estimated glomerular filtration rate (eGFR) and Blood Urea Nitrogen (BUN) show a moderately negative correlation (r = -0.57). This suggests that reduced cardiac function is associated with worsening renal function. PCA might capture this relationship in one of the principal components, highlighting the link between cardiac and renal function.

The correlation between elevated Blood Urea Nitrogen levels and mortality is relatively weak but positive (r = 0.29), indicating that an increase in these markers is associated with a higher likelihood of death. PCA might consider this variable in a principal component that represents cardiovascular risks.

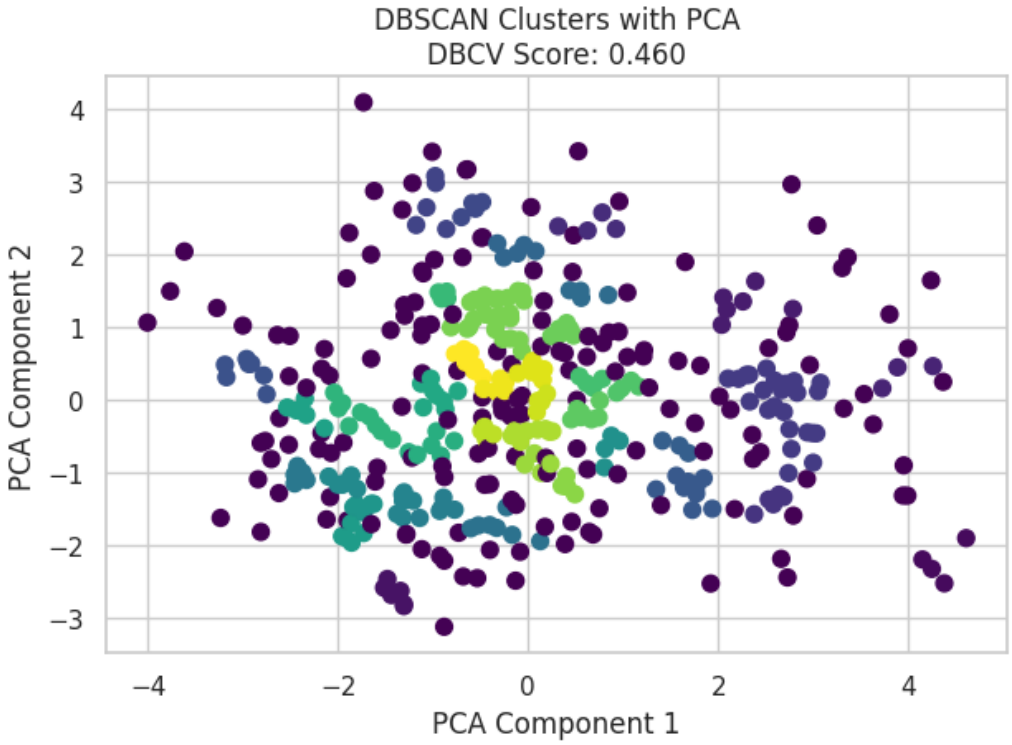

Figure 16: Clustering Plot of Heart failure

The DBCV value of 0.460 suggests that clusters are present and can be identified, but they are not particularly well-separated. This indicates a moderate effectiveness in data clustering, with some confusion or overlap between the clusters. The presence of noise points scattered across the graph and the relatively compact nature of the clusters suggest that the data does not separate clearly. There is likely significant overlap between the clusters, or the data may not follow well-defined density structures. Compared to the DBCV of 0.407, this value of 0.460 represents an improvement, but still indicates that the clusters are not well distinguished. The clusters appear more intertwined and less defined, which could suggest the presence of complex subgroups or intricate structures within the data.

5.5 Cardiac Arrest

The dataset appears to contain information on patients with cardiac arrest in Spain. Below is an explanation of the columns:

- **Exitus**: A binary variable indicating the patient's outcome.
- **sex_woman**: A binary variable indicating the patient's sex.
- **Age_years**: The patient's age in years.
- **Endotracheal_intubation**: A binary variable indicating whether the patient underwent endotracheal intubation during the cardiac arrest.
- **Functional_status**: A score or category indicating the patient's functional status prior to the cardiac arrest. This may reflect the patient's ability to perform daily activities or the degree of disability.
- **Asystole**: A binary variable indicating whether the cardiac rhythm at the time of arrest was asystole.

- **Cardiac_arrest_at_home**: A binary variable indicating whether the cardiac arrest occurred at the patient's home.
- **Bystander**: A binary variable indicating whether a bystander was present during the cardiac arrest.
- **Time_min**: The time in minutes from the onset of cardiac arrest to the first intervention or arrival of emergency services.
- **Cardiogenic**: A binary variable indicating whether the cause of the cardiac arrest was cardiogenic.

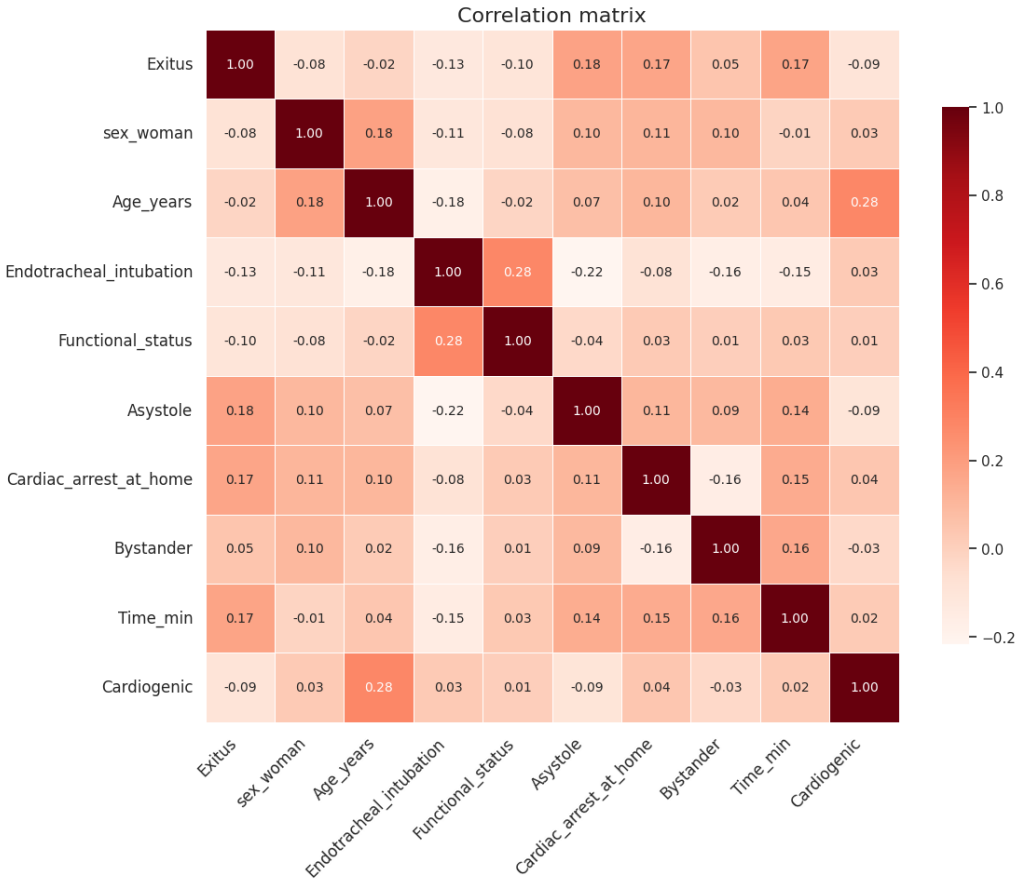

Figure 17: Correlation Matrix of Cardiac Arrests's dataset's columns

Correlazione con "Exitus" (Esito):

The variable "Exitus" has a positive correlation with "Asystole" (0.18) and "Cardiac_arrest_at_home" (0.17). These results suggest that the presence of asystole or a cardiac arrest occurring at home may increase the likelihood of a negative outcome (death).

The correlation with "Endotracheal_intubation" is negative (-0.13), which might indicate that endotracheal intubation was less common among patients with a negative outcome, or it may have been applied at a stage where mortality was already high.

The highest positive correlation is between "Functional_status" and "Endotracheal_intubation" (0.28), indicating a strong association between these two factors.

The variable "Cardiogenic" has a positive correlation with "Age_years" (0.28), suggesting that the cardiogenic cause of the analyzed events increases with age.

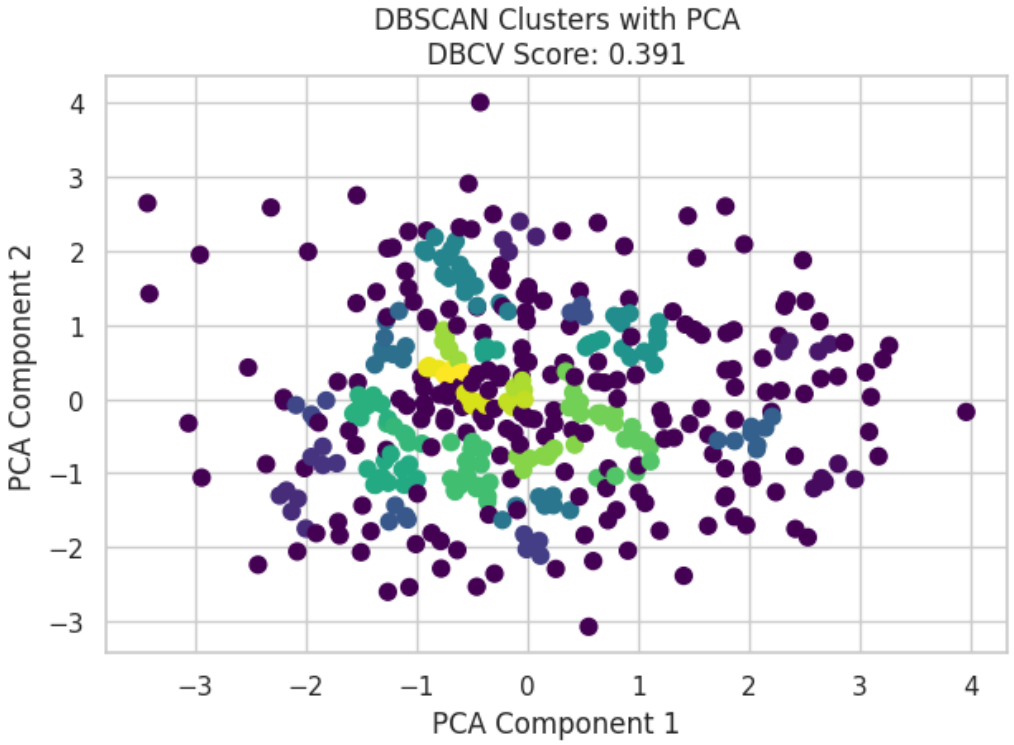

Figure 18: Clustering Plot of Cardiac Arrest

The score of 0.391 is positive, suggesting that the identified clusters have moderate separation and compactness, but are not particularly strong. This value indicates that the clusters are distinguishable, but there are overlaps or the cluster definitions are not perfectly clear.

From the graph, it can be observed that the points are primarily distributed around the center of the two principal components, with some distinct clusters that are not completely separated. The different colors represent the various detected clusters, and there are also points that do not belong to any cluster, considered as noise by the model.

The presence of noise points or overlaps between clusters may contribute to a suboptimal DBCV score. In other words, the cluster structure is not entirely well-defined, which is reflected in a relatively modest DBCV score.

6 Conclusion

In this project, it is explored the efficacy of the DBCV (Density-Based Cluster Validity) metric in evaluating clustering outcomes, particularly in the context of density-based clustering methods like HDBSCAN. Through extensive comparison with other widely-used clustering validation metrics such as silhouette, Dunn, Davies-Bouldin, and Calinski-Harabasz, the DBCV metric demonstrated its sensitivity to variations in cluster density and separation. This characteristic makes DBCV particularly suitable for scenarios where cluster density plays a crucial role in determining clustering quality.

The application of HDBSCAN, combined with PCA for dimensionality reduction, allowed to optimize the clustering process and validate the DBCV metric on real-world data. The parameter validation process highlighted the importance of fine-tuning parameters like minimum cluster size and cluster selection epsilon, as these significantly impact the DBCV score and, consequently, the perceived clustering quality.

Overall, the DBCV metric has proven to be a valuable tool in clustering analysis, offering a nuanced perspective that complements other validation methods. Its focus on density and cohesion provides insights that are particularly relevant in complex clustering scenarios where traditional metrics might

fall short. Future work could involve further testing of DBCV on different types of datasets and clustering methods to better understand its strengths and limitations. Additionally, exploring ways to integrate DBCV with other validation techniques could lead to more comprehensive and accurate clustering evaluations.

References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.

[2] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, *Density-based Clustering Validation*, In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM)*, Philadelphia, PA, 2014.

[3] J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, 1975.

[4] Y. Lu et al., "Subtyping Schizophrenia Using Psychiatric Polygenic Scores," *medRxiv*, 2023.10.12.23296915 (2023).

[5] B. Bozdemir, S. Canard, O. Ermis, H. Möllering, M. Önen, and T. Schneider, "Privacy-preserving density-based clustering," in *ASIACCS*, 2021.

[6] S. Chowdhury and R. Amorim, "An efficient density-based clustering algorithm using reverse nearest neighbour," *Intelligent Computing: Proceedings Of The 2019 Computing Conference*, vol. 2, pp. 29-42, 2019.

[7] E. Werner, J. N. Clark, A. Hepburn, R. S. Bhamber, M. Ambler, C. P. Bourdeaux, C. J. McWilliams, R. Santos-Rodriguez, "Explainable hierarchical clustering for patient subtyping and risk prediction," *Experimental Biology and Medicine*, p. 15353702231214253 (2023).

[8] A. Joshi, H. Li, N. A. Parikh and L. He, "A systematic review of automated methods to perform white matter tract segmentation," *Front. Neurosci.*, vol. 18, p. 1376570, 2024. doi: 10.3389/fnins.2024.1376570

[9] A. Kumar, A. Aggarwal, "An efficient simulated annealing based constrained optimization approach for outlier detection mechanism in RFID-sensor integrated MANET," in *Proceedings Inter. Conf. on Intelligent Systems Design and Applications (ISDA)*, 2018.

[10] S. Zhu, L. Xu, and E. D. Goodman, "Evolutionary multi-objective automatic clustering enhanced with quality metrics and ensemble strategy," *Knowl. Based Syst.*, vol. 188, Jan. 2020, Art. no. 105018.

[11] G. Park, M. Cho, J. Lee, "Leveraging machine learning for automatic topic discovery and forecasting of process mining research: A literature review," *Expert Syst. Appl.*, 2023, p. 122435.

[12] A. Herderich, H. H. Freudenthaler, and D. Garcia, "A Computational Method to Reveal Psychological Constructs from Text Data," *PsyArXiv*, August 9, 2023. https://doi.org/10.31234/osf.io/s64tm.

[13] Y. Liu, "Understanding public perception of societal concerns using social media platforms," (Order No. 30413238). Available from *ProQuest Dissertations & Theses Global*. (2788544790). Retrieved from https://www.proquest.com/dissertations-theses/understanding-public-perception-societal-concerns/docview/2788544790/se-2, 2022.

[14] https://github.com/christopherjenness/DBCV

[15] https://github.com/FelSiq/DBCV

[16] https://hdbscan.readthedocs.io/en/latest/api.html