Revue-IRS

Revue Internationale de la Recherche Scientifique (Revue-IRS)

ISSN: 2958-8413

Vol. 3, No. 3, Juin 2025

Contrasting Classical and Symbolic Approaches to Classification in Data Warehousing Systems: Application to Metabolic Health Profiling

Alain M. KUYUNSA[1], Alidor M. MBAYANDJAMBE[2], Grevi B. NKWIMI[3], Dorotha K. TSHIBOLA[4], Jacques B. TSHINGAMBU[5], Blanchard M. KANGULUMBA[6]

[1,2,3,4]Department of Business Informatics and Business English, Faculty of Economic and Management Sciences, Kinshasa, University of Kinshasa, DR. Congo

[2,5]Department of Mathematics, Statistics and Computer Science, Faculty of Science and Technology, University of Kinshasa, Kinshasa, DR Congo.
[6]Faculty of Science and Technology, Kwango University, Kwango, DR Congo.

Abstract: In the era of Big Data, the diversity and volume of data in enterprise data warehouses pose significant challenges for accurate and robust classification. Data mining encompassing classification, clustering, and association-rule techniques serves as a cornerstone for uncovering actionable insights from these vast repositories. Traditional analytic paradigms, however, often falter when confronted with real-world imperfections, such as multi-valued attributes, measurement imprecision, and aggregated summaries. Symbolic Data Analysis (SDA) addresses these shortcomings by representing entities as complex “symbolic objects” (e.g., intervals, distributions, or sets), thereby preserving inherent variability and uncertainty. A data warehouse based on medical parameters (BMI, blood glucose, etc.) was constructed and analyzed using both approaches. In this work, we first develop a star-schema data warehouse tailored to a representative application domain and implement a robust ETL pipeline to ensure data consistency and integrity. We then apply both classical classification algorithms (e.g., k-means, decision trees, and support vector machines) and a novel symbolic dynamic classification framework where class prototypes are defined as hyperrectangular envelopes of feature intervals to the same datasets. Our evaluation demonstrates that symbolic approaches excel in handling data imprecision and provide richer interpretability, while classical methods remain computationally efficient. The results are validated using accuracy, inertia-based metrics, and clinical interpretability, offering actionable insights for data warehousing applications.

Keywords: Complex data, Tuning, Data warehouse, Symbolic data analysis, Clustering.

Digital Object Identifier (DOI): https://doi.org/10.5281/zenodo.**15657368**

1 Introduction

In the era of Big Data, organizations increasingly rely on data warehousing systems to store, manage, and analyze massive volumes of heterogeneous and complex data. Data mining has emerged as a critical technology for extracting actionable knowledge from such large-scale data repositories. According to Han et al. (2011), data mining encompasses a broad set of techniques such as classification, clustering, and association rule mining, enabling analysts to uncover previously unknown patterns and trends from structured datasets.

Traditional or *classical classification methods*, such as decision trees (Quinlan, 1993), k-nearest neighbors (Cover & Hart, 1967), and support vector machines (Cortes & Vapnik, 1995), have been widely applied in data mining tasks due to their simplicity, interpretability, and effectiveness on well-defined data. These methods assume that each instance in the dataset is represented by a single-valued feature vector, often requiring a high degree of data cleaning and transformation.

However, real-world data are often imperfect, imprecise, and multi-valued, especially in domains involving aggregated, temporal, or symbolic information. In response to these challenges, *symbolic data analysis* (SDA) was introduced as an extension of classical data analysis, capable of handling more complex data types such as intervals, distributions, and categories (Bock & Diday, 2000). Symbolic objects, a core concept in SDA, allow the representation of grouped or summarized entities, facilitating the analysis of aggregated information without loss of essential variability.

The use of SDA in data warehousing systems is particularly relevant for high-level, descriptive knowledge discovery, where one must account for the imprecision and diversity inherent in the summarized data stored within OLAP cubes or multidimensional schemas (Billard & Diday, 2003). Symbolic classifiers such as symbolic k-means or symbolic hierarchical clustering operate on such enriched data types, revealing trends and patterns that may be invisible under classical frameworks.

Recent research has highlighted the advantages of symbolic approaches in domains like customer segmentation, bioinformatics, and temporal data mining (Lechevallier et al., 2001; Brito & Moniz, 2018). Despite their promise, symbolic methods are less mainstream than their classical counterparts, partly due to the complexity of symbolic data representation and a lack of integrated tools within standard data warehousing pipelines.

This paper provides a comparative study of classification methods in the context of data warehousing systems, contrasting classical classification models with symbolic approaches. We begin by designing a data warehouse and conducting data mining tasks using conventional classification algorithms. We then apply symbolic dynamic classification techniques to the same datasets. Our objective is to evaluate the performance, expressiveness, and suitability of both paradigms for knowledge discovery in complex, real-world data environments.

While symbolic methods have been explored in niche domains, their integration into mainstream data warehousing systems remains limited. This study bridges this gap by proposing a scalable symbolic dynamic classification framework tailored for OLAP environments, addressing both theoretical and practical challenges.

2 Literature Review

2.1 Classical Classification Methods in Data Mining

Classical classification methods have long been central to data mining tasks, especially in structured and well-defined environments such as relational databases and traditional data warehouses. These approaches include well-known algorithms such as Decision Trees (Quinlan, 1993), Naïve Bayes, Support Vector Machines (Cortes & Vapnik, 1995), and k-Nearest Neighbors (Cover & Hart, 1967). Their main strength lies in their interpretability, computational efficiency, and availability in most data mining toolkits.

Decision Trees, for instance, have been widely adopted for their ease of use and ability to produce understandable models (Han et al., 2011). However, they often struggle with noisy data or missing values. Support Vector Machines are effective in high-dimensional spaces but require significant tuning of hyperparameters and do not natively handle categorical or symbolic data. The k-Nearest Neighbors algorithm is straightforward but suffers from the curse of dimensionality and sensitivity to irrelevant features.

Classical classification assumes that data instances are described by fixed-size feature vectors containing single numerical or categorical values. This assumption fails to capture the complexity of real-world data, which may involve multi-valued, interval-based, or uncertain information (Diday & Noirhomme-Fraiture, 2008). As a result, such approaches are limited when applied to summarized or aggregated data, as commonly found in OLAP systems.

2.2 Symbolic Data Analysis (SDA) and Symbolic Classification

Symbolic Data Analysis (SDA) emerged in the 1980s as an extension of classical data analysis to handle more complex and structured data types. Introduced by Diday and his collaborators, SDA aims to analyze data described not just by single values but by more complex objects such as intervals, distributions, sets, and even logical formulas (Bock & Diday, 2000).

In the symbolic paradigm, data units called symbolic objects are used to represent aggregated or grouped information. This is particularly suited to data warehousing contexts, where data is often pre-aggregated across dimensions like time, region, or customer segment (Billard & Diday, 2003). For example, rather than a single age value, a customer group might be described by an age interval or by a modal distribution over age categories.

Symbolic classification methods adapt traditional algorithms to work with symbolic objects. Lechevallier et al. (2001) proposed symbolic versions of hierarchical and partitional clustering methods that operate directly on complex data types. Similarly, Brito and Moniz (2018) reviewed clustering methods for symbolic data, showing how symbolic representations enable richer and more realistic modeling, particularly in uncertain or imprecise environments.

The main advantages of symbolic methods include:

- Better handling of heterogeneity and imprecision in data,
- Natural support for aggregated and multivalued information,
- Improved interpretability in decision-support contexts.

However, symbolic methods come with challenges, such as increased algorithmic complexity, fewer available tools, and limited standardization compared to classical data mining.

2.3 Applications in Data Warehousing Systems

Data warehousing systems collect large volumes of transactional data and transform them into multidimensional representations via ETL (Extract, Transform, Load) processes. These multidimensional cubes facilitate OLAP (Online Analytical Processing) but result in data that are often symbolic in nature e.g., summarized by ranges, counts, or distributions.

While classical data mining methods have been applied to data warehouses with some success (Inmon, 2005), symbolic data analysis is particularly well-suited to such contexts. Billard & Diday (2006) demonstrated how SDA can be integrated with OLAP structures to improve the quality of descriptive and predictive modeling. Furthermore, symbolic classifiers can operate on complex features resulting from roll-up and drill-down operations in multidimensional cubes.

Despite the promise of symbolic methods, their integration into mainstream data warehousing and business intelligence workflows remains limited. This is partly due to the lack of mature tools and frameworks, and partly due to the higher expertise required to handle symbolic modeling.

3. Materials and Methods

3.1. Study Design and Data Collection

The dataset used in this study was collected from anonymized medical records of patients admitted to Matete General Reference Hospital between 2020 and 2022. It includes biometric and physiological indicators such as height, weight, heart rate (HR), respiratory rate (RR), cholesterol, and blood glucose. These variables were used to build a multidimensional data warehouse, facilitating classification through both classical and symbolic methods.

Glycemic status was classified into four categories based on fasting glucose values:

- **Hypoglycemia**: < 60 mg/dL
- **Normal**: 60–100 mg/dL
- **Borderline**: 110–126 mg/dL
- **Hyperglycemia**: > 126 mg/dL

Nutritional status was determined by calculating Body Mass Index (BMI = weight/height^2), with categories defined as:

- < 16.5: Famine
- 16.5–18.5: Underweight
- 18.5–25: Normal
- 25–30: Overweight

- 30–35: Moderate Obesity
- 35–40: Severe Obesity
- 40: Morbid Obesity

The objective was to analyze the relationships between glycemic status and nutritional profiles, as well as to identify homogeneous subgroups of individuals using two complementary classification methodologies.

3.2. Methods

3.2.1 Classical Classification Method

Classical classification allowed us to create homogeneous groupings within our individuals based on the observation of descriptors including height, weight, heart rate, respiratory rate, cholesterol, and blood glucose. In classification, it is possible to define a dissimilarity measure d: E x E → IR+ where E is the set of individuals. Various properties are desirable for such a measure, specifically, ∀ (i,j,k)∈E x E x E:

- d(i, i) = 0
- d(i, j) = d(j, i)
- d(i, j) = 0 ⟹ i = j
- d(i, j) ≤ d(i, k) + d(k, j)
- d(i, j) ≤ max (d(i, k), d(k, j))

Depending on the properties verified by d, the terminology differs. Thus, we speak of:

Table 1 : Classical Classification Method

Dissimilarity index	(1)	(2)			
Distance index	(1)	(2)	(3)		
Gap	(1)	(2)		(4)	
Distance	(1)	(2)	(3)	(4)	
Ultrametric gap	(1)	(2)		(4)	(5)
Ultrametric distance	(1)	(2)	(3)	(4)	(5)

The choice of a measure essentially depends on the nature of the descriptors; most often, the choice falls on an appropriate Euclidean distance. Thus, if table X is numerical, we can use the distance M = L. If the set of variables is homogeneous, the inverse distance of variances $M = diag(1/\sigma_j^2)$ if the set is heterogeneous, or the Mahalanobis distance (M = V^-1 where V is the variance-covariance matrix) if we want to sterilize the cloud.

In the following, we will use classical distances that we will denote ‖.‖.

a) Inertia

At the center of the most used methods - having good generalist properties and consistent with factorial analysis methods in particular - we find the notion of inertia. If we have already defined, in the context of PCA, the inertia I of the point cloud, and G its center of inertia, we will now define new notions. If our set of points is grouped into K classes (i.e., K sub-clouds N_k), with weights (M_k = ∑_(i∈N_k) m_i)_1≤k≤K, centers of inertia (G_k)_1≤k≤K, and inertia (I_k)_1≤k≤K, we define:

- The intraclass inertia: I_W = ∑_(k=1)^K I_k
- The interclass inertia: I_B = ∑_(k=1)^K M_k ‖G_k - G‖²

According to Huygens' theorem: I = I_W + I_B

$$I = \sum_{i=1}^{n} m_i \left\|X_i - G\right\|^2 = \sum_{k=1}^{K} \sum_{i \in N_k} m_i \left[\left\|X_i - G_k\right\|^2 + \left\|G_k - G\right\|^2 + 2\left\langle X_i - G_k, G_k - G \right\rangle \right]$$

$$= \sum_{k=1}^{K} I_k + \sum_{k=1}^{K} M_k \left\|G_k - G\right\|^2 = I_w + I_B$$

A usual classification criterion will then be, for fixed K, to minimize the intraclass inertia (i.e., make the classes as homogeneous as possible). This is equivalent to maximizing the interclass inertia (i.e., separating the classes as much as possible).

The quality of a classification can then be evaluated by the ratio I_B/I, interpretable as a proportion of the inertia of the n points explained by their synthesis into K barycenters.

b) Partitioning Algorithms: "Mobile Centers"

Several classification algorithms are geometrically inspired: they are known as partitioning methods, and their principle is to start from an arbitrary partition, improved iteratively until convergence. All require choosing a number of classes a priori.

The best-known of these methods is that of mobile centers, mainly due to Forgy (1965). Its principle is as follows: if we want K classes, we choose K points in the individual space; we then assign each of the n individuals to the one of these K points that is closest to it (in terms of the distance d chosen at the beginning); the K starting points are replaced by the K (or fewer, if a point has not attracted anyone...) barycenters of the individuals assigned to each; then we reiterate the assignment until convergence.

c) Proposition

This algorithm decreases the intraclass variance at each iteration.

Proof

Let N_k^t be the sub-clouds constituted and C_k^t the points to which they are attached (barycenters of N_k^(t-1)), at the t-th iteration.

Consider the following criterion:

$$v(t)=\sum_{k=1}^{K}\sum_{i\in N_k^t} m_i\left\|X_i-C_k^t\right\|^2 .$$

The intraclass inertia is written:

$$I_W(t)=\sum_{k=1}^{K}\sum_{i\in N_k^t} m_i\left\|X_i-C_k^{t+1}\right\|^2 .$$

According to Huygens' theorem, we have:

$$v(t)=I_W(t)+\sum_{k=1}^{K} M_k\left\|C_k^t-C_k^{t+1}\right\|^2 .$$

By minimizing distances for each individual, we also have: $I_W(t)\geq v(t+1)$.

We thus obtain $v(t+1)\leq I_W(t)\leq v(t)$, and $I_W(t+1)\geq I_W(t)$. therefore the intraclass inertia decreases.

This algorithm has been slightly sophisticated in two other methods: k-means (the barycenters are not calculated at the end of assignments, but after each one: the order of appearance of individuals is therefore not neutral), and dynamic clouds (it is no longer a single point that represents a class).

However, these methods have the advantage of converging rapidly towards a local minimum of the intraclass inertia. On the other hand, their two main drawbacks are having to fix K, and especially converging towards a result that depends on the K points initially chosen, often arbitrarily.

3.3. Methodological Framework

To explore the latent structure of the dataset and identify coherent patient profiles, we adopted a hybrid methodological approach combining both classical and symbolic dynamic classification techniques. This two-pronged strategy increases robustness, enabling a deeper analysis of relationships among metabolic indicators.

3.3.1. Classical Classification Approach (Clustering-Based)

We initially applied traditional unsupervised classification methods based on Euclidean and Mahalanobis distances to segment individuals into homogeneous clusters. The following steps were conducted:

- **Data Standardization**: All variables were normalized to ensure comparable scales.
- **Distance Metric Selection**: Depending on variable homogeneity, Euclidean distance or Mahalanobis distance was used.
- **Inertia Analysis**: Total inertia was decomposed into intra-class and inter-class components to evaluate clustering quality.
- **Partitioning Algorithm**: The *Forgy algorithm* (mobile centers) was implemented, initializing K random centroids, followed by iterative assignment and centroid update steps until convergence.
- **Optimization Criterion**: The intraclass inertia IWI_W was minimized, and the proportion of explained inertia IBI\frac{I_B}{I} was used to assess the clustering performance.

This method was useful in identifying general patterns and grouping individuals with similar biometric profiles.

3.3.2 Symbolic Dynamic Classification Method

This is a pure partitioning method, which seeks to improve a partition (in the sense of a criterion W), without changing the number of groups.

The method begins by initializing a random partition into k classes, with k fixed by the user. The iterations are done in two parts: first, the method associates with each class "a representative," called a prototype. This is the representation step. In the case of interval data, the prototype of a class C is the hyperrectangle of gravity L such that $L = \left(\left[\frac{1}{n}\sum_{x_i \in C}\alpha_{\mathrm{i1}}, \frac{1}{n}\sum_{x_i \in C}\beta_{\mathrm{i1}}\right], \ldots, \left[\frac{1}{n}\sum_{x_i \in C}\alpha_{\mathrm{ip}}, \frac{1}{n}\sum_{x_i \in C}\beta_{\mathrm{ip}}\right]\right)$.This method reassigns each individual to the class whose prototype it is closest to; this is the assignment step.

To complement the classical analysis, a symbolic dynamic method was applied to better capture variability in interval and qualitative data. This method involved:

- **Random Initialization**: The dataset was partitioned into *k* symbolic classes.
- **Representation Step**: Each class was represented by a prototype (hyperrectangle of gravity), calculated as the interval enclosing all individuals in the class for each feature.
- **Assignment Step**: Individuals were re-assigned to the class whose prototype was closest according to a symbolic dissimilarity measure.
- **Iteration Until Stability**: The representation and assignment steps were repeated until the partition stabilized.

This approach allowed us to interpret class characteristics in terms of feature intervals and to incorporate uncertainty and variability in physiological measurements.

3.3.3. Comparative Evaluation

To evaluate the coherence and complementarity of the two approaches:

- We compared the clusters obtained with both methods using external validation indices (e.g., silhouette score, Davies-Bouldin index).
- Cross-tabulation of symbolic clusters with classical ones was used to assess convergence and divergence in group structuring.
- Special attention was paid to outlier profiles (e.g., obese subjects with hypoglycemia or normal BMI with hyperglycemia).

3.3.4. Tools and Implementation

The analysis leveraged Microsoft SQL Server 2008's BI stack (SSIS/SSAS/SSRS) for robust data management, implementing optimized star schemas with T-SQL stored procedures for data preparation. For advanced analytics, we utilized Python's scientific stack (NumPy/pandas/scikit-learn) for classical clustering and developed custom symbolic algorithms, with seamless integration through SQL CLR. This hybrid approach combined SQL Server's enterprise-grade data handling (including partition management for performance and MDX cubes for clinical KPIs) with Python's analytical flexibility, while visualization used both matplotlib/seaborn for exploratory analysis and SSRS for production dashboards. Key technical considerations included handling interval-valued data types in the BI environment and optimizing the ETL pipeline for clinical data quality requirements.

4. Results and Discussion

4.1. Construction of the Data Warehouse

In the framework of our study, we were interested in data on patients from the Matete General Reference Hospital with the following parameters: height, weight, heart rate (HR), respiratory rate (RR), cholesterol, and blood glucose.

The data are stored in a data warehouse whose development is presented as follows:

- **Selection of Business Processes to Model**

The business process to model is one of the researched activities. In the health zone activity, decision-makers want to better understand the relationship between subjects' glycemic status and nutritional status. The business process to model is: status (glycemic and nutritional).

- **Grain Declaration**

For patient management, the most granular data are individual transaction lines entered in the form when they are performed. These data consist of age, height, weight, RR, HR, cholesterol, and blood glucose.

- **Choice of Dimensions**

In our study on patient management, we retain only the Location and Time dimensions that we analyze with the patient_management fact.

Table 2 : Location

Attribute	**Data Type**
Id_Location	Numeric
Label	Text

Table 3 : Time:

Attribute	**Data Type**
Id_Time	Numeric
Year	Numeric
Month	Numeric
Quarter	Numeric

Identification of Facts

For our study, the facts collected are ID, height, weight, RR, HR, cholesterol, and blood glucose. The Status Fact table contains the calculable items called measures that we have just mentioned and all the primary keys of all the dimensions cited, which become foreign keys. The description of this Fact table is as follows:

Table 4 : **Identification of Facts**

Attribute	**Data Type**
Id_Status	Numeric
Age	Numeric
HEIGHT	Numeric
HEIGHT m	Numeric
WEIGHT	Numeric
TEMPERATURE	Numeric
RR	Numeric
HR	Numeric
CHOLESTEROL	Numeric
BLOOD GLUCOSE	Numeric

Id_Time	Numeric
Id_Location	Numeric

Star Model of the Data Warehouse

As shown in Figure 1 presents the star model of our data warehouse.

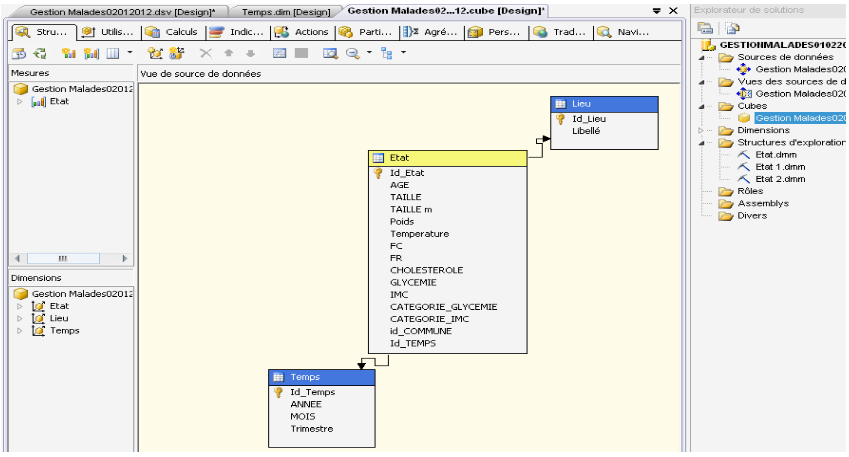
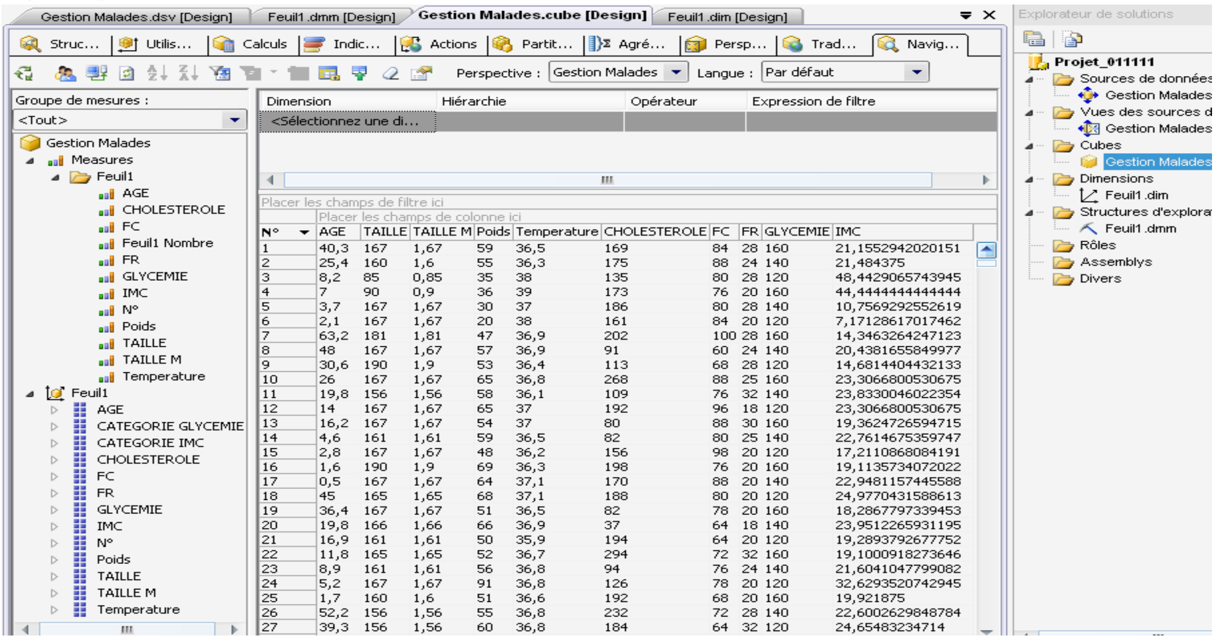


Figure 1: Star model of the data warehouse

Content of the Data Warehouse

Figure 2 presents the content of our data warehouse.



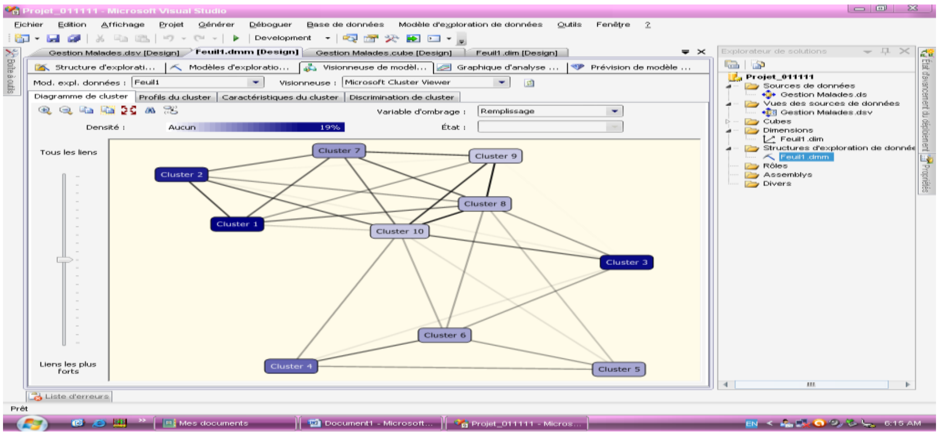
Perspective : Gestion Malades
Langue : Par défaut
Groupe de mesures : <Tout>

N°	AGE	TAILLE	TAILLE M	Poids	Temperature	CHOLESTEROLE	FC	FR	GLYCEMIE	IMC
1	40,3	167	1,67	59	36,5	169	84	28	160	21,1552942020151
2	25,4	160	1,6	55	36,3	175	88	24	140	21,484375
3	8,2	85	0,85	35	38	135	80	28	120	48,4429065743945
4	7	90	0,9	36	39	173	76	20	160	44,4444444444444
5	3,7	167	1,67	30	37	186	80	28	140	10,7569292552619
6	2,1	167	1,67	20	38	161	84	20	120	7,17128617017462
7	63,2	181	1,81	47	36,9	202	100	28	160	14,3463264247123
8	48	167	1,67	57	36,9	91	60	24	140	20,4381655849977
9	30,6	190	1,9	53	36,4	113	68	28	120	14,6814404432133
10	26	167	1,67	65	36,8	268	88	25	160	23,3066800530675
11	19,8	156	1,56	58	36,1	109	76	32	140	23,8330046022354
12	14	167	1,67	65	37	192	96	18	120	23,3066800530675
13	16,2	167	1,67	54	37	80	88	30	160	19,3624726594715
14	4,6	161	1,61	59	36,5	82	80	25	140	22,7614675359747
15	2,8	167	1,67	48	36,2	156	98	20	120	17,2110868084191
16	1,6	190	1,9	69	36,3	198	76	20	160	19,1135734072022
17	0,5	167	1,67	64	37,1	170	88	20	140	22,9481157445588
18	45	165	1,65	68	37,1	188	80	20	120	24,9770431588613
19	36,4	167	1,67	51	36,5	82	78	20	160	18,2867797339453
20	19,8	166	1,66	66	36,9	37	64	18	140	23,9512265931195
21	16,9	161	1,61	50	35,9	194	64	20	120	19,2893792677752
22	11,8	165	1,65	52	36,7	294	72	32	160	19,1000918273646
23	8,9	161	1,61	56	36,8	94	76	24	140	21,6041047799082
24	5,2	167	1,67	91	36,8	126	78	20	120	32,6293520742945
25	1,7	160	1,6	51	36,6	192	68	20	160	19,921875
26	52,2	156	1,56	55	36,8	232	72	28	140	22,6002629848784
27	39,3	156	1,56	60	36,8	184	64	32	120	24,65483234714

Figure 2: Content of the data warehouse

4.2. Application of the Classical Classification Method

This method forms data sets with common characteristics. By applying our method, we found 10 classes (clusters): Figure 3 represents the classes obtained by this method.



Mod. expl. données : Feuil1

Visionneuse : Microsoft Cluster Viewer

Variable d'ombrage : Remplissage

Densité : Aucun 19%

État :

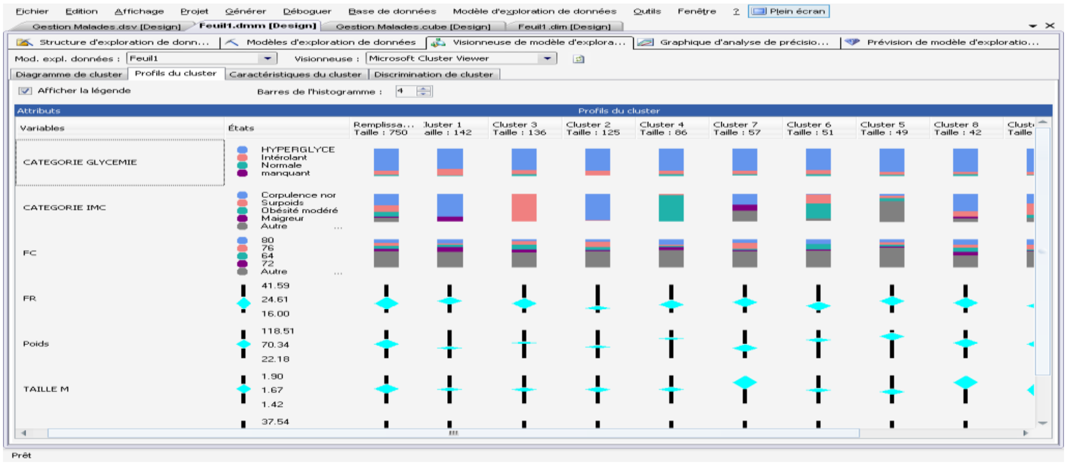
Figure 3: Representation of classes (clusters)

The classes have an influence on blood glucose and BMI. Each class has a given number of attributes, and the attributes of each class have a given number of records:

1. BLOOD GLUCOSE CATEGORY
    - HYPERGLYCEMIA: 591
    - BORDERLINE: 119
    - NORMAL: 39
2. BMI CATEGORY
    - Normal weight: 312
    - Malnutrition or famine: 35
    - Underweight: 47
    - Moderate obesity: 128

4.2.1. CLUSTER PROFILE

This provides the attributes that influence the subjects' membership in each class. It then provides the qualifications and properties of each class. Figure 4 represents the class profiles.



Mod. expl. données : Feuil1

Visionneuse : Microsoft Cluster Viewer

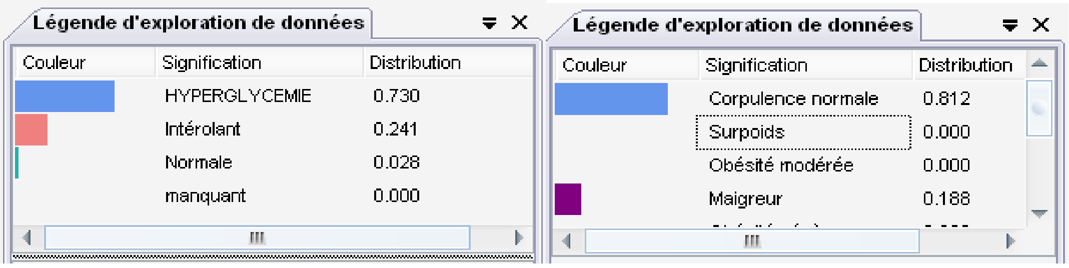
☑ Afficher la légende

Barres de l'histogramme : 4

Figure 4: Class profiles

The class legends are summarized as follows:

CLASS 1



Légende d'exploration de données

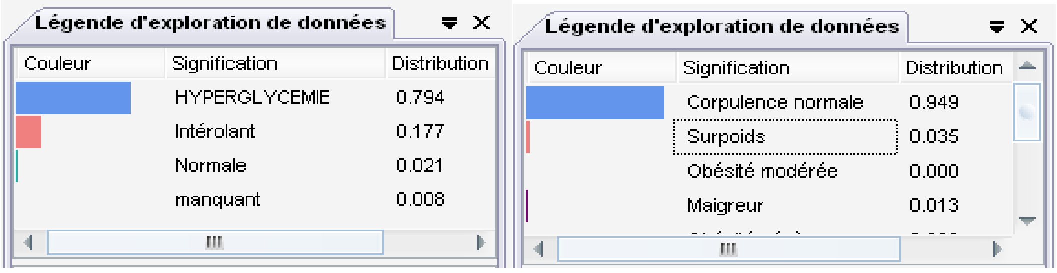
Couleur	Signification	Distribution
🟦	HYPERGLYCEMIE	0.730
🟥	Intérolant	0.241
▏	Normale	0.028
	manquant	0.000

Légende d'exploration de données

Couleur	Signification	Distribution
🟦	Corpulence normale	0.812
	Surpoids	0.000
	Obésité modérée	0.000
🟪	Maigreur	0.188

Class 1: Dominated by subjects with normal BMI (81.2%) and underweight individuals (18.8%). Most subjects also presented with hyperglycemia (73%), followed by borderline glucose levels (24%).

CLASS 2



Légende d'exploration de données

Couleur	Signification	Distribution
🟦	HYPERGLYCEMIE	0.794
🟥	Intérolant	0.177
▏	Normale	0.021
	manquant	0.008

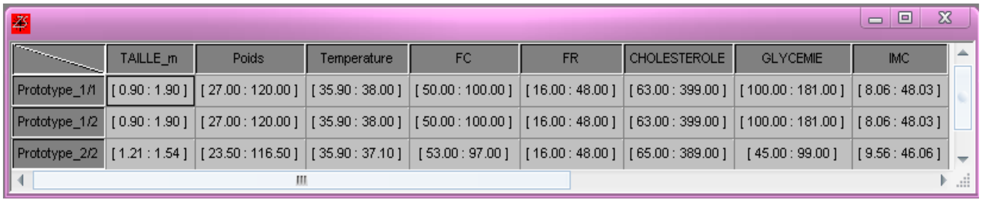
Légende d'exploration de données

Couleur	Signification	Distribution
🟦	Corpulence normale	0.949
▏	Surpoids	0.035
	Obésité modérée	0.000
▏	Maigreur	0.013

Class 2: Predominantly composed of individuals with normal BMI (94.9%), along with a small proportion of overweight (3.5%) and underweight (1%) subjects. The majority of individuals in this class exhibit hyperglycemia (79.4%), with borderline cases (17.7%) and very few normoglycemic profiles (2.1%).

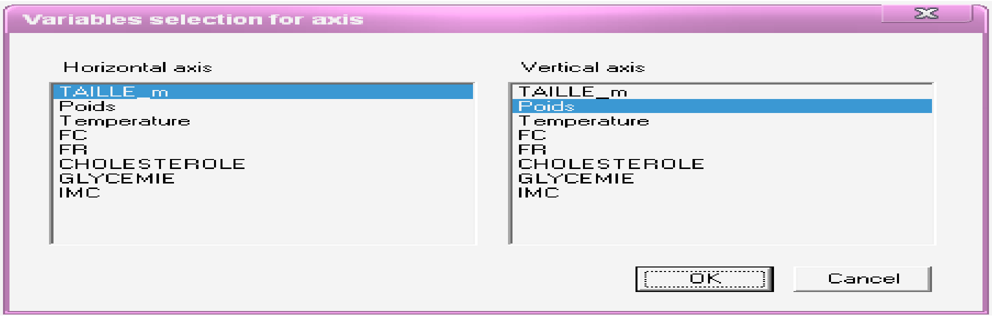
4.3. Application of the Symbolic Dynamic Classification Method

This method obtains as a result 3 prototypes of symbolic objects:



	TAILLE_m	Poids	Temperature	FC	FR	CHOLESTEROLE	GLYCEMIE	IMC
Prototype_1/1	[0.90 : 1.90]	[27.00 : 120.00]	[35.90 : 38.00]	[50.00 : 100.00]	[16.00 : 48.00]	[63.00 : 399.00]	[100.00 : 181.00]	[8.06 : 48.03]
Prototype_1/2	[0.90 : 1.90]	[27.00 : 120.00]	[35.90 : 38.00]	[50.00 : 100.00]	[16.00 : 48.00]	[63.00 : 399.00]	[100.00 : 181.00]	[8.06 : 48.03]
Prototype_2/2	[1.21 : 1.54]	[23.50 : 116.50]	[35.90 : 37.10]	[53.00 : 97.00]	[16.00 : 48.00]	[65.00 : 389.00]	[45.00 : 99.00]	[9.56 : 46.06]

Figure 5: Table representing class prototypes



Variables selection for axis

Horizontal axis: TAILLE_m (selected), Poids, Temperature, FC, FR, CHOLESTEROLE, GLYCEMIE, IMC

Vertical axis: TAILLE_m, Poids (selected), Temperature, FC, FR, CHOLESTEROLE, GLYCEMIE, IMC

Figure 6: Table representing variable axes

Prototypes 1/1 and 1/2 are represented respectively by:

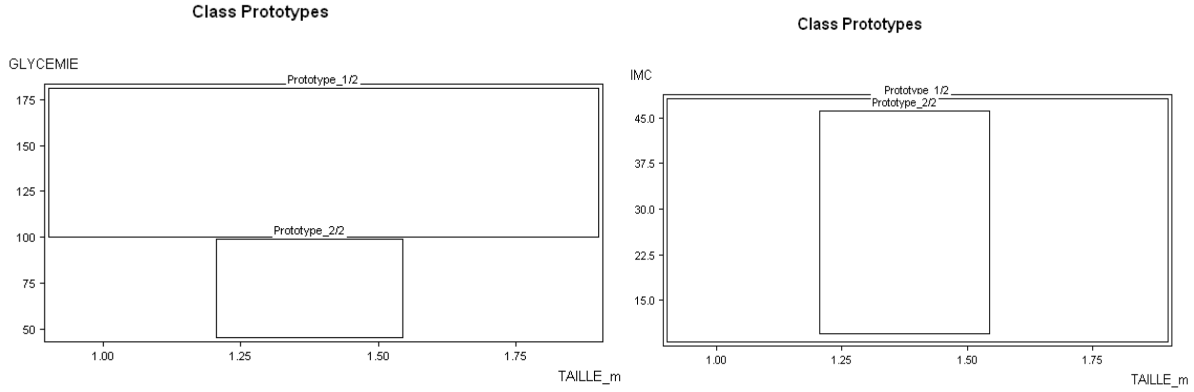
- A height of length 1 m and middle 1.4 m;
- A weight of length 93 m and middle 73.5 m;
- A temperature of length 2.1° and middle 37 m;

- A heart rate of length 50 beats/minute and middle 75 m;
- A respiratory rate of length 32 cycles/minute and middle 32 m;
- A blood glucose of length 81 milligrams/deciliter and middle 140.5 m;
- A BMI of length 40 kg/m² and middle 28 m.

Prototype 2/2 is represented by:

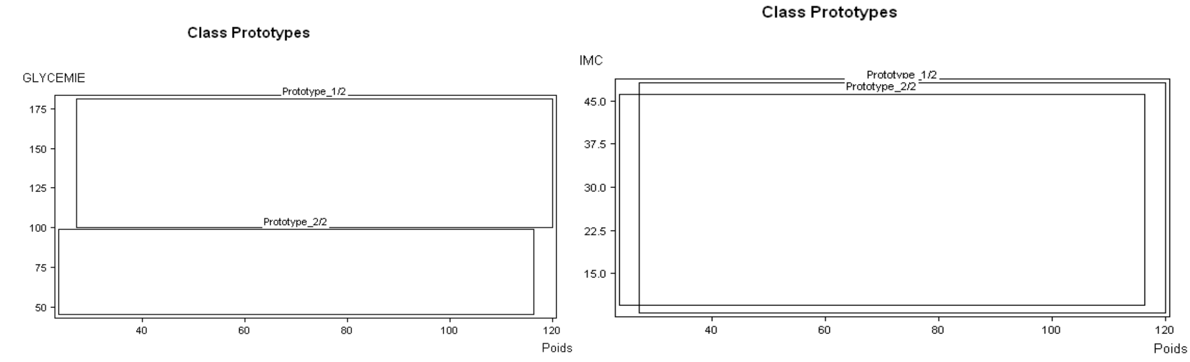
- A height of length 0.3 m and middle 1.4 m;
- A weight of length 93 m and middle 70 m;
- A temperature of length 1.2° and middle 36.5 m;
- A heart rate of length 44 beats/minute and middle 75 m;
- A respiratory rate of length 32 cycles/minute and middle 32 m;
- A blood glucose of length 54 milligrams/deciliter and middle 72 m;
- A BMI of length 36.5 kg/m² and middle 27.8 m.

Figures 7-10 represent histograms showing the evolution of prototypes.



(a) **(b)**

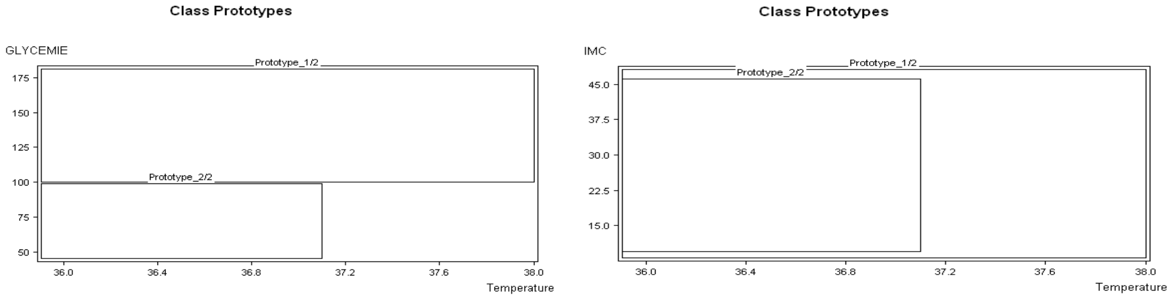
Figure 7: Histogram of prototype evolution with respect to (a) height and blood glucose, and (b) height and BMI. With respect to height and blood glucose, prototype 1/2 presents subjects whose blood glucose is between 100 to 180 milligrams/deciliter and whose height is between 0.7 to 2 m. While prototype 2/2 presents subjects whose blood glucose is between 50 to 100 milligrams/deciliter and whose height is between 1.2 to 1.55 m, as shown in (a).With respect to height and BMI, prototype 1/2 presents subjects whose BMI is between 7.5 to 50 kg/m² and whose height is between 0.7 to 2 m. While prototype 2/2 presents subjects whose BMI is between 8 to 45 kg/m² and whose height is between 1.2 to 1.55 m, as shown in (b).



(a) **(b)**

Figure 8: Histogram of prototype evolution with respect to (a) weight and blood glucose, and (b) weight and BMI. With respect to weight and blood glucose, prototype 1/2 presents subjects whose blood glucose is between 100 to 180 milligrams/deciliter and whose weight is between 25 to 115 kg. While prototype 2/2 presents subjects whose blood glucose is between 45 to 100 milligrams/deciliter and whose weight is between 20 to 120 kg.

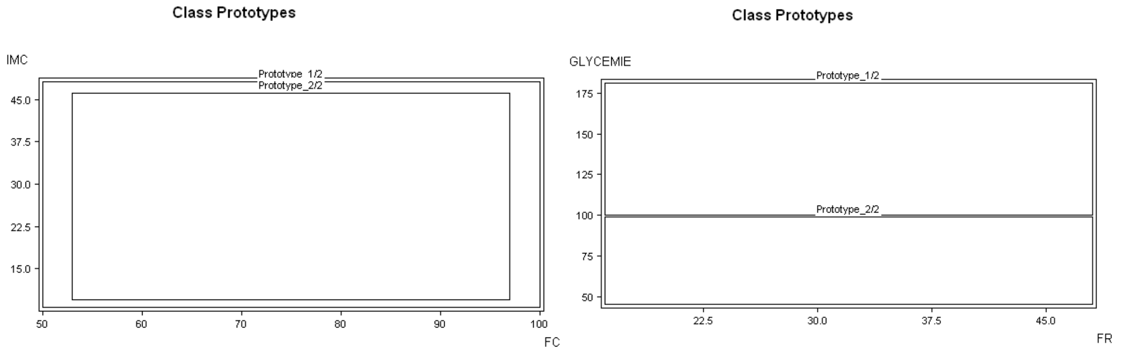
With respect to weight and BMI, prototype 1/2 presents subjects whose BMI is between 7.5 to 48 kg/m² and whose weight is between 25 to 120 kg. While prototype 2/2 presents subjects whose BMI is between 8 to 45 kg/m² and whose weight is between 20 to 115 kg.



(a) (b)

Figure 9: Histogram of prototype evolution with respect to (a) temperature and blood glucose, and (b) temperature and BMI.

With respect to temperature and blood glucose, prototype 1/2 presents subjects whose blood glucose is between 100 to 180 milligrams/deciliter and whose temperature is between 35.8 to 38°C. While prototype 2/2 presents subjects whose blood glucose is between 45 to 100 milligrams/deciliter and whose temperature is between 35.8 to 37°C. With respect to temperature and BMI, prototype 1/2 presents subjects whose BMI is between 10 to 50 kg/m² and whose temperature is between 35.8 to 38°C. While prototype 2/2 presents subjects whose BMI is between 12 to 47 kg/m² and whose temperature is between 35.8 to 37°C.



(a) (b)

Figure 8: Histogram of prototype evolution with respect to (a) heart rate and blood glucose, and (b) heart rate and BMI.With respect to heart rate and blood glucose, prototype 1/2 presents subjects whose blood glucose is between 100 to 180 milligrams/deciliter and whose heart rate is between 50 to 100 beats/minute. While prototype 2/2 presents subjects whose blood glucose is between 45 to 100 milligrams/deciliter and whose heart rate is between 53 to 97 beats/minute. With respect to heart rate and BMI, prototype 1/2 presents subjects whose BMI is between 10 to 50 kg/m² and whose heart rate is between 50 to 100 beats/minute. While prototype 2/2 presents subjects whose BMI is between 12 to 47 kg/m² and whose heart rate is between 53 to 97 beats/minute.

4.4. Comparative Study of Results

4.4.1. Summary of Results

4.4.1.1 Summary of Classical Methods

Table 5 : Results of the classical method at the blood glucose level

Class	**Hyperglycemia**	**Borderline**	**Normal**
Class 1	73%	24.1%	2.8%
Class 2	79.4%	17.7%	2.1%
Class 3	76.8%	14.9%	8.3%
Class 4	79.7%	12.8%	7.5%
Class 5	84.3%	11.3%	4.4%

Class 6	76.6%	17%	6.4%
Class 7	80.8%	14.5%	4.7%
Class 8	84.7%	8.6%	6.7%
Class 9	84.2%	9.7%	6.1%
Class 10	84.7%	12.6%	2.8%

All the classes considered show a high percentage for hyperglycemic subjects, a low percentage for borderline subjects, and an even lower percentage for normal subjects.

Table 6: Results of the classical method at the BMI level

Class	**Normal weight**	**Overweight**	**Moderate obesity**	**Underweight**	**Severe obesity**	**Malnutrition or famine**	**Morbid obesity**
Class 1	81.2%			18.8%			
Class 2	1.5%	98.3%	1%				
Class 3	1.5%	98.3%	1%				
Class 4		6%	94%				
Class 5	8%	10%	12.5%		55.4%		14%
Class 6	2.2%	33.3%	55.2%		9.2%		
Class 7	39.1%	1%		21.7%		39%	
Class 8	61.8%	20.6%		6.9%	2.6%	6.2%	
Class 9	38%	17.2%		12%		26.5%	
Class 10	34.6%	40.6%	11.7%	1%		8.4%	3.7%

The results of the classical method at the BMI level are presented as follows:

- Class 1 shows 81.2% of subjects with normal weight and 18.8% of underweight subjects.
- Class 2 shows 1.5% of subjects with normal weight, 98.3% of overweight subjects, and 1% of subjects with moderate obesity. The same applies to Class 3.
- Class 4 shows 6% of overweight subjects and 94% of subjects with moderate obesity.
- Class 5 shows 8% of subjects with normal weight, 10% of overweight subjects, 12.5% of subjects with moderate obesity, and 55.4% with severe obesity.
- Class 6 shows 2.2% of subjects with normal weight, 33.3% of overweight subjects, 55.2% of subjects with moderate obesity, and 9.2% of subjects with severe obesity.
- Class 7 shows 39.1% of subjects with normal weight, 1% of overweight subjects, 21.7% of underweight subjects, and 39% of subjects with malnutrition or famine.
- Class 8 shows 61.8% of subjects with normal weight, 20.6% of overweight subjects, 6.9% of underweight subjects, 2.6% of subjects with severe obesity, and 6.2% of subjects with malnutrition.
- Class 9 shows 38% of subjects with normal weight, 17.2% of overweight subjects, 12% of underweight subjects, and 26.5% of subjects with malnutrition.
- Class 10 shows 34.6% of subjects with normal weight, 40.6% of overweight subjects, 11.7% of subjects with moderate obesity, 1% of underweight subjects, 8.4% of subjects with malnutrition, and 3.7% of subjects with morbid obesity.

3.4.1.2. Summary of Symbolic Methods

Table 7: Results of the symbolic method at the BMI level

Symbolic Object	**Normal weight**	**Morbid obesity**	**Malnutrition**	**Underweight**	**Moderate obesity**	**Overweight**	**Severe obesity**
Hyperglycemia	41%	1%	5%	7%	17%	23%	6%
Borderline	49%	2%	6%	9%	14%	19%	2%
Normal	30%	2%	7%	6%	24%	28%	4%

Each symbolic object is dominated by subjects with normal weight. Next are overweight subjects, subjects with moderate obesity, underweight subjects, subjects with malnutrition, and subjects with severe obesity. In these symbolic objects, the normal weight characteristic prevails. Results confirmed by classical methods.

Table 8: Results of the symbolic method at the blood glucose level.

Category_blood glucose	
Hyperglycemia	HYPER(1.00)
Borderline	INTER(1.00)
Normal	NORM(1.00)

The table confirms that all hyperglycemia, borderline, and normal are blood glucose (characteristics). The three symbolic data analysis methods have similarities in terms of partition into three classes. They present the same symbolic objects: borderline, normal, and hyperglycemia.

The various analyses used give results presented differently but complementary. From a statistical point of view, the classical method explores the data in detail but does not adapt to the analysis of complex data.

While symbolic methods present objects in a synthetic or global manner and cause a loss of information.

In all classes, subjects are dominated by hyperglycemia followed by borderline subjects and normal subjects (Classical method). This is confirmed by the symbolic method, see tables 6 and 7. However, symbolic methods are more interpretative because of graphical representations.

4.5.Comparative Discussion of the Two Classification Methods

The classical and symbolic dynamic classification methods were applied to group individuals based on their biometric characteristics (height, weight, temperature, heart rate, respiratory rate, blood glucose, and BMI). The results reveal significant differences in how each method structures data, identifies subject profiles, and interprets health phenomena.

4.5.1. Structure and Granularity of Classification

The classical method formed 10 distinct clusters, providing fine-grained population segmentation. Each cluster is characterized by specific proportions of glucose levels (hyperglycemia, borderline, normal) and BMI categories (normal weight, moderate obesity, malnutrition, etc.). While this granularity enables targeted analysis, it may sometimes lead to redundancy or over-segmentation, particularly when multiple clusters show very similar distributions (e.g., Clusters 2 and 3).

In contrast, the Symbolic Dynamic Classification Method (SDCM) produces three symbolic class prototypes. Each prototype is defined by intervals (ranges) and central values (medians) for each variable. This condensed representation provides a synthetic yet dynamic view of typical profiles while accounting for measurement variability. It thus enables more flexible and symbolic modeling of diverse health states.

4.5. 2. Interpretation of Glucose Results

Results from the classical method (Table 5) show that all clusters contain a majority of hyperglycemic subjects (>70% in most cases). This reflects both the high prevalence of hyperglycemia in the sample and a potential limitation of the method in identifying clusters rich in normoglycemic or borderline subjects. This could bias clinical interpretation or mask certain at-risk subpopulations.

The SDCM enables evolutionary analysis through histograms (Figures 7-10). We observe that:

- Prototype 1/2 primarily represents subjects with glucose levels between 100-180 mg/dL
- Prototype 2/2 better represents subjects with glucose levels between 45-100 mg/dL

This distinction models state transitions (from normal to hyperglycemic states), which is difficult to achieve with the rigid clusters of classical clustering.

4.5.3. Analysis of BMI Profiles

The classical method clearly differentiates clusters by BMI categories, for example:

- Cluster 1: 81.2% normal weight
- Cluster 5: 55.4% severe obesity
- Cluster 6: 55.2% moderate obesity

This classification is effective for **static identification** of weight-based groups. However, it provides no information about intra-cluster variation amplitude or potential evolutions.

The SDCM, through interval ranges and centers for BMI, illustrates **internal group variability**. For example:

- Prototype 1/2 shows BMI between **7.5-50 kg/m²**, covering a wide spectrum from malnutrition to morbid obesity.

This provides better sensitivity to individual case diversity, particularly valuable in clinical contexts where measurements may fluctuate (e.g., during treatments or hospitalizations).

4.5.4.Comparative Advantages and Limitations

To better highlight the respective strengths and limitations of each approach, Table 8 provides a comparative overview based on key analytical and practical criteria. This summary facilitates a clearer understanding of how each method performs in terms of granularity, interpretability, adaptability to variability, and clinical relevance.

Table 9: Comparative Summary of Classical and Symbolic Classification Approaches

Criterion	**Classical Method**	**Dynamic Symbolic Method (DSCM)**
Granularity	Very fine (10 distinct classes)	Synthetic (3 prototypes)
Readability	Simple class-by-class interpretation	More complex symbolic representation
Intra-class variability	Not represented	Accounted for via interval lengths
Temporal evolution	Not modeled	Possible representation via histograms
Clinical orientation	Good for typed diagnoses	Good for monitoring and state transitions
Risk of over-segmentation	High	Low

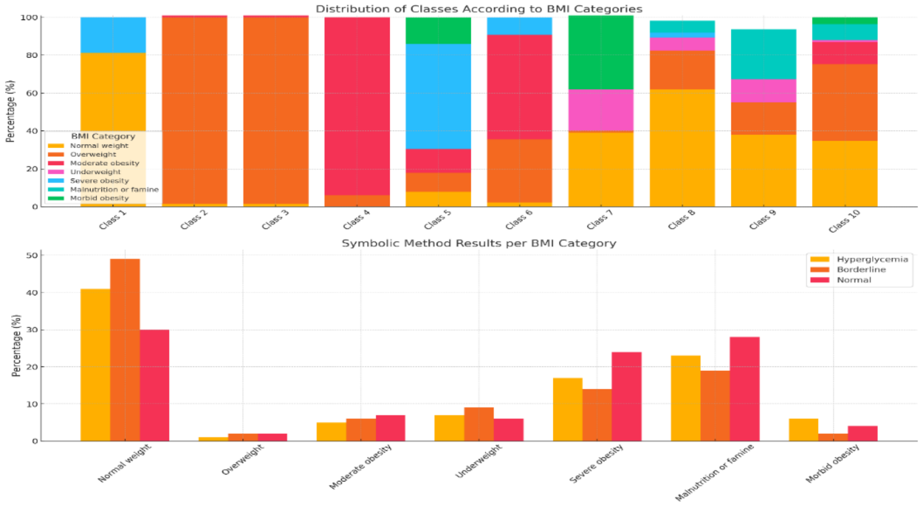


Figure 11 : Top plot: Shows the distribution of each Class across different BMI categories and Bottom plot: Compares the Symbolic method results (Hyperglycemia, Borderline, Normal) across the same BMI categories.

5. Conclusion

Our work focused on the comparative study of classical and symbolic classification of data in a data warehouse. The objective pursued in our work was to design a data warehouse, using a tool for its analysis in order to make good decisions and then apply it in the medical context in relation to glycemic status and nutritional status, making analyses based on two approaches: classical and symbolic.

The classical approach allowed the classification of subjects according to their classification profiles in relation to all indicators (Blood Glucose, BMI, etc.). It also allowed the formation of sets of classes with common characteristics. In our case, the characteristics found were the measures or properties of the indicators used to distinguish the classes. Once the classes were established on the sample, any new object should be assigned, thanks

to its membership in a homogeneous class (whose behavior is known), by examining its characteristics, to allow decision-makers to apply treatments according to its class.
The symbolic approach was applied to more complex data. It started from symbolic data (variables with multiple values, interval, histogram, probability distribution, etc.) equipped with rules and taxonomies and was able to provide new knowledge in the form of symbolic objects as output. This approach made it possible to reduce the data in the form of intervals, frequency distributions, graphical representations expressing, among other things, the internal variation of symbolic descriptions.
In view of the results, we note that the two approaches each have their particularities, but they find complementary results, allowing for more refined analyses.

REFERENCES

[1] Bock, H.-H., & Diday, E. (Eds.). (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer.

[2] Billard, L., & Diday, E. (2003). Symbolic data analysis: Conceptual statistics and data mining. *Journal of Statistical Planning and Inference*, 100(2), 457–471.

[3] Brito, P., & Moniz, H. (2018). An overview of clustering methods for symbolic data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1248.

[4] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

[5] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

[6] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[7] Lechevallier, Y., Vrain, C., & Bock, H.-H. (2001). Clustering symbolic objects. In Diday, E. et al. (Eds.), *Symbolic Data Analysis and the SODAS Software* (pp. 51–77). Springer.

[8] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[9] Billard, L., & Diday, E. (2003). Symbolic data analysis: Conceptual statistics and data mining. *Journal of Statistical Planning and Inference*, 100(2), 457–471.

[10] Billard, L., & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley.

[11] Bock, H.-H., & Diday, E. (Eds.). (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer.

[12] Brito, P., & Moniz, H. (2018). An overview of clustering methods for symbolic data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1248.

[13] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

[14] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.

[15] Diday, E., & Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

[16] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

[17] Inmon, W. H. (2005). *Building the Data Warehouse* (4th ed.). Wiley.

[18] Lechevallier, Y., Vrain, C., & Bock, H.-H. (2001). Clustering symbolic objects. In E. Diday et al. (Eds.), *Symbolic Data Analysis and the SODAS Software* (pp. 51–77). Springer.

[19] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

[20] AMBAPOUR S, Introduction à l'Analyse des Données, Bureau d'application des méthodes statistiques et informatiques, Bamsi Reprint, April 2003.

[21] BERTRAND B, Business Intelligence avec SQL SERVER 2008, Edition Dunod, Paris, 2009.

[22] DURAND J.F, Elément d'Analyse Factorielle, Université Montpellier II, 2002.

[23] DIDAY E et KODRATOFF Y, Des Objets de l'Analyse de Données à ceux de l'Analyse des Connaissances, Induction Symbolique et Numérique à partir de Données, Edition Cépaduès, Université Paris Dauphine, 1991.

[24] DIDAY E, LEMAINE J, POUGET J et TESTU F, Eléments d'Analyse de Données, Edition Dunod, 1982.

[25] DIDAY E, L'Analyse des Données Symboliques, un Cadre Théorique et des Outils, cahiers CEREMADE, Université Paris Dauphine, 1998.

[26] DIDAY E, Introduction à l'Approche Symbolique en Analyse de Données, RAIRO (Revue, d'Automatique, d'Informatique et Recherche Opérationnelle), Edition Dunod, 1989.

[27] GARDARIN G : Internet/Intranet et Base des Données (Data Web, Data media, Data Warehouse, Data Mining), Edition Eyrolles, 2000