ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

Project Data analysis:

Library Data analysis, IA and cyber-security

Present by:

Fomin William

Professor: Dr.EL MKHALET MOUNA

School year: 2025/2026

Table of Contents

List of tables

List of figures

I- General Introduction

1- Presentation

Libraries, centuries-old institutions dedicated to the preservation and dissemination of knowledge, have undergone a profound transformation in their role and mission over the past several decades. Traditionally centers for preserving printed collections, they are now evolving into multipurpose spaces integrating digital resources, innovative services, and expanded socio-cultural functions. This transformation is taking place within a global context characterized by the digital revolution, the explosion of information, and the evolution of reading and research practices.

Modern libraries are now defined less by the size of their collections than by their ability to meet the diverse needs of their users. They must reconcile sometimes conflicting requirements: maintaining optimal physical conditions (temperature, humidity) for document preservation, offering suitable workspaces, managing increasing visitor flows, and all within a frequently constrained budget. The management of these institutions therefore relies on a complex balance between eight fundamental parameters: the size of the collections (Books), seating capacity (Seats), visitor numbers (Visitors/day), preservation conditions (Temperature, Humidity), human resources (Staff), financial resources (Budget), and opening hours (Opening Time).

The diversity of situations observed internationally is considerable, as illustrated by the extreme ranges in our sample: from model libraries with 1,000 books to mega-libraries containing 10 million, from institutions welcoming 10 visitors daily to others receiving 20,000, with budgets varying from €10,000 to €50 million. This heterogeneity raises important questions of comparative analysis and categorization.

2- Problematic

Given this diversity, how can we rigorously analyze, compare, and categorize libraries using multidimensional quantitative criteria? More specifically, our research revolves around four main questions:

How can we identify the latent structures underlying the eight measured variables, and what synthetic dimensions best summarize the information contained in these data?

Can we establish a reasoned typology of libraries based on their quantitative characteristics, and how can we determine the optimal number of groups in this classification?

Which variables have the most decisive influence on a library's profile, and how should these different criteria be weighted for a balanced evaluation?

To what extent can data analysis methods be transposed to other fields, such as cybersecurity, and what methodological synergies can be identified between these seemingly disparate fields?

3- Objective

This project aims to answer these questions by implementing a comprehensive analytical approach structured around four axes:

First, apply a Principal Component Analysis (PCA) to reduce the dimensionality of the data, visualize similarities between libraries, and identify linear combinations of variables that best explain the observed variance.

Secondly, implement clustering algorithms with a number of groups ranging from 3 to 7, in order to establish a robust typology of libraries, then use a Random Forest classifier to validate this classification and identify the most discriminating variables.

Third, analyze the impact of variable weighting on the results obtained, then apply a Correspondence Factor Analysis (CFA) to study the relationships between categorized variables.

Fourth, extend the methodological reflection to a related field by applying anomaly detection algorithms (Isolation Forest and LOF) to a cybersecurity-oriented dataset, thus demonstrating the transferability of the analytical skills acquired.

Chapter I: Principal Component Analysis (PCA) applied to Moroccan libraries

This initial chapter carries out an exploratory analysis of the network of 100 Moroccan libraries through Principal Component Analysis (PCA). It pursues a dual objective: First, to reduce the dimensionality of the 8 variables characterizing each library (Books, Spaces, Visitors, Temperature, Humidity, Staff, Budget, Opening Hours) and then to identify the structuring axes and key factors that differentiate the libraries from one another.

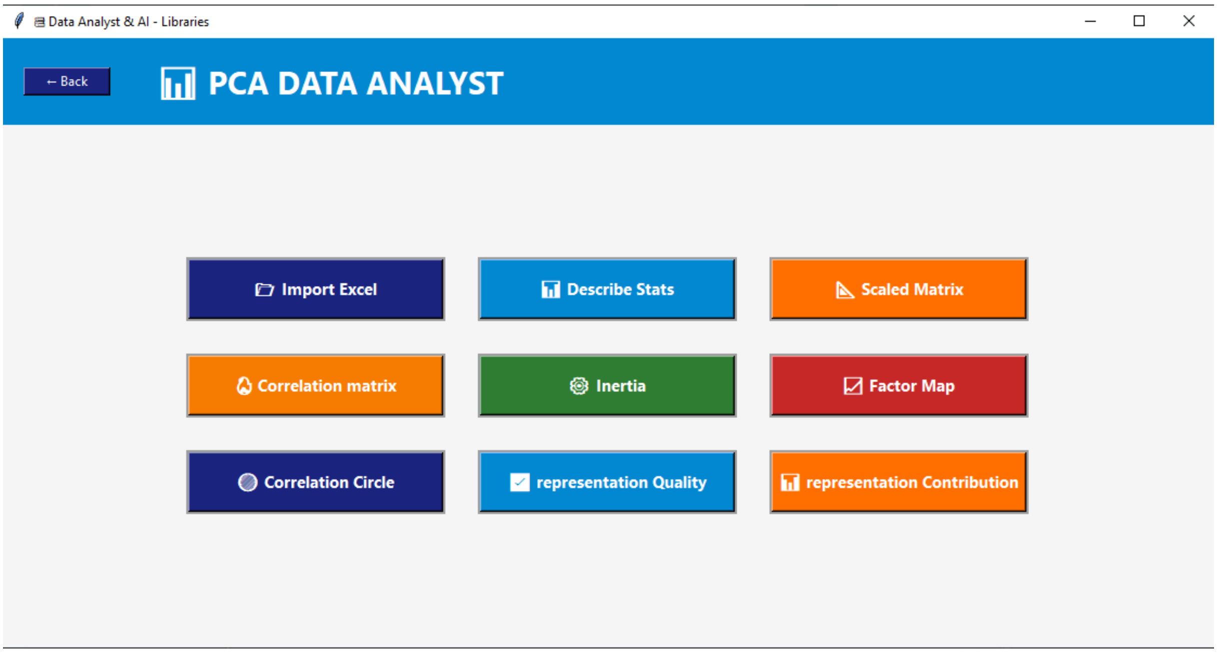
The method employed follows a rigorous approach: data standardization, calculation of principal components, and interpretation of results via correlation circles and individual projections. This approach reveals the main differences between libraries and measures the relative influence of each variable.

The analysis is not an end in itself but serves as an essential foundation for subsequent processing. By condensing the information into synthetic dimensions, it directly prepares the classification work of Chapter II while also offering strategic insights into the organization of the national library network.

This exploratory phase thus answers fundamental questions about the specificities, operational compromises and relative positioning of Moroccan libraries, establishing a quantitative and rigorous understanding of their structural diversity.

Furthermore, the graphical interface representing this part in our desktop application is as follows:

Figure 1: PCA interface


Data Analyst & AI - Libraries
← Back
PCA DATA ANALYST
Import Excel
Describe Stats
Scaled Matrix
Correlation matrix
Inertia
Factor Map
Correlation Circle
representation Quality
representation Contribution

I- Presentation of the topic in the Moroccan context

The Moroccan library network exhibits marked diversity, ranging from institutions of national scope to local structures with more limited resources. Take, for example, the Mohammed VI University Library in Ben Guerir, affiliated with Mohammed VI Polytechnic University. Inaugurated in 2017, it embodies the recent modernization of the sector with its innovative architecture, collections focused on science and technology, and strong digital integration. With a capacity of several hundred seats, a substantial budget, and trained staff, it meets international standards for infrastructure and services. However, this reality contrasts sharply with that of many municipal or school libraries in rural or peri-urban areas, which have far fewer resources, limited collections, and restricted opening hours. This duality – between well-endowed centers of excellence and a territorial network with unequal capacities – underlines the structural heterogeneity of the landscape and justifies a systematic quantitative analysis to grasp its logics, measure the gaps and identify levers for action adapted to each type of establishment.

Next, we will begin principal component analysis, which is a very crucial step in data analysis.

II- Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method for reducing dimensionality that transforms a set of correlated variables into a smaller number of uncorrelated variables, called principal components. Its usefulness in data analysis lies in its ability to synthesize information, visualize data structure in a reduced space, and identify the main sources of variation within a multidimensional dataset. By concentrating relevant information in the primary components, PCA facilitates interpretation, reveals underlying correlations, and prepares the ground for subsequent analyses such as classification or modeling.

1- Mean and standard deviation

1-1- Definitions

The mean and standard deviation are fundamental indicators in descriptive statistics that allow us to characterize the distribution of our variables in a concise way. The mean provides a measure of central tendency, indicating the "typical" value around which the observations are concentrated. The standard deviation, on the other hand, measures the dispersion of the data around this mean: the higher it is, the more heterogeneous the values. To complete this analysis, the coefficient of variation (CV)—the ratio between the standard deviation and the mean—allows us to compare the relative variability between variables with different units. A high CV (> 30%) indicates strong heterogeneity, while a low CV (< 15%) indicates relative homogeneity. See the table of statistics for the different variables below.

Table 1: Descriptive Statistics

	Mean	Std Dev	CV (%)
Books	4832968.020	3298987.991	68.260
Places	2567.870	1298.352	50.561
Visitors_day	10520.110	6280.842	59.703
Temperature_C	20.914	0.871	4.167

Humidity_%	45.217	6.248	13.817
Personnel	239.400	118.916	49.672
Budget_k€	25666.476	14972.985	58.337
Opening_h_d	14.725	4.707	31.965

1-2- Interpretation of descriptive statistics

Based on the statistics obtained in the table above, we can give these different interpretations for each variable.

- **Books**The average is**4,832,968 pounds**with a high standard deviation (**3 298 988**), indicating a high degree of dispersion among libraries. The coefficient of variation (CV) of**68,26 %**confirms a significant variability around the average, which may reflect marked differences in size, attendance or acquisition policy.
- **Places** The average is**2 568 places**, with a standard deviation of**1 298**The CV of**50,56 %**shows a moderate to high dispersion, suggesting a diversity in the capacity of the establishments.
- **Visitors/day**The average is**10,520 visitors/day**, with a standard deviation of**6 281**The CV of**59,70 %**reveals considerable variability, probably linked to the location, size or attractiveness of the libraries.
- **Temperature (°C)**The average is**20,91 °C**, with a low standard deviation (**0,87**The very low CV (**4,17 %**) indicates a remarkable stability of temperatures between sites, reflecting a homogeneous climate control.
- **Humidity (%)** The average is**45,22 %**, with a standard deviation of**6,25**The CV of**13,82 %**signals moderate variability, which may be related to differences in regulatory systems or the local environment.
- **Personnel** The average is**239 agents**, with a standard deviation of**119**The CV of**49,67 %**shows a significant dispersion, revealing significant disparities in human resources between institutions.

- **Budget (k€)**The average is**25 666 k€**, with a high standard deviation (**14 973**). The CV of**58,34 %**confirms a strong budgetary heterogeneity, probably correlated with the size and missions of the libraries.
- **Opening hours (h/d):**The average is**14.73 hours/day**, with a standard deviation of**4,71**The CV of**31,97 %**indicates moderate variability in opening hours, which may reflect differentiated reception policies.

2- Centered-reduced matrix

1-1- Definition

The standardized matrix is an essential statistical transformation in multivariate analysis, particularly before the application of PCA. It is obtained by applying the following transformation to each variable X:

$$Z = \frac{X - \bar{X}}{\sigma_X}$$

Or :

- *X*¯ is the average of the variable
- σX is the standard deviation of the variable

Data standardization, resulting in a standardized matrix, is an essential step before any Principal Component Analysis (PCA). This transformation addresses several fundamental methodological requirements. First, it eliminates units of measurement, making all variables dimensionless and therefore directly comparable. This is crucial for comparing quantities as heterogeneous as the number of books and the temperature. Second, it neutralizes scale bias: without this operation, variables with large numerical values (such as the number of books, expressed in millions) would artificially dominate the analysis at the expense of variables with a smaller scale (such as temperature in degrees). Third, by centering each variable on zero (zero mean) and assigning it a unit variance, we ensure an equitable weighting of all dimensions in the construction of the principal components. This homogenization thus guarantees that the

PCA captures the true covariance structures between variables, and not artifacts related to their initial units or orders of magnitude.

See below for the interface of the standardized matrix. We will only use the first 10 to avoid making the report too large. The rest can be seen in the report.

Table 2: centered-reduced matrix

Library	Books	Places	Visitors_day	Temperature_C	Humidity_%	Personnel	Budget_k€	Opening_h_d
Al Andalus Library of Rabat	0.172	0.201	-0.355	0.676	0.737	0.233	-0.059	0.165
Ibn Khaldoun Library of Rabat	0.431	0.344	0.092	-1.400	-0.984	0.368	0.438	0.294
Ibn Battuta Library of Rabat	0.448	-0.189	-0.813	0.676	0.625	-0.088	0.329	-0.945
Al Quaraouiyine Library of Rabat	-0.245	-0.253	-1.090	-1.054	-0.984	-0.680	-0.514	-1.137
Hassan II Library of Rabat	1.270	1.233	0.802	-0.593	-1.000	0.876	1.633	1.233
Mohammed VI Library of Rabat	1.574	1.147	0.009	-0.362	-0.099	1.138	1.065	0.080
Al Manar Library in Rabat	1.047	1.538	1.517	1.368	1.638	1.070	1.132	1.084
Al Fikr Library of Rabat	1.065	0.739	0.673	-0.593	0.544	1.298	0.850	0.571
Al Nour Library of Rabat	-0.622	-0.930	-1.226	-1.400	0.271	-0.899	-0.803	-1.052
Dar Al Hikma Library of Rabat	0.273	-0.226	-0.915	-0.016	-0.357	0.157	0.203	-0.924

3- Correlation matrix

3-1- Definition

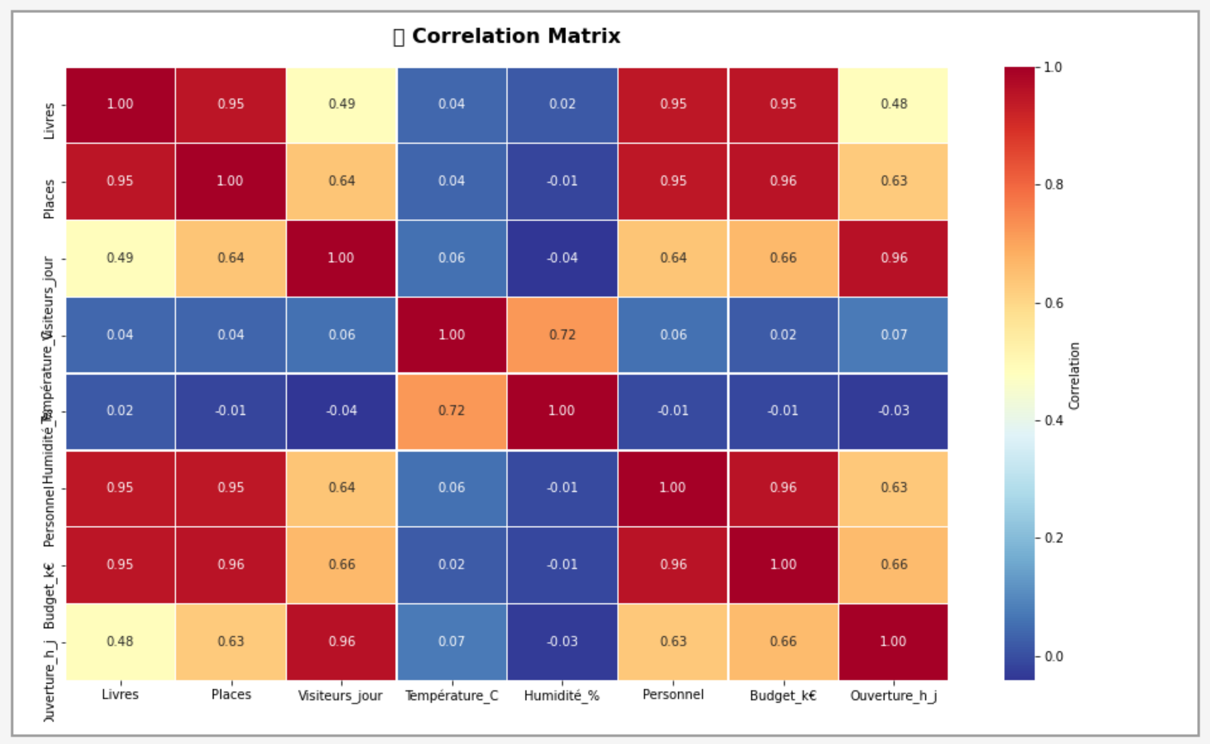
The **correlation matrix**is a symmetric square table that measures the linear relationships between all pairs of variables in a dataset. Each element*row*represents the Pearson correlation coefficient between the variable i*i*and the variable j*j*, varying from -1 to +1.

It is important before PCA for the following reasons:

- **Prerequisite**PCA looks for directions of maximum variance in the data, which requires understanding how variables covariate with each other.
- **Redundancy detection**: Identifies highly correlated variables that could measure similar dimensions.
- **Validation of the PCA**: If the correlations are too weak, PCA may not be relevant.
- **Interpretation of the components**: Allows us to understand which variables "travel together" in the factorial space.

Below is the heatmap-style graphical interface of our correlation matrix.

Figure 2: Graphical interface of the correlation matrix



□ **Correlation Matrix**

	Livres	Places	Visiteurs_jour	Température_C	Humidité_%	Personnel	Budget_k€	Ouverture_h_j
Livres	1.00	0.95	0.49	0.04	0.02	0.95	0.95	0.48
Places	0.95	1.00	0.64	0.04	-0.01	0.95	0.96	0.63
Visiteurs_jour	0.49	0.64	1.00	0.06	-0.04	0.64	0.66	0.96
Température_C	0.04	0.04	0.06	1.00	0.72	0.06	0.02	0.07
Humidité_%	0.02	-0.01	-0.04	0.72	1.00	-0.01	-0.01	-0.03
Personnel	0.95	0.95	0.64	0.06	-0.01	1.00	0.96	0.63
Budget_k€	0.95	0.96	0.66	0.02	-0.01	0.96	1.00	0.66
Ouverture_h_j	0.48	0.63	0.96	0.07	-0.03	0.63	0.66	1.00

3-2- Interpretation of correlation levels

Books ↔ Seats (0.95): Strong positive correlation

*Interpretation:*Libraries with more books have significantly more seating, indicating a close relationship between the size of the collection and seating capacity.

Books ↔ Personal (0.95): Strong positive correlation

*Interpretation:*The more books a library has, the more staff it employs, which suggests that human resources are managed proportionally to the size of the collection.

Livres ↔ Budget (0.95) : Strong positive correlation

*Interpretation:*Libraries with large collections also benefit from higher budgets, reflecting a financial allocation consistent with the size of the collections.

Books ↔ Visitors/day (0.49): Moderate positive correlation

*Interpretation:*A larger collection tends to attract more visitors, although other factors also influence attendance.

Books ↔ Opening/h_d (0.48): Moderate positive correlation

*Interpretation:*Libraries with more books are open slightly longer, perhaps to meet increased demand.

Places ↔ Personnel (0.95) : Strong positive correlation

*Interpretation:*A larger capacity is associated with a larger staff, probably to ensure supervision and services.

Places ↔ Budget (0.96) : Strong positive correlation

*Interpretation:*Libraries with more space have larger budgets, highlighting the link between infrastructure and financial resources.

Places ↔ Visitors/day (0.64): Moderate to strong positive correlation

*Interpretation:*More seating is accompanied by higher attendance, although the relationship is not perfectly linear.

Places ↔ Opening/h_j (0.63): Moderate to strong positive correlation

*Interpretation:*Establishments with a larger capacity tend to have longer opening hours.

Visitors/day ↔ Opening/h_d (0.96): Very strong positive correlation *Interpretation:*Longer opening hours are strongly associated with higher daily visitor numbers, indicating a direct impact of accessibility on the visit.

Visitors/day ↔ Staff (0.64): Moderate to strong positive correlation *Interpretation:*High visitor numbers are generally accompanied by a larger staff, perhaps to manage the crowds and services.

Visitors/day ↔ Budget (0.66): Moderate to strong positive correlation *Interpretation:*Libraries with higher visitor numbers benefit from larger budgets, possibly to support increased activities and services.

Temperature (°C) ↔ Humidity (%) (0.72): Strong positive correlation *Interpretation:*Higher temperature is associated with higher relative humidity, in accordance with common physical climatic relationships.

Staff ↔ Budget (0.96): Very strong positive correlation *Interpretation:*Staffing levels and budgets are closely linked, with human resources representing a significant portion of expenditure.

Staff ↔ Opening/h_d (0.63): Moderate to strong positive correlation *Interpretation:*A larger staff allows or justifies longer opening hours.

Budget ↔ Opening/h_y (0.66): Moderate to strong positive correlation *Interpretation:*Better-funded libraries tend to stay open longer, as their budgets allow them to cover extended operating costs.

Books ↔ Temperature (0.04): Negligible correlation *Interpretation:*There is no significant relationship between the size of the collection and the ambient temperature. Climatic conditions do not depend on the number of books.

Books ↔ Humidity (0.02): Negligible correlation *Interpretation:*Humidity levels are not related to the size of the collection. Book preservation requires humidity control independent of the size of the collection.

Places ↔ Temperature (0.04): Negligible correlation *Interpretation:*The occupancy capacity does not influence the indoor temperature, which is regulated by air conditioning systems independent of occupancy.

Places ↔ Humidity (-0.01): Negligible correlation *Interpretation:*Relative humidity is not correlated with the number of available seats. These two variables evolve independently.

Visitors/day ↔ Temperature (0.06): Negligible correlation *Interpretation:*Daily attendance is not affected by the ambient temperature in the libraries, as they are maintained at a constant level of comfort.

Visitors/day ↔ Humidity (-0.04): Negligible correlation *Interpretation:*The humidity level has no measurable impact on visitor numbers.

Temperature ↔ Personnel (0.06): Negligible correlation *Interpretation:*The size of the workforce does not influence the indoor temperature, which is controlled by independent technical systems.

Temperature ↔ Budget (0.02) : Negligible correlation *Interpretation:*The budget allocated to a library is not related to its average temperature. Air conditioning costs do not appear to be proportional to the overall budget.

Temperature ↔ Opening/h_d (0.07): Negligible correlation *Interpretation:*The daily opening time has no impact on the ambient temperature, as this is regulated automatically.

Humidity ↔ Personnel (-0.01): Negligible correlation *Interpretation:*The number of employees does not affect the humidity level, which depends primarily on the ventilation and dehumidification systems.

Humidity ↔ Budget (-0.01): Negligible correlation *Interpretation:*Larger budgets do not translate into better humidity control, suggesting that all libraries maintain similar levels regardless of their funding.

Humidity ↔ Opening/h_d (-0.03): Negligible correlation *Interpretation:*Opening hours have no influence on the humidity level, which remains stable regardless of operating hours.

4- Inertia

4-1- Definitions

In Principal Component Analysis, inertia represents the quantity of information or total variance contained in the dataset. More specifically:

- **Total inertia**: Sum of variances of all standardized variables
- **Inertia of an axis**Variance explained by each principal component
- **Percentage of inertia**Proportion of total variance captured by each axis

It plays a crucial role in PCA because it forms the basis of several major analytical decisions. First, it serves as an objective selection criterion to determine the number of principal components to retain, thus avoiding excessive or insufficient dimensionality reduction. Second, it allows for the evaluation of the quality of the resulting representation by measuring the proportion of original information retained after projection into the new factor space. Third, it informs the interpretability of the results by indicating the relative importance of each synthetic dimension within the overall dataset. Finally, it guides statistical decision-making, offering practical rules (such as Kaiser's criterion or the explained variance threshold) for rigorously selecting the optimal number of components, thereby ensuring a balance between simplification and information preservation.

4-2- Interpretation of the different results

Let us now move on to the interpretations of the different results we obtained.

First, the eigenvalues. The table below shows the results of the values obtained.

Table 3: Eigenvalues and inertias

Axis	Inertia	Percentage (%)	Cumulative (%)
F1	4.920	60.890	60.890
F2	1.739	21.523	82.413
F3	1.000	12.374	94.787
F4	0.275	3.403	98.190
F5	0.051	0.629	98.820

F6	0.044	0.546	99.365
F7	0.035	0.435	99.800
F8	0.016	0.200	100.000

F1 (Axis 1): Inertia = 4.920, Percentage = 60.890%, Cumulative = 60.890% *Interpretation:*The first axis (F1) alone captures 60.89% of the total variance in the data. This means that this axis summarizes most of the structural information about the libraries. It most likely represents the "size/magnitude" dimension."establishments, combining highly correlated variables such as the number of books, places, staff and budget.

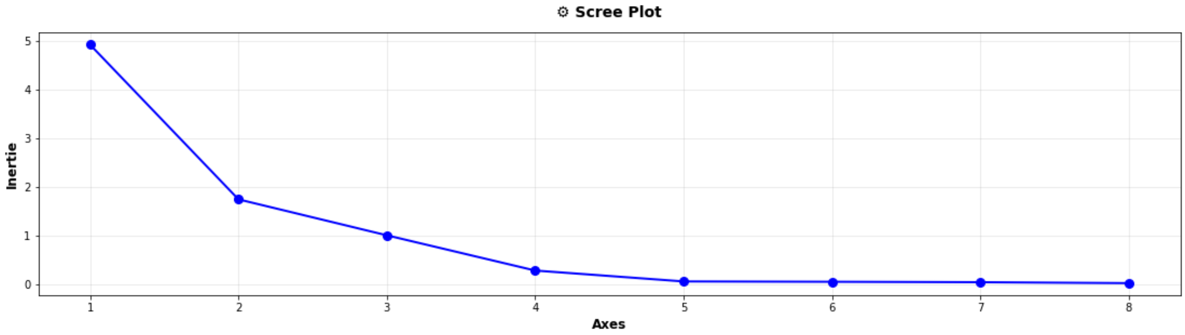
F2 (Axis 2): Inertia = 1.739, Percentage = 21.523%, Cumulative = 82.413% *Interpretation:*The second axis explains an additional 21.52% of the variance, bringing the cumulative total to 82.41%. This axis likely represents a dimension complementary to size, such as visitor numbers and accessibility (visitors/day and opening hours), which are strongly correlated with each other but less directly with pure size variables.

F3 (Axis 3): Inertia = 1.000, Percentage = 12.374%, Cumulative = 94.787% *Interpretation:*The third axis contributes 12.37% of the variance, bringing the total to 94.79%. This axis could reflect environmental or conservation factors (temperature, humidity), which are uncorrelated with the structural dimensions but are nonetheless related. It captures independent information not explained by the first two axes.

F4 to F8 (Axes 4 to 8): Low inertias (0.275 to 0.016), Marginal percentages (3.403% to 0.200%)
*Interpretation:*These axes capture very small proportions of variance (less than 3.5% each). They mainly represent statistical noise, specific individual variations, or unstructured residual interactions between variables. Their contribution to the overall explanation is negligible.

Next we have the***scree plot*** below

Figure 3: sree plot



The**"elbow"**clearly comes after**F3**It is therefore justified to retain the first three axes for the analysis, as they explain the majority of the information (cumulative > 94%). However, we will only retain two axes for the factorial plane. Axes beyond F3 can be ignored for the substantive interpretation, as they do not introduce any new significant dimensions.

5- Factorial plane

5-1- Definitions

The**factorial plane**is a graphical visualization obtained by projecting individuals (libraries) and/or variables into the reduced space formed by two principal components. In concrete terms, it is a two-dimensional graph where:

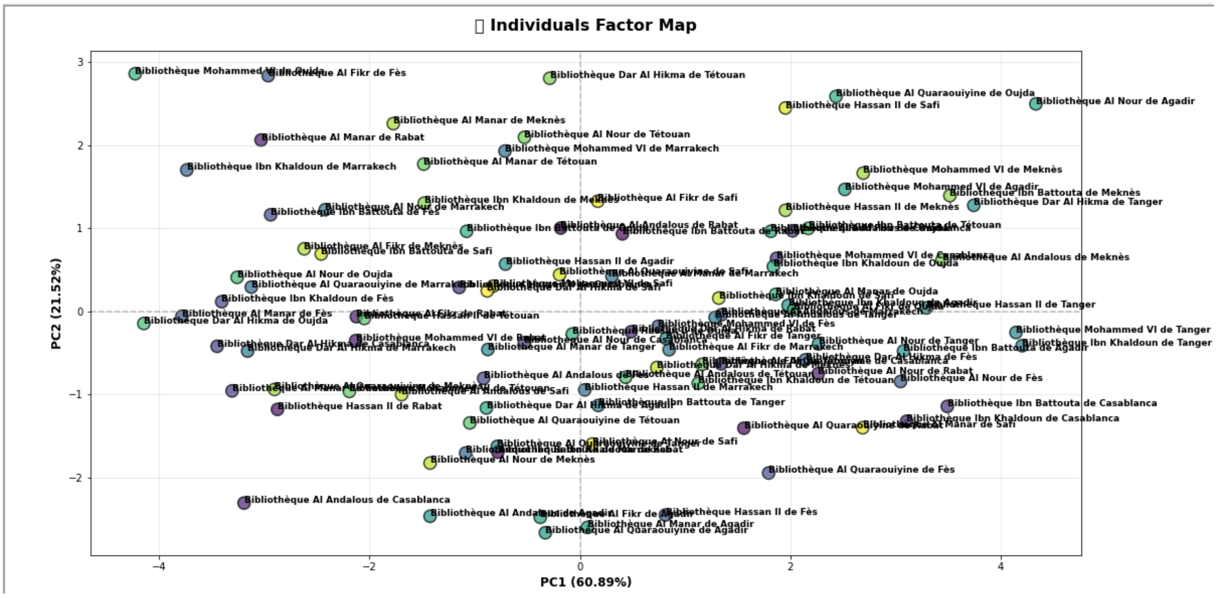
- **L'axe horizontal**generally represents the first principal component (F1)
- **L'axe vertical**represents the second principal component (F2)
- **Each point**corresponds to a library
- **Each vector**(arrow) corresponds to a variable

It's like taking a 2D picture of a point cloud that actually exists in an 8-dimensional space (our 8 variables).

5-2- Interpretation

See below the representative factorial plane of our dataset.

Figure 4: Factorial plane



The Individuals Map (ICM) presents a two-dimensional reduction where ICM1 explains 60.89% and ICM2 21.52% of the total variance (approximately 82% combined), allowing for a quick understanding of library profiles. ICM1 constitutes the primary axis of differentiation: libraries aligned with this axis share similar dominant characteristics, while ICM2 provides a secondary dimension that distinguishes complementary profiles. In practice, nearby points indicate similar offerings and users, while distant points reflect different profiles. This simplification suggests encouraging exchanges between nearby libraries to share best practices and analyzing discrepancies among more distant libraries to identify targeted actions (for example, based on location or type of offering).

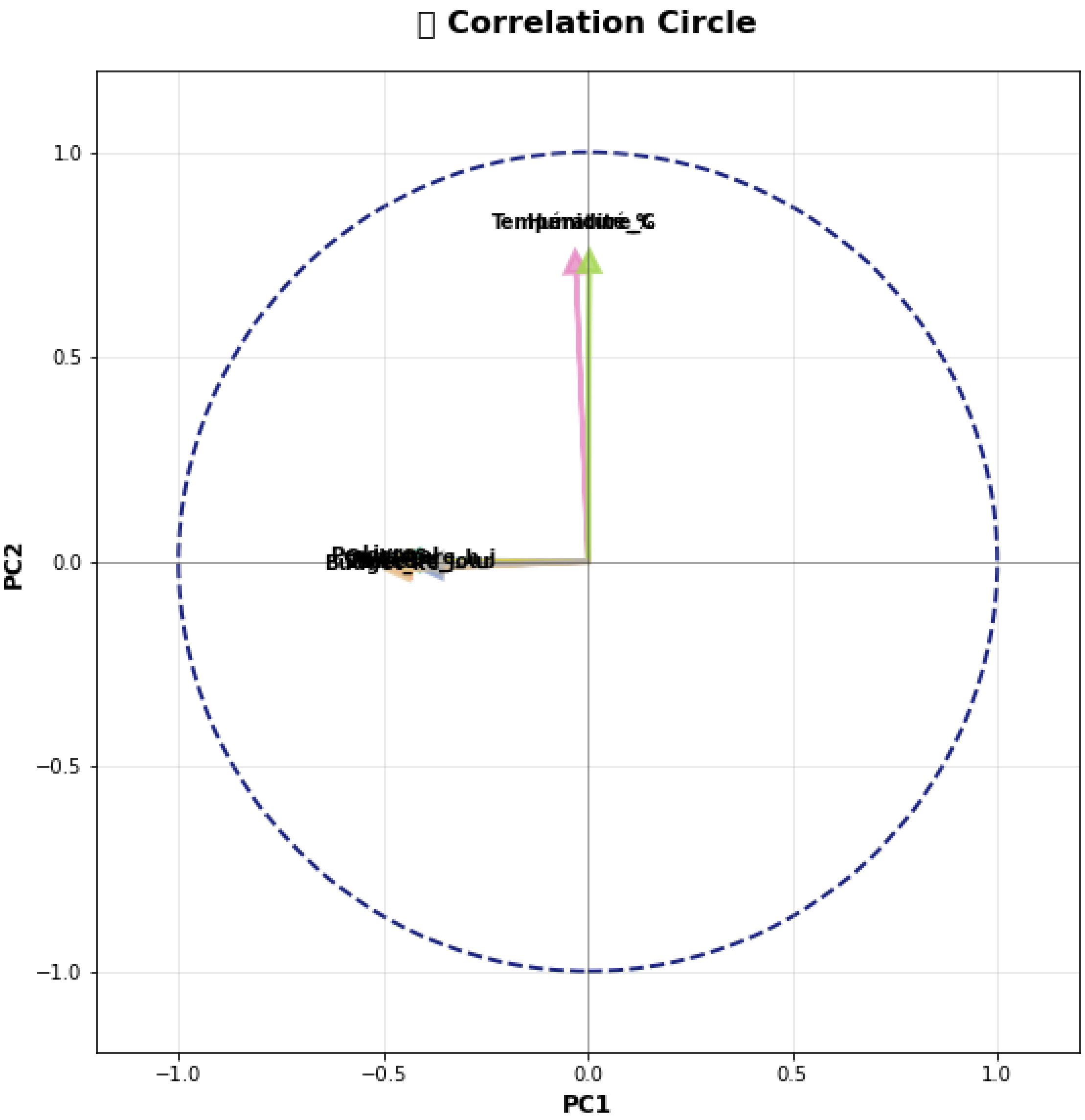
6- Correlation circle

6-1- Definition

The**correlation circle**is a graphical representation that shows the relationships between the original variables and the principal components (factor axes). It is an essential tool for understandingwhat they measureThe axes of PCA.

Below is the correlation circle:

Figure 5: Circle of this correlation



6-2- Interpretation

The correlation circle above highlights two main dimensions: a dimension**“resources/attractiveness”**(budget, staff, places, books, opening hours, visitors) and a dimension**"climate"**(temperature, humidity). Resource variables are strongly correlated with each other and associated with visitor numbers, while climatic variables evolve independently. Operational levers for increasing visitors are primarily on the side of...**resources**and**the opening range**.

7- Quality of representation

7-1- Definition

The **quality of representation**(or cos²) measures how well an individual (library) is represented in the factorial plane. It is a value between 0 and 1 that indicates the proportion of the individual's actual position in multidimensional space that is captured by its projection onto the F1-F2 plane. Here is the table ofThe quality of representation below is from just 10 libraries.

Table 4: Quality of representation

Library	PC1	PC2	Quality_PC1_PC2
Al Andalus Library of Rabat	0.026	0.784	0.810
Ibn Khaldoun Library of Rabat	0.167	0.789	0.955
Ibn Battuta Library of Rabat	0.059	0.316	0.375
Al Quaraouiyine Library of Rabat	0.449	0.367	0.816
Hassan II Library of Rabat	0.820	0.139	0.958
Mohammed VI Library of Rabat	0.714	0.020	0.733
Al Manar Library in Rabat	0.660	0.306	0.966
Al Fikr Library of Rabat	0.819	0.001	0.820
Al Nour Library of Rabat	0.698	0.077	0.775

Dar Al Hikma Library of Rabat	0.118	0.030	0.148

7-2- Interpretation

The individuals well represented on PC1 are:

- **Mohammed VI Library of Tangier (PC1 = 0.980)**
- **Ibn Khaldoun Library of Tangier (PC1 = 0.960)**
- **Dar Al Hikma Library in Oujda (PC1 = 0.976)**
- **Al Nour Library of Oujda (PC1 = 0.949)**
- **Ibn Khaldoun Library of Fez (PC1 = 0.953)**
- **Al Manar Library in Fez (PC1 = 0.973)**
- **Dar Al Hikma Library in Casablanca (PC1 = 0.906)**

Interpretation: These libraries strongly structure the main axis (PC1) and are very well represented. They typically reflect the size/resources dimension, probably corresponding to the largest libraries in terms of books, spaces, staff and budget.

The individuals who are underrepresented on PC1 are:

- **Hassan II Library of Marrakech (PC1 = 0.000)**
- **Al Manar Library in Marrakech (PC1 = 0.032)**
- **Bibliothèque Al Fikr de Casablanca (PC1 = 0.003)**
- **Dar Al Hikma Library of Tetouan (PC1 = 0.009)**
- **Al Andalus Library of Tetouan (PC1 = 0.054)**
- **Hassan II Library of Oujda (PC1 = 0.008)**
- **Dar Al Hikma Library of Agadir (PC1 = 0.140)**

Interpretation: These libraries have very low coordinates on PC1, meaning they do not contribute to the structuring of the main axis. They are not characterized by the size/resource attributes that PC1 represents.

The individuals well represented on PC2 are:

- **Al Andalus Library of Agadir (PC2 = 0.709)**
- **Ibn Battouta Library of Tangier (PC2 = 0.671)**
- **Hassan II Library of Fez (PC2 = 0.640)**
- **Al Andalus Library of Rabat (PC2 = 0.784)**
- **Ibn Khaldoun Library of Rabat (PC2 = 0.789)**
- **Al Quaraouiyine Library of Agadir (PC2 = 0.842)**
- **Al Manar Library in Agadir (PC2 = 0.857)**
- **Al Fikr Library in Agadir (PC2 = 0.881)**
- **Al Nour Library of Tetouan (PC2 = 0.891)**
- **Dar Al Hikma Library of Tetouan (PC2 = 0.892)**

Interpretation: These libraries are strongly associated with the second dimension (PC2), which probably represents the **visitor numbers/accessibility** (visitors/day, opening hours). These establishments are distinguished by their dynamic approach to welcoming visitors or their opening policy.

The individuals who are poorly represented on PC2 are:

- **Al Manar Library in Fez (PC2 = 0.000)**
- **Hassan II Library of Tangier (PC2 = 0.000)**
- **Mohammed VI Library of Tangier (PC2 = 0.004)**
- **Dar Al Hikma Library in Oujda (PC2 = 0.001)**
- **Al Nour Library of Oujda (PC2 = 0.015)**
- **Ibn Khaldoun Library of Agadir (PC2 = 0.001)**
- **Ibn Battuta Library of Agadir (PC2 = 0.022)**

Interpretation: These libraries do not contribute significantly to axis PC2. Their profile is not characterized by the usage/accessibility attributes that this axis represents.

The individuals who are very well represented in terms of factorial distribution (PC1 + PC2) are:

- **Mohammed VI Library of Oujda (Quality = 0.989)**

- **Al Andalus Library of Casablanca (Quality = 0.986)**
- **Mohammed VI Library of Tangier (Quality = 0.983)**
- **Bibliothèque Al Manar de Fez (Quality = 0.973)**
- **Ibn Battuta Library of Marrakech (Quality = 0.975)**
- **Dar Al Hikma Library of Tangier (Quality = 0.975)**
- **Al Andalus Library of Meknes (Quality = 0.975)**
- **Al Quaraouiyine Library of Meknes (Quality = 0.975)**

Interpretation: These libraries have a representation quality greater than 0.97 on the PC1-PC2 plane, which means that they can be interpreted unambiguously in this 2D projection. They are well explained by the two principal dimensions.

The individuals poorly represented on the map (PC1 + PC2) are:

- **Dar Al Hikma Library in Rabat (Quality = 0.148)**
- **Hassan II Library of Oujda (Quality = 0.128)**
- **Al Andalus Library of Safi (Quality = 0.124)**
- **Al Nour Library in Casablanca (Quality = 0.124)**
- **Al Manar Library in Marrakech (Quality = 0.092)**
- **Bibliothèque Al Fikr de Marrakech (Quality = 0.184)**
- **Dar Al Hikma Library of Agadir (Quality = 0.380)**

Interpretation: These libraries have a representation quality of less than 0.40, indicating that they are not well represented in the PC1-PC2 plane. Their positioning in this plane is unreliable and would require examination of additional axes (such as PC3) for proper interpretation.

8- Contribution of individuals and variables

8-1- Contribution of variables

Here is the table showing the contribution of the variables:

Table 5: Contribution of variables

Variable	PC1	PC2
Books	16.98	0.00
Places	18.92	0.01
Visitors_day	12.94	0.02
Temperature_C	0.09	49.82
Humidity_%	0.00	50.11
Personnel	18.89	0.00
Budget_k€	19.36	0.03
Opening_h_d	12.82	0.00

The variables that**contribute the most**PC1 includes:

- **Budget_k€** (19.36%)
- **Places** (18.92%)
- **Personnel** (18.89%)
- **Books** (16.98%)
- **Visitors_day** (12.94%)
- **Opening_h_d** (12.82%)

Interpretation: PC1 clearly represents the **dimension "Size and Resources"** Libraries. This axis is primarily defined by structural and economic variables: libraries with a large budget also have more space, more staff, more books, more visitors, and longer opening hours. The near-zero contribution of Temperature (0.09%) and Humidity (0.00%) confirms that PC1 is independent of environmental conditions.

The variables that contribute significantly to PC2 are:

- **Humidity_%** (50.11%)
- **Temperature_C** (49.82%)

Other variables contribute negligibly (< 0.03%).

Interpretation: PC2 represents almost exclusively the **Environmental dimension**This axis is entirely driven by the internal climatic conditions of libraries (temperature and humidity). The fact that all other variables have a near-zero contribution means that PC2 is uncorrelated with organizational, structural, or economic characteristics.

8-2- Contribution of individuals

The individuals who contribute significantly to PC1 are:

- **Ibn Khaldoun Library of Tangier** : 3.63%
- **Mohammed VI Library of Tangier** : 3.53%
- **Library Dar Al Hikma of Oujda** : 3.53%
- **Al Nour Library of Agadir**: 3.86% (the largest contribution to PC1)
- **Mohammed VI Library of Oujda** : 3.67%
- **Al Manar Library in Fez** : 2.94%
- **Dar Al Hikma Library of Tangier** : 2.88%
- **Ibn Khaldoun Library of Marrakech** : 2.87%

Interpretation: These libraries strongly "pull" the PC1 axis. They represent institutions with extreme characteristics in terms of size and resources. Since PC1 is the "Size/Resources" dimension axis, these libraries are either very large and well-equipped (high book collections, ample space, staff, and budget), or conversely, very Small and poorly funded, but in both cases they define the extremities of this axis. The libraries of Tangier**,** Oujda and Agadir are particularly influential here, suggesting that they concentrate the most marked deviations from the national average in terms of resources.

The individuals who contribute significantly to PC2 are:

- **Mohammed VI Library of Oujda**4.76% (absolute record)
- **Dar Al Hikma Library of Tetouan** : 4.58%
- **Al Fikr Library in Fez** : 4.68%
- **Al Quaraouiyine Library of Agadir** : 4.14%

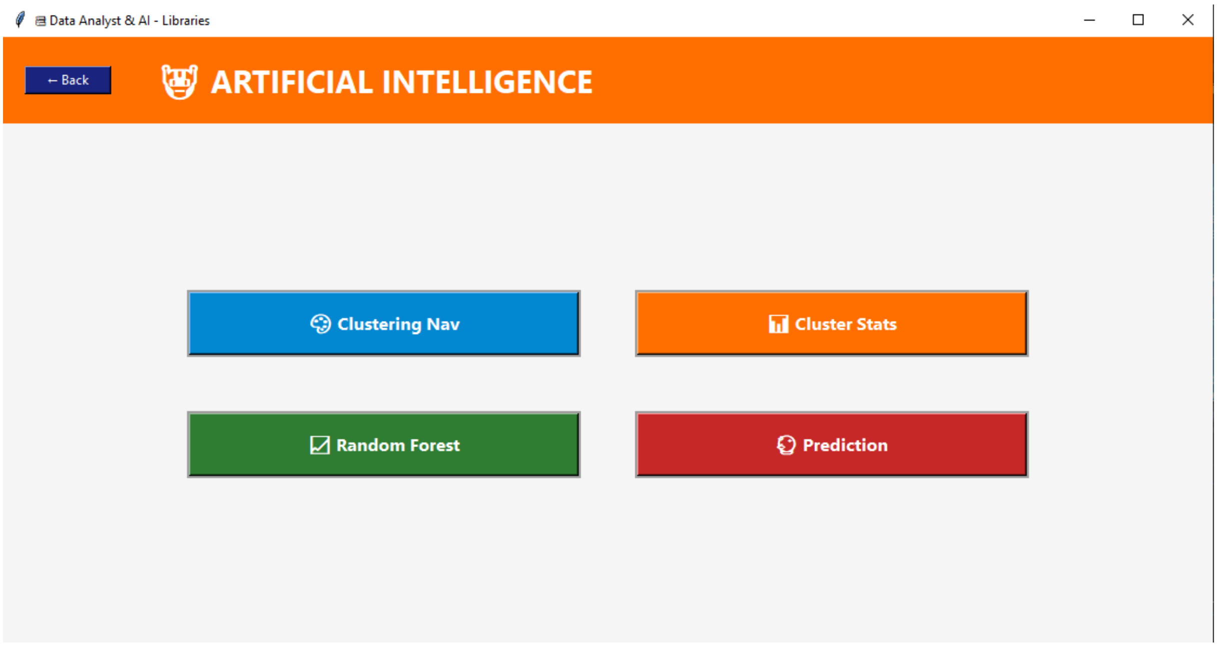
- **Al Manar Library in Agadir** : 3.94%
- **Al Quaraouiyine Library of Oujda** : 3.89%
- **Al Fikr Library in Agadir** : 3.58%
- **Al Nour Library of Agadir** : 3.63%

Interpretation: These libraries define the poles of the PC2 axis, which represents the Environmental/Climatic dimension. They are characterized by temperature and/or humidity conditions far removed from the norm. This can indicate either exceptionally well-controlled indoor environments (ideal for preservation) or, conversely, atypical or problematic conditions. The libraries of Agadir**,** Oujda and Tetouan are the most represented, which could reflect specific regional climate constraints or strategies, or simply different regulatory systems. They are the most decisive in explaining the variations in conservation conditions between libraries.

Chapter II: Library segmentation by clustering (k=3 to k=7) and Random Forest modeling

This chapter uses the results of PCA to construct a classification of Moroccan libraries. It combines two approaches: first, aunsupervised clustering(testing 3 to 7 groups) to identify similar profiles, then the algorithm Random Forest to validate this classification, here is the interface for the AI section.

Figure 6: Artificial Intelligence Interface



I- Clusturing (k=3 to k=7)

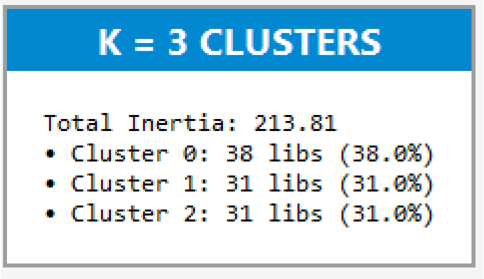
The clustering (or "grouping" in French) is an unsupervised machine learning technique that consists of automatically group individuals (here, the libraries) inhomogeneous groups called clusters, depending on their similarities.

Next, we will talk about clusters for values from k = 3 up to k = 7.

1- Clusturing K-means (k=3)

Regarding the statistics for this cluster, see below:

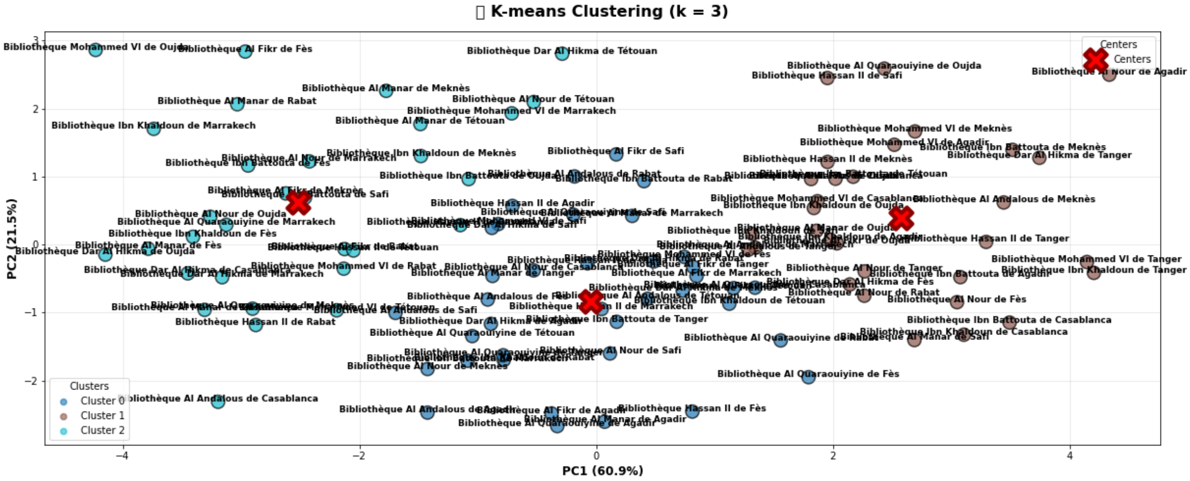
Figure 7: Clustering statistics k = 3



K = 3 CLUSTERS

```
Total Inertia: 213.81
• Cluster 0: 38 libs (38.0%)
• Cluster 1: 31 libs (31.0%)
• Cluster 2: 31 libs (31.0%)
```

Figure 8: k-means pour k = 3

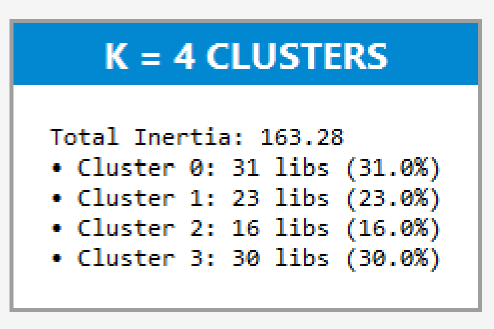


We can note for the **cluster 0**We have 38 libraries, the **cluster 1 with** 31 libraries and cluster 2 (31 libraries)

2- Clusturing K-means (k=4)

Regarding the statistics for this clustering, see below:

Figure 9: Clustering statistics k = 4

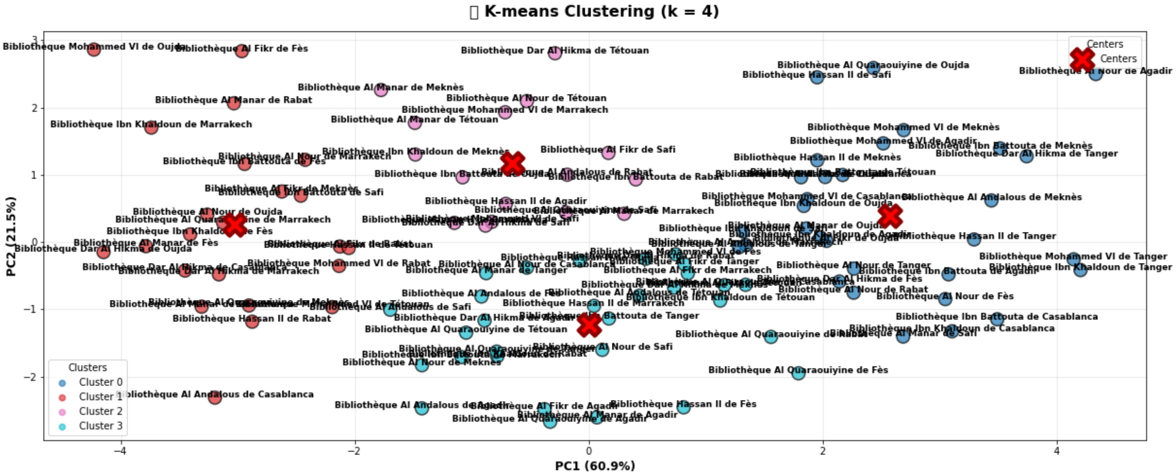


K = 4 CLUSTERS

```
Total Inertia: 163.28
• Cluster 0: 31 libs (31.0%)
• Cluster 1: 23 libs (23.0%)
• Cluster 2: 16 libs (16.0%)
• Cluster 3: 30 libs (30.0%)
```

Here is the graph representing the clustering k = 4:

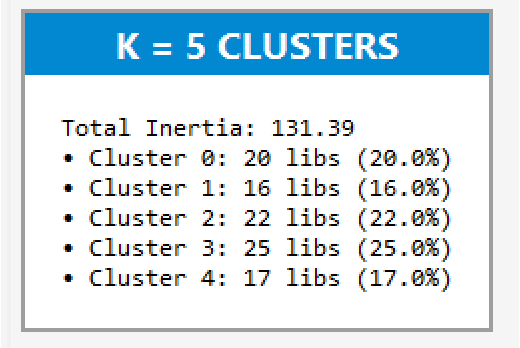
Figure 10: k-means pour k = 4



3- Clusturing K-means (k=5)

Regarding the statistics for this clustering, see below:

Figure 11: Clustering statistics k = 5

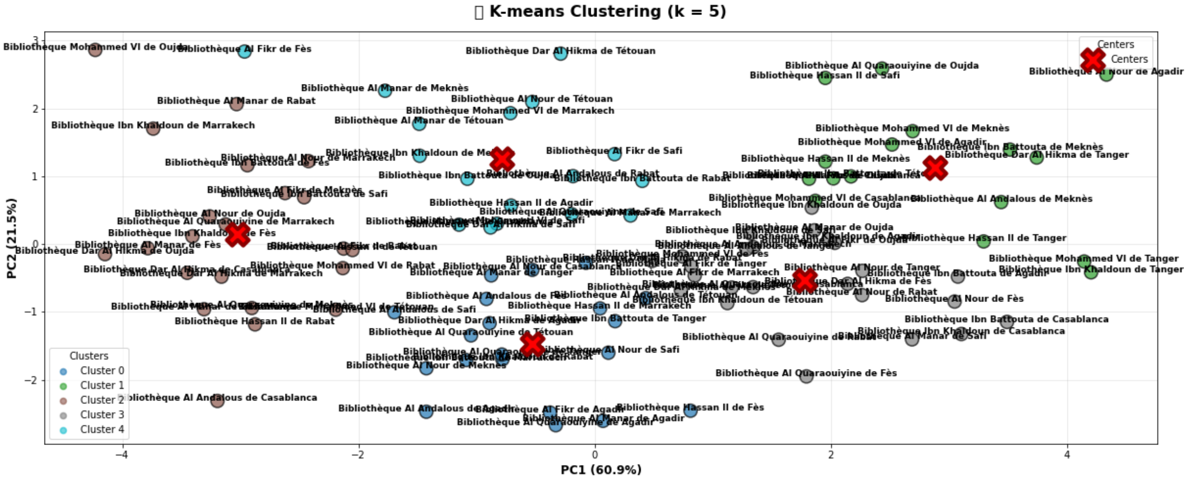


K = 5 CLUSTERS

```
Total Inertia: 131.39
• Cluster 0: 20 libs (20.0%)
• Cluster 1: 16 libs (16.0%)
• Cluster 2: 22 libs (22.0%)
• Cluster 3: 25 libs (25.0%)
• Cluster 4: 17 libs (17.0%)
```

Here is the graph representing the clustering k = 5:

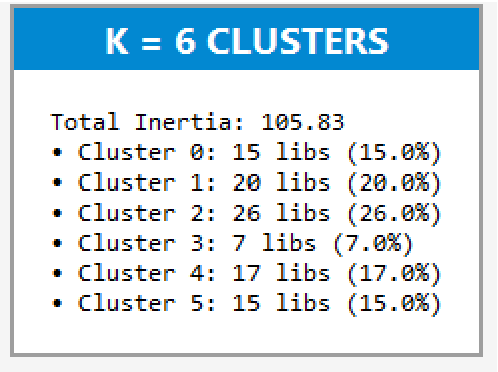
Figure 12: k-means pour k = 5



4- Clusturing K-means (k=6)

Regarding the statistics for this clustering, see below:

Figure 13: Clustering statistics k = 6

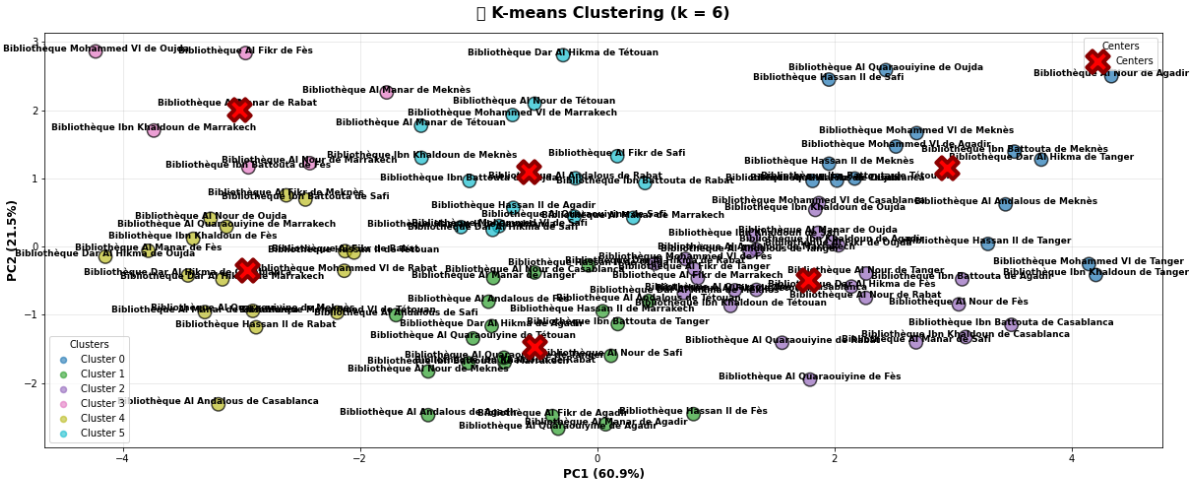


K = 6 CLUSTERS

```
Total Inertia: 105.83
• Cluster 0: 15 libs (15.0%)
• Cluster 1: 20 libs (20.0%)
• Cluster 2: 26 libs (26.0%)
• Cluster 3: 7 libs (7.0%)
• Cluster 4: 17 libs (17.0%)
• Cluster 5: 15 libs (15.0%)
```

And here is the graph representing the clustering k = 6:

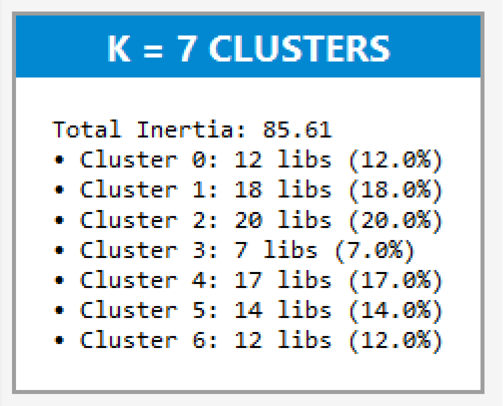
Figure 14: k-means pour k = 6



5- Clusturing K-means (k=7)

Regarding the statistics for this clustering, see below:

Figure 15: Clustering statistics k = 7

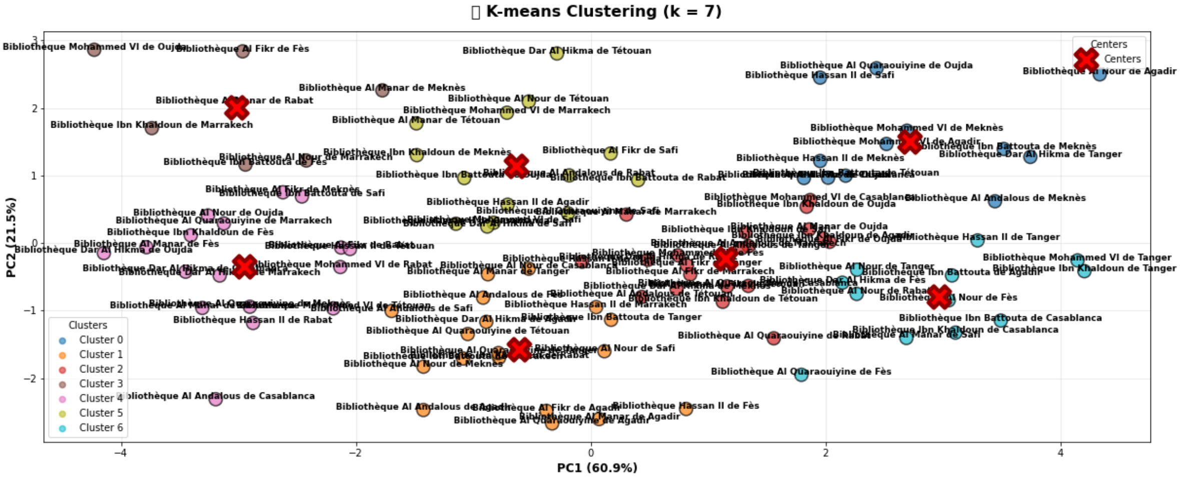


K = 7 CLUSTERS

```
Total Inertia: 85.61
• Cluster 0: 12 libs (12.0%)
• Cluster 1: 18 libs (18.0%)
• Cluster 2: 20 libs (20.0%)
• Cluster 3: 7 libs (7.0%)
• Cluster 4: 17 libs (17.0%)
• Cluster 5: 14 libs (14.0%)
• Cluster 6: 12 libs (12.0%)
```

And here is the graph representing the clustering k = 7:

Figure 16: k-means pour k = 7



II- Random Forest Modeling

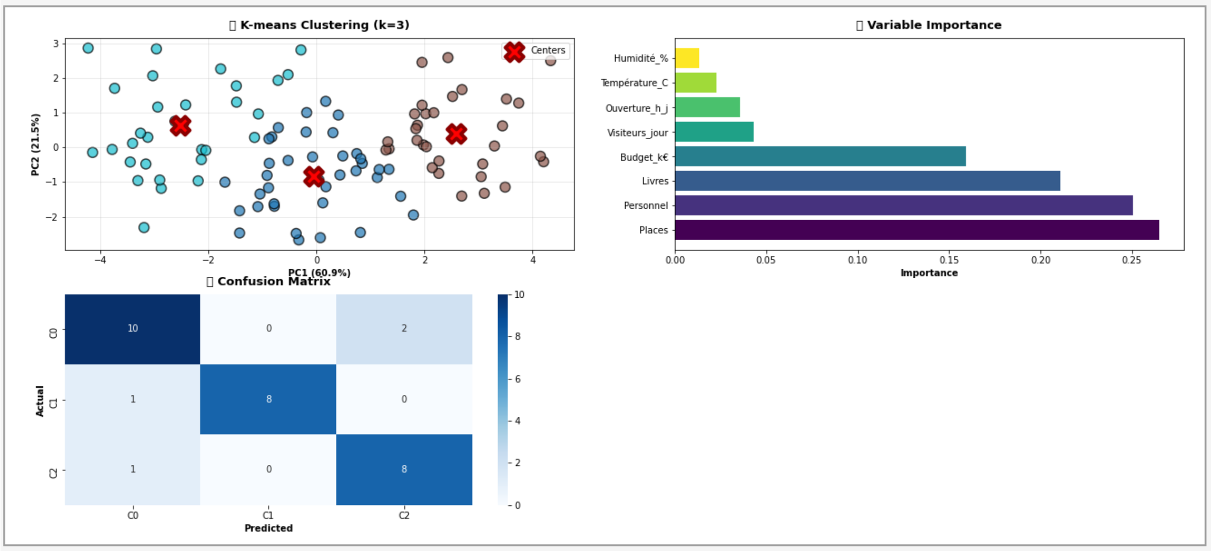
The **Random Forest** (Random Forest) is a machine learning algorithm supervised which builds a multitude of decision trees during training and combines their predictions to improve accuracy and control overfitting.

In our dataset (Libraries), we applied random forest and clustering to predict clusters. The various results we obtained will be presented and interpreted later.

1- Random Forest Result

Regarding the random forest modeling on our dataset, here are the results obtained:

Figure 17:Random Forest Results



The diagram above shows the importance of the variables used by the **Random Forest Classifier** to predict cluster membership.

The three most important variables are: Seats (≈ 0.27), Staff (≈ 0.25), and Books (≈ 0.22). This trio corresponds to major structural indicators: capacity, staffing levels, and documentary resources. The Budget_k€ comes next, which is consistent because it influences the first three. Visitors_day, Opening_h_j, Temperature_C, Humidity_% has a smaller impact on the prediction.

Therefore, we can conclude that the classification is primarily determined by resources. Internal library factors (spaces, staff, books). Climatic variables contribute little to distinguishing the clusters.

For the confusion matrix, it compares the predictions of the Random Forest to the actual K-means clusters**.**

We can see that:

- **C0**is very well recognized:
  - 10 libraries correctly classified, 2 incorrectly classified.
- **C1**is almost perfect:
  - 8 correctly ranked, 1 error.
- **C2**is also performing well:
  - 8 correctly ranked, 1 error.

The model makes very few errors, which shows a good natural separation of clusters, a good generalization capacity of the Random Forest.

2- Report by Cluster

Table 6: Report by cluster

Class	Precision	Recall	F1-score	Support
Cluster_0	0.833	0.833	0.833	12
Cluster_1	1.000	0.889	0.941	9
Cluster_2	0.800	0.889	0.842	9
Accuracy	0.867	0.867	0.867	0.867
Macro_avg	0.878	0.870	0.872	30
Weighted_avg	0.873	0.867	0.868	30

Cluster_1, which is theThe best recognized. Here's what we can say:

- **Precision** = 1,000 → *no false positives*: when it predicts C1, the model is never wrong.

- **Recall** = 0,889 → *only one false negative* out of 9 true C1s (≈ 8/9 correctly detected).
- **F1**= 0,941→ excellent combination of accuracy/recall.

C1 has a very clear signature for the model. It's just missing a small number of genuine, unrecognized C1s.

Cluster_0, which exhibits p stable and balanced performance with:

- **Precision =** 0,833 And **Recall =** 0,833 → approximately10 genuine positives**,** 2 false positives and 2 false negatives (out of 12).
- F1 = 0.833→ Class correctly captured, but with some confusion.

So C0 is clearly identified but there is still a little bit of noise.

And finally, cluster_2, which presents p more incoming confusion with:

- **Precision** = 0,800→ there is false positives **(**other predicted classes C2).
- **Recall =** 0,889 → 1 false negative out of 9 (≈ 8/9 correctly detected).
- **F1 = 0.842**→ slightly lower than C0/C1.

So C2 is often recognized, more absorbs samples from other classes (incorrectly predicted C2), hence the lower accuracy.

3- Global metrics

Tableau 7: Metrics

Metric	Value
Train Accuracy	0.943
Test Accuracy	0.867
Precision	0.873
Recall	0.867

F1-Score	0.868
Train-Test Gap	0.076
K-means Inertia	213.806

Train Accuracy (0,943): very good accuracy on the train, but something to keep an eye on to prevent the model from memorizing too many specific patterns.

Test Accuracy (0,867): good out-of-sample performance, in line with what you observed in cross-validation (≈ 0.90 average).

Precision (0,873): when the model announces a class, it is correct ~87% of the time → few false positives.

Recall (0,867): it retrieves ~87% of the real examples, which implies few false negatives.

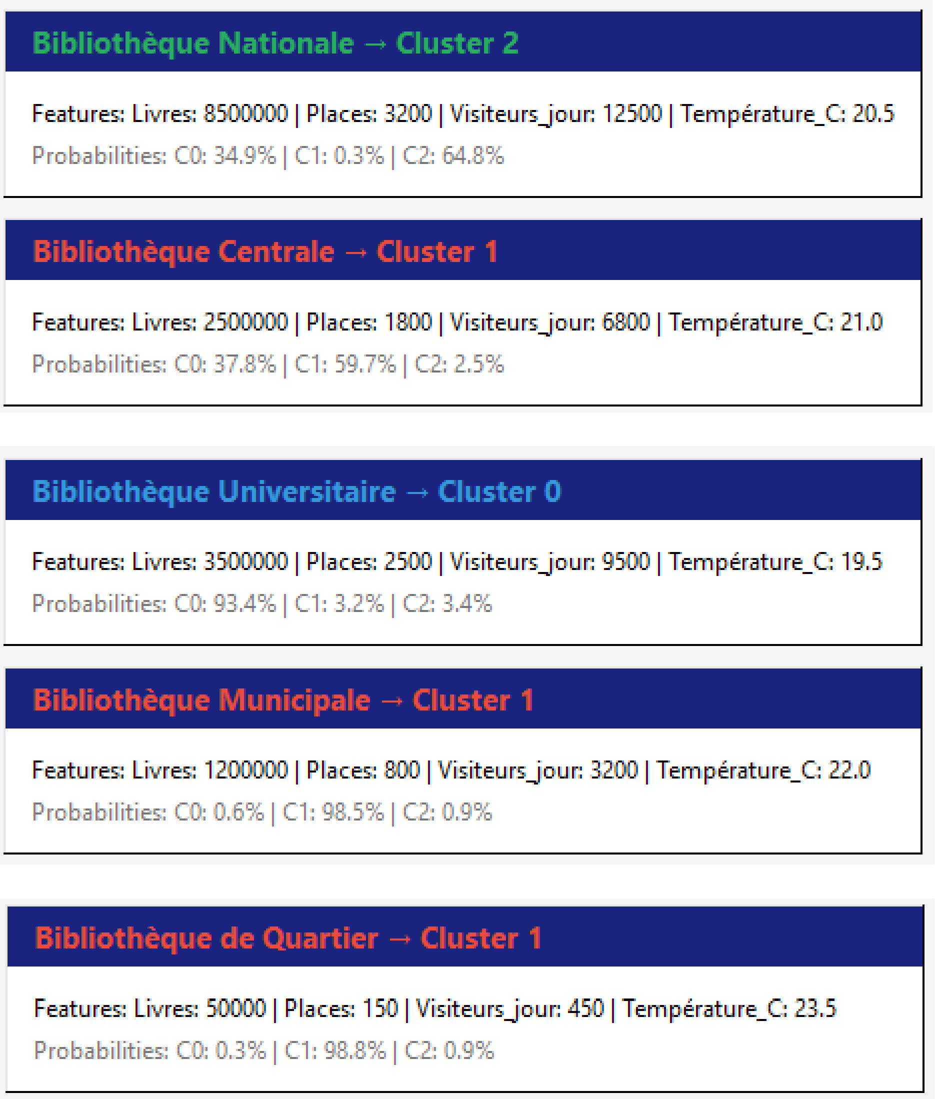
F1-Score (0,868): balance between **precision** and **recall**; no clear bias towards one class.

Train–Test Gap (0,076): measured gap, suggesting strengthening the regularization and/or stability of folds (stratification, more estimators, etc.).

4- Prediction of new individuals

To test our model, we will assign it unknown individuals (libraries) so that it can predict the clusters. In our case, we provided 5 libraries, and here are the results we obtained below:

Figure 18: Prediction of new libraries



Bibliothèque Nationale → Cluster 2

Features: Livres: 8500000 | Places: 3200 | Visiteurs_jour: 12500 | Température_C: 20.5

Probabilities: C0: 34.9% | C1: 0.3% | C2: 64.8%

Bibliothèque Centrale → Cluster 1

Features: Livres: 2500000 | Places: 1800 | Visiteurs_jour: 6800 | Température_C: 21.0

Probabilities: C0: 37.8% | C1: 59.7% | C2: 2.5%

Bibliothèque Universitaire → Cluster 0

Features: Livres: 3500000 | Places: 2500 | Visiteurs_jour: 9500 | Température_C: 19.5

Probabilities: C0: 93.4% | C1: 3.2% | C2: 3.4%

Bibliothèque Municipale → Cluster 1

Features: Livres: 1200000 | Places: 800 | Visiteurs_jour: 3200 | Température_C: 22.0

Probabilities: C0: 0.6% | C1: 98.5% | C2: 0.9%

Bibliothèque de Quartier → Cluster 1

Features: Livres: 50000 | Places: 150 | Visiteurs_jour: 450 | Température_C: 23.5

Probabilities: C0: 0.3% | C1: 98.8% | C2: 0.9%

Above we can see 5 libraries that have been predicted, along with their different probabilities.

Chapter IV: Analysis of the weighting data of the 8 variables for all individuals and application of Correspondence Factor Analysis (CFA)

This chapter is dedicated to the in-depth analysis of the weighting data relating to the eight variables studied for all individuals. After the data collection and preparation phase, it becomes essential to explore the existing relationships between individuals and variables in order to better understand the overall structure of the dataset and to identify the main profiles that emerge from it.

The application of correspondence analysis in this chapter aims to analyze the contribution of the eight variables to the structuring of the factor space, to study the quality of representation of individuals, and to interpret the principal factor axes from a contingency table.

Regarding our contingency table with a weighting between 1 and 10 for the variables for 8 individuals and the 8 criteria, here it is below:

Table 8: Contingency Table

Library	Books	Places	Visitors_day	Personnel	Budget_k€	Opening_h/d	Temperature_C	Humidity_%
BNRM	9	8	7	9	9	8	7	6
Hassania Health Library	6	5	6	7	8	7	8	7
Mohammed VI University Library	7	7	5	6	6	6	6	5
Villa des Arts Library	5	4	8	5	7	9	9	8
Moulay El Hassan Library	8	6	9	8	8	8	8	7
Abdelkrim Khattabi Library	6	5	4	5	5	5	5	4
Lalla Hasnaa Library	7	7	6	7	6	7	7	6
Ibn Zohr Library	5	6	5	5	5	6	9	9

I- Frequency matrix

In AFC, the **frequency matrix** (or correspondence matrix) is obtained by dividing each element of the contingency table by the total sum of all the elements.

Either *N* The grand total:

$$N = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$$

The frequency matrix *F* East :

$$f_{ij} = \frac{x_{ij}}{N}$$

Here is the frequency matrix corresponding to our contingency table:

Tableau 9:Frequency Matrix

Library	Books	Places	Visitors_day	Personnel	Budget_k€	Opening hours/day	Temperature_C	Humidity_%
BNRM	0.021	0.019	0.016	0.021	0.021	0.019	0.016	0.014
Hassania Health Library	0.014	0.012	0.014	0.016	0.019	0.016	0.019	0.016
Mohammed VI University Library	0.016	0.016	0.012	0.014	0.014	0.014	0.014	0.012
Villa des Arts Library	0.012	0.009	0.019	0.012	0.016	0.021	0.021	0.019
Moulay El Hassan Library	0.019	0.014	0.021	0.019	0.019	0.019	0.019	0.016
Abdelkrim Khattabi Library	0.014	0.012	0.009	0.012	0.012	0.012	0.012	0.009

Lalla Hasnaa Library	0.016	0.016	0.014	0.016	0.014	0.016	0.016	0.014
Ibn Zohr Library	0.012	0.014	0.012	0.012	0.012	0.014	0.021	0.021

The **frequency matrix** is crucial in AFC because it standardizes data to eliminate the effect of marginal totals, allowing comparison of relative profiles.

II- Chi-square (χ^2) analysis

1- Definitions

The test of **Khi- square (χ^2)** The independence test assesses whether there is a significant association between two categorical variables in a contingency table. He compares the observed numbers to theoretical numbers (expected under the assumption of independence).

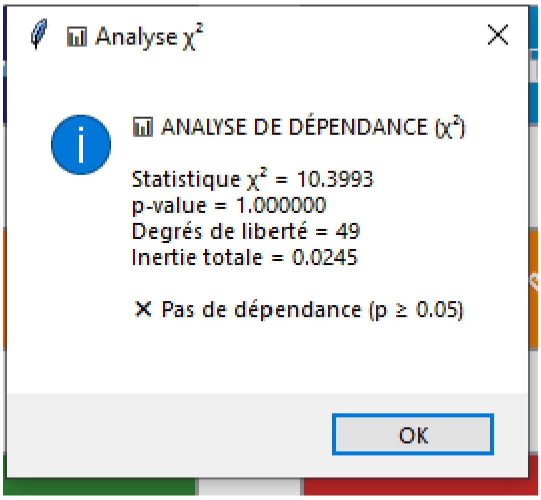
The chi-square test allows us to:

- Detect a significant relationship between the rows and columns of the table.
- To be the preliminary step before a Correspondence Factor Analysis **(AFC)**.
- **Total inertia**: which measures the overall intensity of the dependence (value between 0 and 1).

2- Interpretation

The results obtained in this case are shown below:

Figure 19: Result of the Chi-square (χ^2) analysis



ANALYSE DE DÉPENDANCE (χ^2)

Statistique χ^2 = 10.3993
p-value = 1.000000
Degrés de liberté = 49
Inertie totale = 0.0245

✕ Pas de dépendance ($p \geq 0.05$)

The **degrees of freedom (df)** correspond to the number of independent cells in the contingency table that can vary freely while respecting the marginal totals (sums of rows and columns). It is calculated using the formula (I−1)×(J−1)(I−1) \times (J−1)(I−1)×(J−1), where

III represents the number of rows and JJJ the number of columns. In our case, the table contains 8 libraries and 8 variables, resulting in (8−1)×(8−1)=49(8−1) \times (8−1) = 49(8−1)×(8−1)=49 degrees of freedom. A high number of degrees of freedom implies that the χ^2 distribution is relatively spread out, thus requiring a large χ^2 value to obtain a statistically significant result.

Next, **p-value** represents the probability of observing a deviation from independence at least as large as that measured ($\chi^2 = 10.3993$), under the null hypothesis of independence between rows and columns. A p-value less than 0.05 generally leads to the rejection of this hypothesis. In our case, the p-value is 1.000, indicating that the observed χ^2 value is perfectly consistent with the hypothesis of independence. In other words, the library profiles according to the different variables exhibit very low heterogeneity, making it impossible to identify a statistically significant relationship between rows and columns.

Total inertia, also called ϕ2\phi^2ϕ2, it is defined as the ratio of χ^2 to the total number of data points (NNN). It measures the overall strength of the association between rows and columns, independent of the sample size. Its value ranges from 0 (perfect independence) to a maximum that depends on the size of the table. In this study, the total inertia is 0.0245, a value very close to zero, confirming the weakness of the overall data structure. In Correspondence Analysis, this inertia is distributed across the factorial axes, meaning that the structured information to be explained is relatively limited. Correspondence Analysis is thus used here as a descriptive method for analyzing the relative profiles of libraries and variables, rather than as an inferential testing tool.

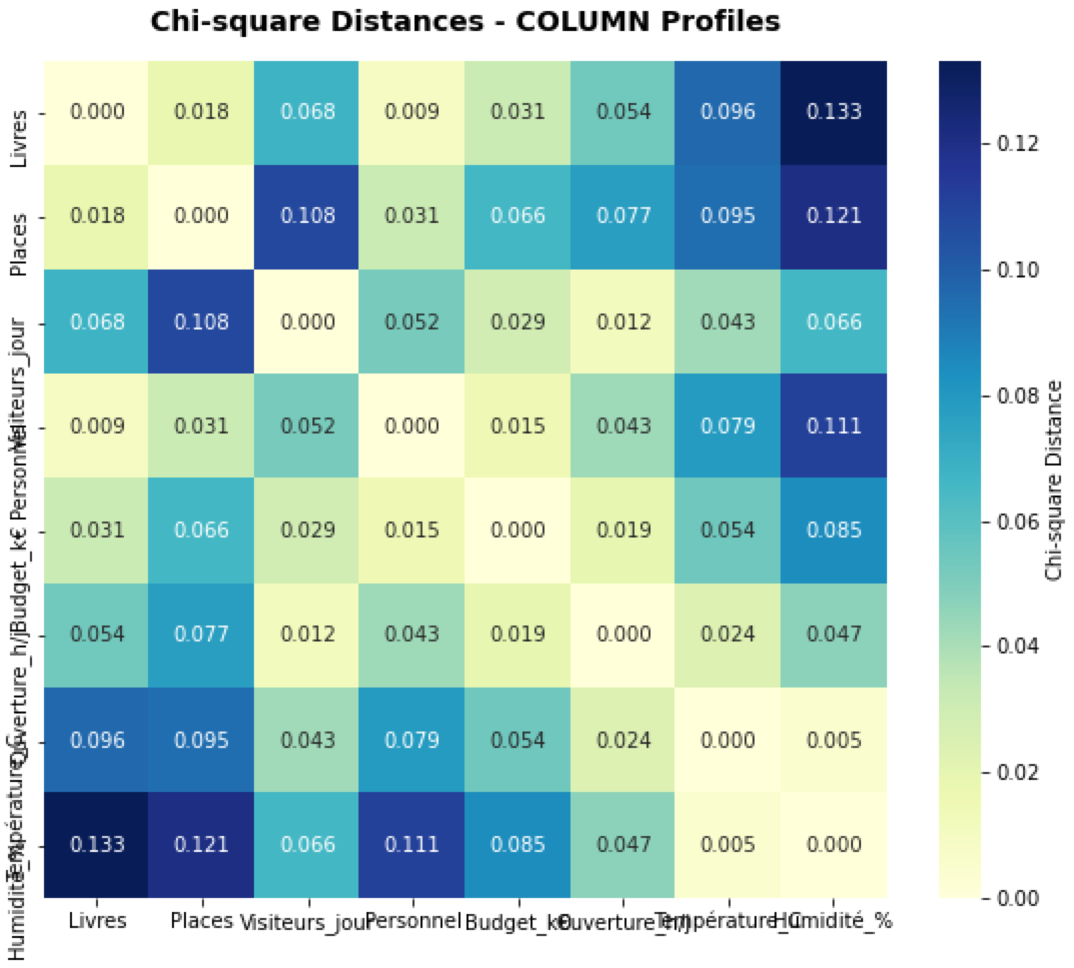
III- Distance Khi-deux (χ^2)

The **Chi-square distance** is a measure of dissimilarity specifically designed to compare profiles (relative distributions) in a contingency table. It measures the difference between two profiles, taking into account the marginal weights columns.

1- Distance Chi-square (χ^2) column profiles

Here is the heatmap of the chi-square distance on columns:

Figure 20: Chi-square (χ^2) distance between column profiles



We can see that the most similar variables are:

Temperature_C ↔ Humidity_% : **distance = 0.005**(very low). These two environmental variables vary in exactly the same way according to the libraries. Probably physically correlated.

Furthermore, the following variables are quite similar:

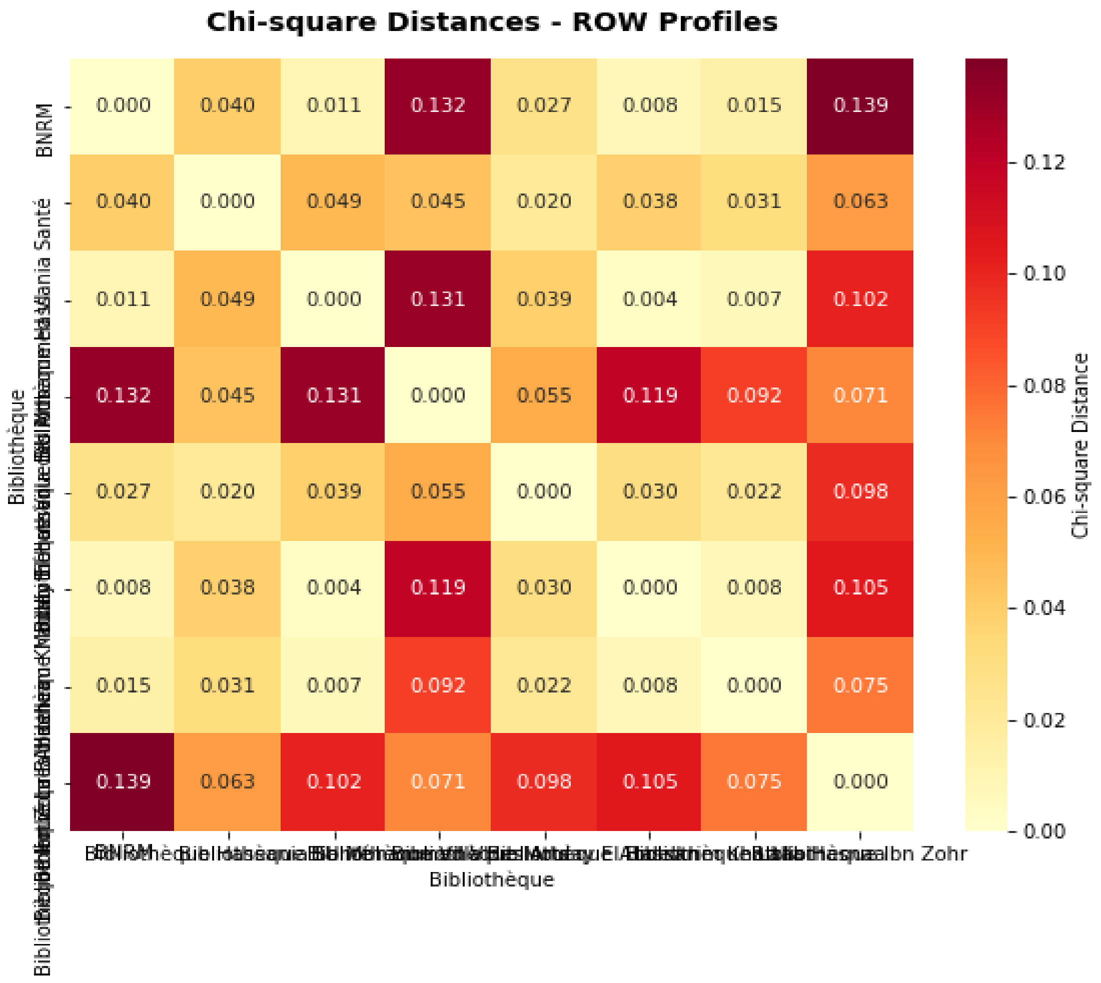
- Visitors_day ↔ Opening_h/d : **distance = 0.012** : The number of visitors and opening hours have similar profiles.
- Personnel ↔ Budget_k€ : **distance = 0.015** Staff and budget are logically linked.
- Books ↔ Personal : **distance = 0.009**(very low): Libraries with more books also have more staff.
- Books ↔ Seats : **distance = 0.018**: the capacity (books and seats) is consistent.

2- Distance Chi-square (χ^2) profile lines

Analyzing the χ^2 distances between libraries allows us to identify levels of similarity and differentiation between their profiles.

Here is the heatmap of the chi-square distance on rows:

Figure 21: Chi-square (χ^2) distance of line profiles



- **Libraries very similar (almost identical profiles)**

The smallest distances are observed between the Mohammed VI University Library and the Abdelkrim Khattabi Library(distance = 0.004). This minimum value indicates that these two libraries have almost identical profiles across all the variables studied, reflecting a very similar structure and operation.

Short distances are also observed between the BNRM and the Abdelkrim Khattabi Library (distance = 0.008), as well as between the BNRM and the Mohammed VI University Library (distance = 0.011). This shows that the BNRM shares a profile similar to these libraries, particularly in terms of resources and general organization.

Furthermore, the Lalla Hasnaa Library is very close to the Abdelkrim Library Khattabi (distance = 0.008) and the Mohammed VI University Library (distance = 0.007). These libraries thus form a homogeneous core characterized by very similar profiles.

- **Libraries with moderate proximity**

The Moulay El Hassan Library is moderately close to the Hassania Health Library (distance = 0.020) and with the Lalla Hasnaa Library (distance = 0.022). These distances indicate that Moulay El Hassan occupies an intermediate position, with a profile relatively close to that of balanced functioning libraries, without being strictly identical.

- **Remote libraries (strong oppositions)**

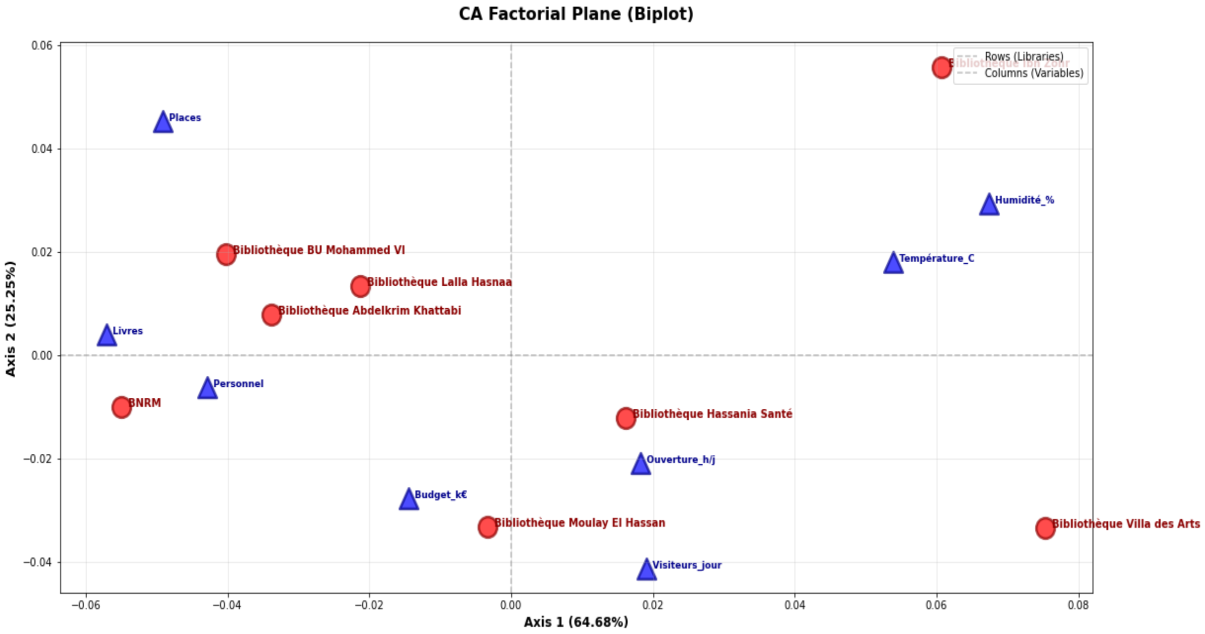
Large distances are observed between the BNRM and the Villa des Arts Library (distance = 0.132), as well as between the Mohammed VI University Library and the Villa Library Arts Library (distance = 0.131). These results show that the Villa des Arts Library presents a clearly distinct profile, in contrast to libraries with a structural dominance.

The maximum distance is observed between the Ibn Library Zohr and the BNRM (distance = 0.139). A similarly high distance is observed between the Ibn Zohr Library and the Abdelkrim Khattabi Library (distance = 0.105). This indicates that the Ibn Zohr Library stands out significantly from other libraries, particularly due to specific characteristics related to its use and comfort.

IV- Factorial Plan

The factorial plane simultaneously represents the **libraries (lines)**and the **variables (columns)** on the first two factorial axes.
Axis 1 explains**64,68 %**of the total inertia, while axis 2 explains**25,25 %**Together, these two axes restore **nearly 90% of the information.** This allows for a reliable interpretation of the plan. See the figure below.

Figure 22: Factorial plane



1- Interpretation of axis 1 (PC1)

Axis 1 is the main axis for structuring the data.

To the right of axis 1 are the variables **Temperature_C** And **Humidity_%**, as well as libraries **Villa des Arts** and **Ibn Zohr**.

HAS To the left of axis 1 are the variables **Books**, **Places** and **Personnel** accompanied by libraries **BNRM**, **BU Mohammed VI**, **Abdelkrim Khattabi** and **Lalla Hasnaa**. So we can say that axis 1 opposes libraries oriented towards the **environmental comfort** to those characterized by **structural and human resources**.

2- Interpretation of axis 2 (PC2)

The variables that appear at the top of axis 2 are the most prominent. **Places**, **Temperature_C** And **Humidity_%**, as well as libraries **BU Mohammed VI**, **Lalla Hasnaa** and **Ibn Zohr**.

The variables are positioned at the bottom of axis 2. **Visitors_day**, **Opening hours/day** And **Budget_k€**, with the libraries **Moulay El Hassan**, **Villa des Arts** and **Hassania Health**.

We can say that axis 2 distinguishes libraries from **capacity- and physical condition-oriented organization** of those more closely related to the **dynamics of use and openness**.

3- Library placement

BNRM: located to the left of the plan, close to the variables **Books** and **Personnel.** It is characterized by a strong focus on structural resources.

BU Mohammed VI, **Abdelkrim Khattabi** and **Lalla Hasnaa** Close to each other, they present similar and balanced profiles, dominated by resources and reception capacity.

Moulay El Hassan And **Hassania Health**: intermediate position, associated with the variables **Budget_k€**, **Opening_h/d** And **Visitors_day**, reflecting an orientation towards operation and activity.

Villa des Arts and **Ibn Zohr**: located to the right of the plane, strongly associated with the variables **Temperature_C** and **Humidity_%** they present profiles **atypical**, focused on comfort and conditions of use.

4- Reading rows and columns together

The proximity between a library and a variable indicates a strong association. So:

- **Villa des Arts** and **Ibn Zohr** are strongly associated with environmental comfort.
- **BNRM**is primarily associated with human and documentary resources**.**
- **Moulay El Hassan** is associated with attendance and openness.

Chapter IV: Cybersecurity

The first three chapters were devoted to analyzing the network of Moroccan libraries using different statistical methods: PCA for exploring underlying structures, clustering for unsupervised classification, and Random Forest for predictive modeling. These approaches, although specific to the library field, are based on methodological foundations. general principles that can be transposed to other contexts.

This fourth and final chapter aims precisely to demonstrate this transferability methodologically, by applying similar techniques to a radically different field: cybersecurity. We will show how the data analysis skills acquired on a concrete case (libraries) can be reinvested to address crucial contemporary problems, such as the detection of intrusions and malicious activities in computer systems.

We will rely on a cybersecurity-oriented dataset with 20 libraries and the following 8 criteria:

- Number of PCs: Number of computer workstations
- Firewall_Active: Indicator 1/0 if the firewall is active
- Server Update: % of servers up to date (0–100)
- Public Network Access: Number of Wi-Fi/public access points
- Antivirus_Installed: Number of workstations with antivirus software
- Phishing Attempts: Number of phishing attempts detected this month
- Logs_Analyzes: % of logs analyzed regularly
- Security Incidents: Number of security incidents recorded

To achieve this, we will implement two specialized anomaly detection algorithms: **Isolation Forest** and the **Local Outlier Factor (LOF)**These methods, although conceptually distinct from the clustering applied previously, share the same logic: that of identify what deviates from the norm in a dataset.

I- Isolation Forest

1- Definition and importance

Isolation Forest (Isolation Forest) is an unsupervised machine learning algorithm specifically designed for anomaly detection .Unlike traditional methods that model normal behavior, Isolation Forest is based on a simple but powerful idea: anomalies are rare and different, therefore easier to isolate.

Forest Isolation is important because:

- **Computational efficiency** Linear complexity: O(n), Fast on large datasets, Less memory-intensive
- **No need for strict standardization** Works well with different scales, less sensitive to outliers than other methods
- **Ideal for tall dimensions:** Performs well even with many variables; does not suffer from the "scourge of dimensionality".
- **Practical applications:** Fraud detection, network monitoring, medical diagnostics, predictive maintenance

Below is the Forest isolation table applied to our dataset:

Table 10: table isolation forest

Library	Status	Anomaly_Score	Possible_Reason
Hassania Health	ANOMALY	-0.587	Normal configuration
Agadir South	NORMAL	-0.551	Normal configuration
Meknes West	ANOMALY	-0.551	Normal configuration
Tangier Nord	NORMAL	-0.551	Normal configuration

BNRM	NORMAL	-0.536	Normal configuration
Casablanca Centre	NORMAL	-0.533	Normal configuration
North Nador	NORMAL	-0.527	Normal configuration
Discount City	NORMAL	-0.517	Normal configuration
Fes Medina	NORMAL	-0.508	Normal configuration
Abdelkrim Khattabi	NORMAL	-0.504	Normal configuration
Larache City	NORMAL	-0.503	Normal configuration
Kenitra Ville	NORMAL	-0.498	Normal configuration
Tetouan City	NORMAL	-0.490	Normal configuration
Marrakech Ville	NORMAL	-0.489	Normal configuration
Mohammed VI	NORMAL	-0.487	Normal configuration
Sidi Kacem West	NORMAL	-0.486	Normal configuration

Taza Centre	NORMAL	-0.482	Normal configuration
Oujda East	NORMAL	-0.475	Normal configuration
Safi Centre	NORMAL	-0.474	Normal configuration
The South Jadida	NORMAL	-0.457	Normal configuration

2- Interpretation

The Isolation Forest algorithm above analyzed 20 libraries and identified 2 anomalous cases. based on their safety indicators, i.e. 10% of the sample. The majority of establishments (18 out of 20) have security profiles that comply with the expected standards.

The libraries detected as abnormal are:

- **"Hassania Santé" Library** with a score of -0.587 (the lowest in the dataset), this library exhibits significantly different security indicators from the rest. Its very low score suggests that it is "easy to isolate" from the others, a typical characteristic of outliers. This could indicate either serious vulnerabilities or, conversely, an exceptionally strict security policy that sets it apart.

- **"Meknes West" Library**: with a Score: -0.551 (close to the detection threshold), this second library reported, with a score slightly higher than "Hassania Santé". Its borderline position suggests moderate deviations from standards, requiring investigation but not necessarily a critical emergency.

For standard libraries we have:

- **Libraries that are very compliant (scores > -0.50)** : **The South Jadida**(-0.457) is the best score, very standard configuration, **Oujda East**(-0.475) and **Safi Centre**(-0.474): Very typical profiles. These libraries represent the core of the "normal" behavior of the network.

- **Libraries with moderate compliance** (scores -0.50 to -0.52): **Discount City** (-0.517), **BNRM** (-0.536), **Casablanca Centre**(-0.533): Slightly below average. They represent the silent majority of the network.
- **Libraries operating at the limit of normality** (scores < -0.52) : **Tangier Nord** (-0.551), **Agadir South** (-0.551).

II- Local Outlier Factor (LOF)

1- Definitions and importance

The **Local Outlier Factor (LOF)**is an unsupervised machine learning algorithm for anomaly detection that measures the **local detour** from a point relative to its neighbors. Unlike global methods, the LOF adopts a **perspective locale** to identify anomalies.

Below is the LOF figure applied to our dataset:

Table 11:Local Outlier Factor

Library	Status	LOF_Score	Risk_Level	Problematic_Features
Hassania Health	ANOMALY	1.136	Critical	Security OK
North Nador	ANOMALY	1.101	Critical	Security OK
Fes Medina	NORMAL	1.073	Critical	Security OK
Meknes West	NORMAL	1.070	Critical	Security OK
Kenitra Ville	NORMAL	1.050	Critical	Security OK
Agadir South	NORMAL	1.046	High	Security OK
Casablanca Centre	NORMAL	1.041	High	Security OK
Tetouan City	NORMAL	1.021	High	Security OK

Larache City	NORMAL	1.016	High	Security OK
Safi Centre	NORMAL	1.012	High	Security OK
BNRM	NORMAL	1.009	Medium	Security OK
Discount City	NORMAL	1.005	Medium	Security OK
Mohammed VI	NORMAL	1.001	Medium	Security OK
Marrakech Ville	NORMAL	0.994	Medium	Security OK
Taza Centre	NORMAL	0.993	Medium	Security OK
Abdelkrim Khattabi	NORMAL	0.990	Low	Security OK
Oujda East	NORMAL	0.973	Low	Security OK
The South Jadida	NORMAL	0.958	Low	Security OK
Tangier Nord	NORMAL	0.956	Low	Security OK
Sidi Kacem West	NORMAL	0.950	Low	Security OK

2- Interpretation

The Local Outlier Factor analysis above assessed the 20 libraries according to a four-level risk scale, revealing a significant distribution:

Table 12: Summary table of the LOF algorithm

Risk Level	Number of Libraries	Percentage	General Characteristics
Critique	5	25%	LOF scores > 1.05, multiple vulnerabilities
Pupil	5	25%	Scores 1.01-1.05, specific problems marked
AVERAGE	6	30%	Scores 0.99-1.01, moderate risk
Weak	4	20%	Scores < 0.99, relatively safe configurations

This distribution indicates that half of the network(10 out of 20 libraries) presents a critical or high level of risk, requiring special attention within the framework of IT security policies.

The major anomalies identified are:

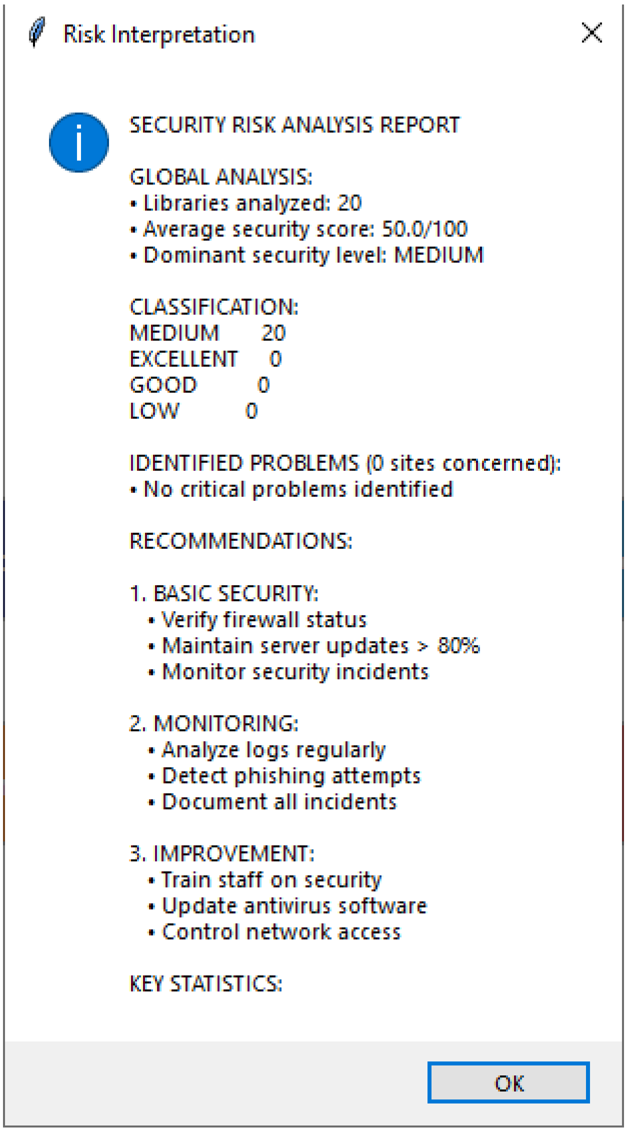
- **"Hassania Santé" Library:** The "Hassania Santé" library represents a particularly interesting case of a contextual anomaly. Despite the presence of only two identified vulnerabilities, their specific combination creates a statistically rare risk profile. The high LOF score (1.136) indicates that this library is located in a **low-density region** compared to its most similar neighbors. In other words, few comparable libraries simultaneously exhibit these two weaknesses, making this profile exceptionally deviant in its local context. This situation illustrates the fundamental principle of LOF: it is not necessarily the number of vulnerabilities that determines the anomaly, but rather their specific combination in a given context.

- **"Nador North" Library**: with identified vulnerabilities (High public access, pending critical updates, Multiple unresolved incidents, Active phishing activity. Unlike "Hassania Santé", "Nador Nord" presents an anomaly profile based on the **quantitative accumulation** of vulnerabilities. With four major simultaneous problems, this library stands out for the exceptional density of its security weaknesses.

III- Risk analysis and recommendations

Below is the risk analysis and recommendations:

Figure 23: Risk interpretation and recommendations



```
SECURITY RISK ANALYSIS REPORT

GLOBAL ANALYSIS:
• Libraries analyzed: 20
• Average security score: 50.0/100
• Dominant security level: MEDIUM

CLASSIFICATION:
MEDIUM      20
EXCELLENT    0
GOOD         0
LOW          0

IDENTIFIED PROBLEMS (0 sites concerned):
• No critical problems identified

RECOMMENDATIONS:

1. BASIC SECURITY:
  • Verify firewall status
  • Maintain server updates > 80%
  • Monitor security incidents

2. MONITORING:
  • Analyze logs regularly
  • Detect phishing attempts
  • Document all incidents

3. IMPROVEMENT:
  • Train staff on security
  • Update antivirus software
  • Control network access

KEY STATISTICS:
```

The security analysis conducted on 20 libraries reveals an average score of 46.5/100, with an overall security level assessed as **GOOD.** The distribution of sites shows that 7 present a level**GOOD**5, one level**WEAK**4, one level **AVERAGE**, and 3 a level **EXCELLENT**.

Among the problems identified, 19 sites are affected by at least one vulnerability, including:

- Disabling firewalls on multiple sites,
- A high number of security incidents,
- And a server update rate that is often lower than recommended.

The main recommendations focus on:

- **Basic security**: firewall checks, maintaining server updates above 80%, and increased incident monitoring.
- **Surveillance**: regular analysis of logs, detection of phishing attempts and systematic documentation of incidents.
- **Improvement**: staff training, antivirus updates and network access control.

Key statistics highlight that:

- The average server update rate is 74.2%.
- 13 sites have a rate below 80%.
- A total of 94 incidents were recorded.
- 12 sites recorded more than 3 incidents.
- Only 55% of firewalls are active (11 out of 20).

Corrective actions are needed to strengthen the security posture, particularly at sites with a low level and a high number of incidents.

Conclusion

This study demonstrated the feasibility and usefulness of a data-driven approach to analyzing and optimizing the network of Moroccan libraries. Through a progressive methodology ranging from principal component analysis (PCA) to unsupervised clustering, then predictive modeling (Random Forest), and finally cybersecurity anomaly detection (Isolation Forest, LOF), we revealed the multidimensional structure of this institutional landscape. The results highlight a clear typology of four to five library profiles, identify the most discriminating variables, and pinpoint critical security vulnerabilities affecting nearly half of the libraries. This analysis thus provides an objective and actionable diagnosis to guide public policy, while also validating the transferability of data science skills to diverse fields, from cultural management to cybersecurity.