



MedTech
Mediterranean
Institute of Technology



BIAT
Engagés
avec vous

ENGINEERING PROGRAM

INT 102 FINAL REPORT

Unsupervised Learning for Credit Risk Profiling: Application of Clustering Techniques at BIAT

BY

Sahar Bahloul

ACADEMIC SUPERVISOR

Walid Ben Haj Othmen

INSTITUTION SUPERVISOR

Nader Trigui

Banque Internationale Arabe de Tunisie

Tunis, 2025-2026

Approval

Approval

APPROVED BY

ACADEMIC SUPERVISOR

Name

Signature

Date

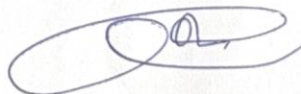
COMPANY SUPERVISOR

Name

Signature

Date

Trigun Noden



17/03 2021

ACADEMIC EVALUATOR

Name

Signature

Date

Declaration

I certify that I am the author of this project and that any assistance I received in its preparation is fully acknowledged and disclosed in this project. I have also cited any source from which I used data, ideas, or words, either quoted or paraphrased. Further, this report meets all the rules of quotation and referencing in use at MedTech, as well as adheres to the fraud policies listed in the MedTech honor code.

No portion of the work referred to in this study has been submitted in support of an application for another degree or qualification to this or any other university or institution of learning.

Student Name

Date

Signature

Work Term Release

I hereby state and verify by my signature that I have reviewed this report. I hereby affirm that the report contains

☐ no confidential data/information, and I authorize it to be released.

☐ confidential data/information, and I do not authorize it to be released.

COMPANY SUPERVISOR

Name

Signature

Date

Internship Agreement and Certificate



Internship Agreement and Certificate Submission

Summer 2025

The Career Services confirms that the student

Name: Sahar Bahloul

ID: 211-4313

Submitted the internship Agreement and Certificate of the

☐ INT 101

☒ INT 102

Internship entitled

.....

Career Services



Abstract

Keywords: Clustering, DBSCAN, K-means, PCA, unsupervised learning, ratios, outliers, Gaussian mixture model (GMM), Pearson filtering, missing values, CRISP-DM, count, mean, standard deviation, interquartile range, 99% percentile, flag, abnormal values, Spearman, correlation matrix, filtering.

This report is the result of a seven-week internship at BIAT within the risk department. This research applies the CRISP-DM methodology, encompassing systematic data preprocessing, outlier treatment via percentile capping, and domain-informed imputation strategies for missing values. A Python-based technique was developed that uses ratio checks and mode comparisons to identify and improve abnormalities automatically. This method standardized the data, ensuring consistency between firms and years. The next step was to fill in the missing values to get the dataset ready for clustering. After standardizing the data, dimensionality reduction was implemented using either PCA or a correlation matrix to remove strongly linked variables. As an experimental measure, different techniques were used, such as K-means and GMM (Gaussian mixture model). To evaluate the robustness of the clustering results, further tests were conducted using different methodologies, such as the silhouette score. Additional experiments explored advanced clustering methods like Trimmed K-Medoids, DBSCAN with Wasserstein Distance, Spectral Clustering on Distributions, and Wasserstein Barycenter Visualization. This internship helped me enrich my skills, make new connections, and apply theoretical knowledge in a practical setting, touching both two trendy worlds: finance and technology.

Acknowledgements

First and foremost, I want to express my heartfelt gratitude to my supervisor, Mr. Nader Trigui, for his invaluable support, constructive feedback, and constant encouragement. His knowledge, expertise, and guidance played a crucial role in shaping the direction of this project.

Special thanks to Mr. Ahmed Azzabi for his insightful assessment and valuable support, and to Miss Sinda Drira for welcoming me for the last two weeks of the internship as part of her team, allowing me to gain exposure to the financial side of the risk department

I extend my gratitude to the entire BIAT team for their patience, support, and for welcoming me as a valued team member right from the start

I am also thankful to my family for their constant support and understanding during the ups and downs of this journey.

To all of you, thank you deeply and sincerely.

Table of Contents

Approval	ii
Declaration	iii
Work Term Release	iv
Internship Agreement and Certificate	v
Abstract	vi
Acknowledgements	vii
List of Figures	11
List of Equations	12
1. EXECUTIVE SUMMARY	13
2. INTRODUCTION.	14
3. COMPANY CONTEXT	15
3.1. Description Of the Company	15
3.2. Mission and Objectives	15
3.2.1. Vision.....	15
3.2.2. Mission	16
3.2.3. Goals	16
3.2.4. Values.....	16
3.2.5. Objectives.....	16
3.3. Industry Structure	17
3.4. Market Structure.....	17
3.4.1. Group Structure and Business Lines	18
3.4.2. Presentation of The Risk Department	19
3.4.3. The Credit Risk Treatment Process	20
4. INTERNSHIP DESCRIPTION	21

4.1.	Internship Context.....	21
4.2.	General and specific objectives of the internship	21
4.3.	Challenges and Obstacles.....	22
4.4.	Assigned Tasks and Responsibilities	25
4.5.	Key Learnings and Observations	26
5.	LITERATURE REVIEW	27
5.1.	Challenges of Credit Risk	27
5.2.	Traditional Management of Credit Risk	27
5.3.	Unsupervised Versus Supervised Learning	28
5.3.1.	The Credit Risk Treatment Process	28
5.3.2.	What is Supervised Learning?	29
5.3.3.	Comparison	30
5.4.	Unsupervised Methods In Credit Risk	30
5.4.1.	K-means Clustering	30
5.4.2.	DBSCAN (Density-Based Clustering)	30
5.4.3.	Gaussian Mixture Models (GMM).....	31
5.4.4.	Advanced Methods	32
5.4.4.1.	Trimmed K-Medoids	32
5.4.4.2.	Weighted k-medoids	32
5.4.4.3.	Energy distance k-medoids	32
5.4.4.4.	MMD k-medoids (Maximum Mean Discrepancy)	33
5.4.4.5.	Wasserstein barycentres robustes (trimmed barycentres)	33
5.4.4.6.	Spectral clustering.....	33
5.4.5.	Anomaly Detection in Credit Risk	33
5.5.	Dimensionality Reduction in Risk Data	33
5.6.	Hybrid Approaches and BIAT's Strategy	34

6. METHODOLOGY	35
6.1. Business Understanding	35
6.2. Data Understanding.....	36
6.2.1. Variables	36
6.2.2. Data Characteristics.....	38
6.3. Data Preparation	38
6.3.1. Descriptive Statistics.....	38
6.3.2. Unit Consistency Check with K_social	40
6.3.3. Correlation-based validation of flagging	41
6.3.4. Outlier treatment	43
6.3.5. Ratio recalculation	44
6.3.6. Handling missing values	44
6.3.7. Standardization	44
6.3.8. Feature Selection, Dimensionality Reduction and Clustering	45
6.3.9. Approach A: PCA only + K-means	45
6.3.9.1. Approach B: PCA + UMAP + K-Means.....	45
6.3.9.2. Gaussian Mixture Model (GMM) on PCA	46
6.3.9.3. Pearson p-Value Filtering + K-Means	46
7. RESULTS AND FINDINGS	47
8. RECOMMENDATIONS	55
9. CONCLUSIONS	56
REFERENCES	57

List of Figures

Figure 1: Banking Industry under the supervision of the Central Bank of Tunisia	17
Figure 2: Structure of the Tunisian banking market	18
Figure 3: BIAT Subsidiaries and Range of Activities	19
Figure 4: Organizational Chart of the Risk Department	19
Figure 5: Credit Risk File Treatment Process.....	20
Figure 6: Interpretation of the tests	42

List of Equations

Equation 5.1 : Expected Loss (Basel formula)	27
Equation 5.2 Logistic Regression Probability of Default	27
Equation 5.3 : K-Means Objective Function.....	30
Equation 5.4 : DBSCAN Core Condition	30
Equation 6.1: ratio check	40
Equation 6.2: mode check	40
Equation 6.3: absolute difference between correlations	41
Equation 6.4: Tukey bounds	43
Equation 6.5: z-score standardization.....	45
Equation 6.6: PCA formula	45
Equation 6.7: Pearson correlation coefficient	46

1. EXECUTIVE SUMMARY

During my seven weeks internship at the headquarters of BIAT carried at the Risk Management department under the supervision of Mr. Nader Trigui, I had the opportunity to get exposure to how banks work and manage their credit risk. The project focused on applying unsupervised techniques on a real dataset.

The work followed CRISP-DM methodology with a big focus on the data preparation and preprocessing. The key challenges were understanding the dataset's variables, translating it from French to English, treat unit problems, cap outliers and handle missing values. Custom algorithms were developed for each step.

Serval approaches were tested such as PCA, UMAP, Pearson filtering, dimensionality reduction using correlation matrix and p-value, as well as serval clustering techniques such as K-mean, DBSCAN and Gaussian Mixture Model.

Among these, the PCA-based K-Means model delivered the most stable and interpretable results. It produced four clear profiles, ranging from financially healthy companies to likely stressed ones. The result showed that unsupervised learning can clearly take a role in BIAT's traditional credit risk assessment tool by mapping a company into a category based on its financial sheet.

Based on these findings, serval recommendations can be made to strengthen BIAT's credit risk assessment approach. First, the data quality must get a better check at the source level, since the data contained a lot of inconsistencies and missing values. Second, the models would benefit from integrating additional variables beyond accounting ratios, such as sector-specific indicators and client behavioral data. Finally, incorporating macroeconomic factors like inflation, interest rates, or market shocks would allow the system to adapt more effectively to changing economic conditions. All these factors together can lead to a more performant and reliable system.

2. INTRODUCTION.

Artificial intelligence is playing a crucial role in today's world. It is shaping every aspect of our lives, from daily tasks to more complicated ones. And the banking sector is no exception, as it is being transformed by this new technological generation.

Conventional risk models are changing to become data-driven, dynamic systems. These systems identify hidden weaknesses, find irregularities early, and predict new risks before they become real. Financial institutions can transition from reactive troubleshooting to proactive resilience building thanks to this change.

I was particularly interested to dive into machine learning and AI as I had taken this course last semester with Miss Dorra Louati, and I was deeply curious about how to apply different dimensionality reduction and clustering techniques in the real world.

It was within this context that I had the opportunity to carry out my INT102 Engineering Internship at the Banque Internationale Arabe de Tunisie (BIAT), in the Risk Management Department, Credit Risk Unit. My mission was to clean the real data I was given, following the CRISP-DM methodology and researching and applying different clustering techniques.

The goal was to classify the companies that BIAT works with into three categories and create a model that does it automatically. With the use of this classification framework, the bank would be able to classify any new customers into the proper category and then determine whether they qualify for credit.

The first part of the internship consisted of conducting profound research about unsupervised learning, from traditional models to the most up-to-date version, to move to the practical implementation step.

This report provides an overview of my seven-week internship. It starts with a company presentation followed by the tasks that I was given and accomplished during this time. A literature review is presented to help frame the theoretical notions underpinning my work. The technique and models used are then described, followed by a discussion of the findings.

In addition, I will briefly describe the obstacles encountered and the tactics used to overcome them. Finally, I will give some recommendations for future work.

3. COMPANY CONTEXT

3.1. Description Of the Company

BIAT – Banque Internationale Arabe de Tunisie is a universal bank founded in 1976 and today ranks among the country's leading financial institutions. It operates a wide branch network and a diversified group of subsidiaries in insurance, asset management, private equity, stock-market intermediation, consulting, and international payments.

My internship took place in the Risk Department, within the Strategy and Risk Policy Division. The working environment is structured, compliance-driven, and collaborative: analysts in retail/corporate banking coordinate with risk analysts, the credit committee, and administrative teams (T24 core banking) along a documented, multi-level validation process.

Indicator	Value
Net Banking Income	1,479.7 MDT
Equity	2,226.4 MDT
Net Result	357.8 MDT
Customer Deposits	20,814.1 MDT
Net Loans Outstanding	12,806.9 MDT
Employees	2,415
Branches	206

Table 1: BIAT Key Benchmarks at the End of 2024

3.2. Mission and Objectives

3.2.1. Vision

BIAT aims to become a leading financial institution in Tunisia and the North African area. It is known for providing excellent customer service and actively contributing to economic growth.

3.2.2. Mission

The bank aims to deliver a wide range of innovative and integrated financial solutions tailored to the diverse needs of its clients, while actively supporting Tunisia's economic and social advancement

3.2.3. Goals

- Consolidate its position as a market leader in Tunisia
- Accelerate the development and adoption of digital banking platforms
- Strengthening customer loyalty through enhanced service quality
- Promote long-term, inclusive economic development

3.2.4. Values

BIAT operates on the principles of dedication, transparency, customer focus, innovation, and social responsibility.

3.2.5. Objectives

- Expand the bank's asset base and improve financial performance
- Increase efficiency by leveraging digital tools and processes
- Encourage a culture centered on innovation and continuous improvement
- Uphold rigorous standards in governance and risk oversight

3.3. Industry Structure

At the top sits the Central Bank of Tunisia (CBT), which supervises all institutions. Beneath it, the industry splits into:

- Lending institutions, which include:

Banks (21), Financial institutions (13), under which you commonly find Leasing companies (9), Factoring companies (2), and Merchant banks (2).

- Specialized banks/banks with particular status, including Offshore Banks (8).
- Agencies representing foreign banks in Tunisia (09).

This structure reflects Tunisia's universal-banking model with specialized layers for non-bank finance and internationally oriented activities (offshore), all under CBT prudential oversight.

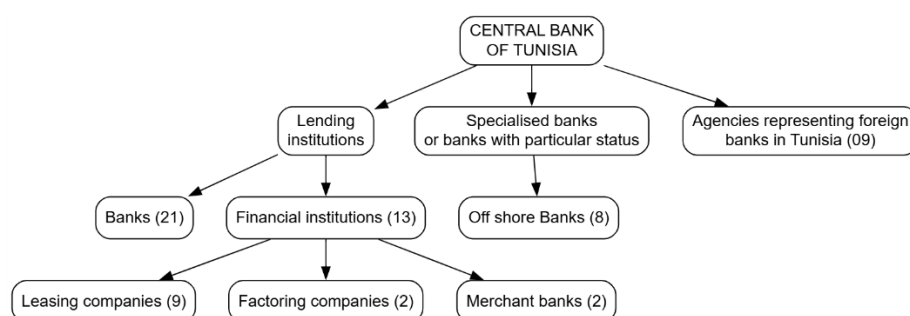


Figure 1: Banking Industry under the supervision of the Central Bank of Tunisia

3.4. Market Structure

The Tunisian financial system is largely dominated by banks, which remain the main source of financing for the economy. By the end of 2023, the sector consisted of 29 licensed institutions, including 22 local banks and 7 foreign branches.

In terms of size, the sector's total assets stood at TND 160.7 billion, while customer deposits reached TND 103.9 billion, showing an annual growth rate of around 6%. This reflects steady expansion despite economic pressures.

Public banks continue to control more than a third of sector assets and loans, yet private banks such as BIAT remain strong competitors and continue to consolidate their positions through innovation, digital transformation, and customer service improvements.

From a stability perspective, the system is considered financially solid, with a capital adequacy ratio of 14.5% and a Tier 1 ratio of 11.5%, both comfortably above the regulatory thresholds imposed by the Central Bank of Tunisia. On the stock market, bank equities remain among the most profitable, confirming investor confidence in the sector.

BIAT is considered to have a very significant position. With TND 18.8 billion in deposits representing almost 18% of the market share, we can confidently say that the bank is one of the largest in Tunisia. In 2024, BIAT ranked among the top five banks by net profit, alongside Attijari, UIB (Amen), BNA, and BT. Beyond its constant presence, BIAT also operates internationally through BIAT France and relies on specialized subsidiaries such as BIAT Asset Management and BIAT Capital Risque, which strengthen its role as a universal bank offering integrated solutions across the market.

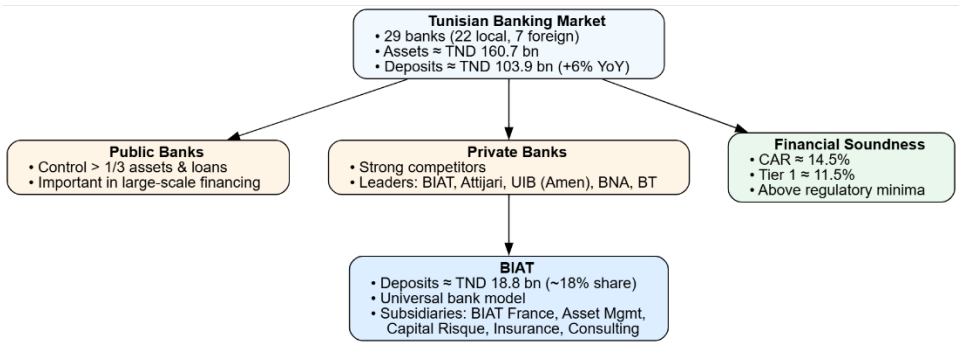


Figure 2: Structure of the Tunisian banking market

3.4.1. Group Structure and Business Lines

BIAT's specialty subsidiaries have created a diverse structure that strengthens its function as a universal bank. The bank's operations extend beyond traditional banking through companies such as BIAT France, BIAT Consulting, BIAT Capital Risk, Tunisie Valeurs, and BIAT Insurance. They together provide a wide range of services, including ordinary banking operations and loans, insurance, asset management, financial engineering, capital investment, and overseas consultation. Thanks to this arrangement, BIAT can effectively handle the diverse needs of individuals, SMEs, large corporations, and institutions both locally and globally.

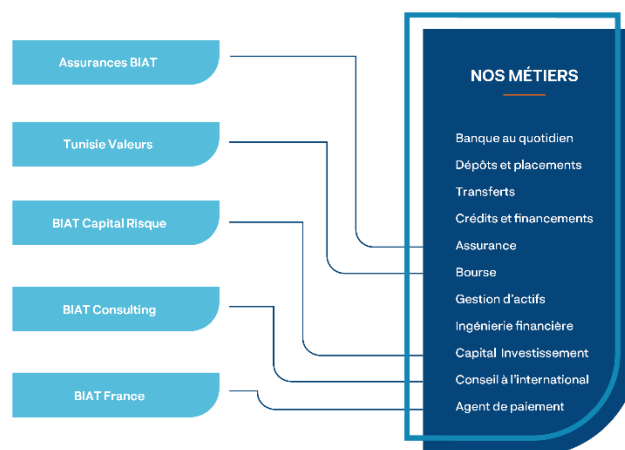


Figure 3: BIAT Subsidiaries and Range of Activities

3.4.2. Presentation of The Risk Department

The risk department's responsibility is to ensure the bank's strength and financial stability. It is responsible for discovering, analyzing, and managing all risks, including market and operational exposures, as well as credit risk. It serves as a protection against possible losses by trying to predict the probability of default attached to each loan given.

The Credit Risk Department focuses on evaluating borrowers' repayment capacity, while the Market Risk Division oversees risks linked to fluctuations in interest rates, exchange rates, and securities. The Operational Risk & Permanent Control Division monitors internal processes, fraud, and human errors. Meanwhile, the Risk Strategy & Policies Division designs the overall risk framework, and the Risk Modeling & Portfolio Analytics Division provides statistical models and stress tests to help make decisions. Finally, Risk Reporting & Regulatory Control ensures transparency with authorities, and Risk Projects & Systems drives innovation through digital solutions. Together, these units form a comprehensive structure that aligns risk management with the bank's strategic objectives.

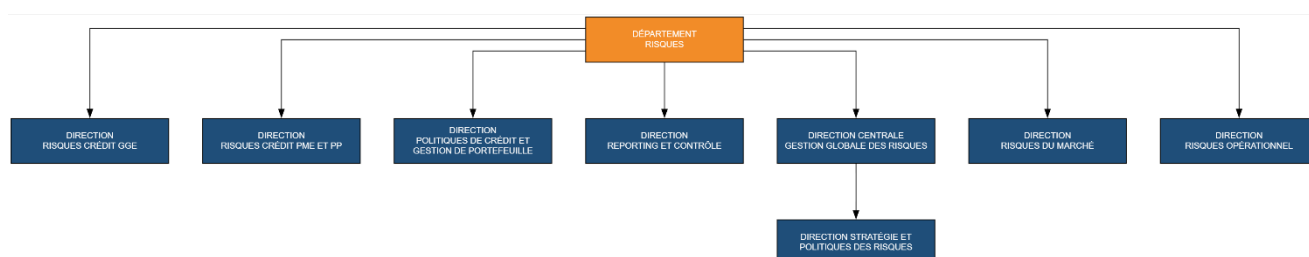


Figure 4: Organizational Chart of the Risk Department

3.4.3. The Credit Risk Treatment Process

The credit approval process is one of the most critical functions of the Risk Department, ensuring that loans are granted responsibly and sustainably. The process begins with the submission of the client's request and required documentation. An initial review is then performed by the business manager using RiskPro, a decision-support tool. The case is forwarded to the Risk Department for a thorough analysis, which includes scoring, rating, and providing a contradictory opinion to ensure objectivity.

Once the evaluation is complete, the application is presented to the Credit Committee, which makes the final decision. If approved, the guarantees are established and verified before the decision is transmitted to the Central Bank of Tunisia and other banks. The client is then notified, and the administrative team sets up the loan in the T24 core banking system.

Finally, the funds are disbursed. If the application is rejected, the client is informed, and the file is closed. This structured process reflects the department's mission to maintain a balance between business growth and prudent risk management.

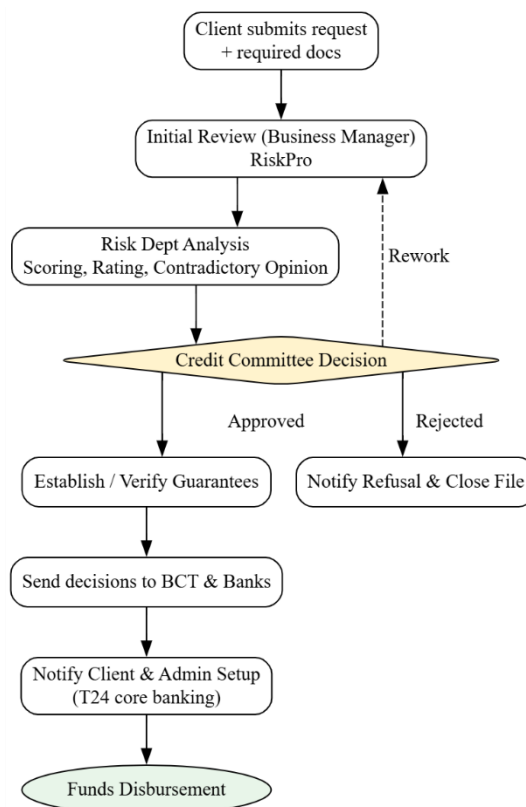


Figure 5: Credit Risk File Treatment Process

4. INTERNSHIP DESCRIPTION

4.1. Internship Context

My internship was conducted at the Risk Department of BIAT, under the supervision of Mr. Nader Trigui. It lasted seven weeks, from mid-May to the first week of July. The project consisted of applying unsupervised learning techniques to the financial data. The goal was to explore and understand how clustering techniques could be used to group different companies based on some common characteristics without being given a target value.

I had to work with a large and complex dataset of 109,381 rows and 157 columns, which were expressed in French financial abbreviations. The work, therefore, required a deep understanding of financial ratios, as well as translation and deep data cleaning before any modeling could be performed.

4.2. General and specific objectives of the internship

The primary objective was to investigate the application of unsupervised learning techniques in credit risk analysis and their effectiveness in improving BIAT's client segmentation.

At the specific level, my first objective was to prepare a full presentation to my supervisor that encompasses all aspects of unsupervised learning, from clustering techniques such as K-Means, DBSCAN, Hierarchical clustering, GMM, Spectral clustering, Mean Shift, and OPTICS, to anomaly detection such as Isolation Forest, One-Class SVM, LOF, and Autoencoders. I also dived into dimensionality reduction techniques such as PCA, t-SNE, UMAP, ICA, and LLE. The goal was to obtain a general idea about this domain to be able to choose which technique would be best adapted to my project.

The next step was to translate and understand the 157 variables. I built my own dictionary that contains all the formulas in the French and English version as well as their interpretation. This gave me a clearer idea of what I am working with. Understanding data is a crucial step because, without it, it can lead to a misunderstanding of the requirements or even non-useful work.

Another objective was to clean and preprocess the dataset. This included standardizing reporting units, because in some years they used thousands while others used millions, as well as handling missing values. A particular challenge was that many missing values were due to division by zero in financial ratios, and in this case, using classic imputation like KNN (k-nearest neighbor) was not allowed, because that would contradict the coherence and the reliability

of the real data. As my supervisor said, we never impute artificial data in a financial context. Therefore, I had to create a custom flagging and detection method to identify unrealistic jumps or inconsistencies without creating artificial values, and my own algorithm that fills in missing values based on the financial ratio type.

The core objectives of the project involved applying clustering algorithms such as K-Means, DBSCAN, and Gaussian Mixture Models, to moving to the step of dimensionality reduction techniques like PCA and UMAP to simplify the data and visualize clusters.

Lastly, studying BIAT's IFR9 categorization system, which goes from 0 (performing customers) to 4 (default), was one of my goals. Although I didn't use it in clustering, learning about it gave me crucial background information on how banks internally categorize risk.

Another fundamental goal was managing the work's computational complexity. Running models frequently requires a long time because of the size of the dataset and the nature of methods like DBSCAN or GMM, and the flagging algorithms that I have created. That's why I had to figure out how to evaluate them on smaller samples before applying techniques to the entire dataset

4.3. Challenges and Obstacles

One of the most difficult parts of my internship was the data preprocessing stage. I used the CRISP-DM method, which included:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation

Therefore, before I could run any clustering or anomaly detection methods, I had to spend an enormous amount ensuring the dataset itself was reliable. I faced several key challenges: translation, inconsistent units, missing values, and outliers.

The first challenge was translating the variables. All the features were labeled with abbreviated French accounting and banking terms. At first, I found it overwhelming, because I couldn't immediately connect the abbreviations to the ratios they represented. For example, "CAF"

stood for *Capacité d'Auto financement* (self-financing capacity), "FRNG" for *Fonds de Roulement Net Global* (net working capital), and so on. In MedTech, we took the introduction to finance course in English, which made it difficult for me to distinguish which French ratio name translates to which one in English.

To solve this, I created my own dictionary of variables where I translated every abbreviation into English and wrote down its full formula, meaning, and interpretation. This process took time, but it gave me a much better understanding of what I was working with, allowing me to avoid mistakes later.

The second challenge was inconsistent reporting units across the dataset. Some companies reported values in thousands of dinars in some years, while reporting in millions in other years. Therefore, for the same company, one row could look "tiny" next to another row just because of the scale, not because of a real difference. To handle this, I designed a method that compares values year by year for each company. If a value suddenly jumped by an unrealistic factor (for example, $1000\times$ larger or smaller from one year to the next), I flagged it as suspicious. I also compared values to the most frequent valid value (the statistical mode) within each company's history. This way, I could detect when a unit inconsistency was likely the cause of the problem and make corrections without losing the financial logic of the data.

The third obstacle was addressing the problem of outliers. Some variables had extreme values that could completely distort clustering. Also, the data was a mix of very big companies and startups, which could appear as outliers, but they just don't belong to the same group. Instead of removing them, I decided to correct them in a structured way. I grouped companies by their sector of activity and by size class (small, medium, or large, depending on their revenue). Then, for each group, I used Tukey's method based on quartiles: values below $Q1 - 1.5 \times IQR$ were flagged as too low, and values above $Q3 + 1.5 \times IQR$ were flagged as too high. Instead of deleting these observations, I replaced them with the nearest quartile value ($Q1$ or $Q3$). This ensured data consistency while decreasing the impact of extreme points.

Finally, I handled the issue of missing values caused by division by zero. Many financial ratios, such as return on equity (ROE) or solvency ratios, can't be computed when the denominator is zero. These weren't "missing" in the usual sense but were mathematically undefined.

Using a method like KNN imputation would have created values that made no financial sense. To fix this, I created a safe division function in Python.

The idea was simple: whenever a denominator was zero, I didn't just return infinity or NaN. Instead, the function checked the context:

- If the numerator was positive and the denominator was zero, it replaced the result with the maximum valid ratio in the dataset.
- If the numerator was negative and the denominator was zero, it replaced the result with the minimum valid ratio.
- If both the numerator and denominator were zero, it set the result to 0.

This way, every undefined case was handled consistently, and the replacement values still made sense compared to the rest of the data. It also allowed me to calculate dozens of derived ratios (liquidity ratios, profitability ratios, leverage ratios, etc.) without leaving holes in the dataset.

4.4. Assigned Tasks and Responsibilities

Phase	Week(s)	Main tasks	Algorithms/focus	Deliverables
Presentation	Week 1	Delivered an in-depth presentation about unsupervised learning, from clustering techniques to anomaly detection to dimensionality reduction	Conceptual overview	PowerPoint slides and oral presentation
Data understanding	Week 2	Translated 154 financial variables and ratios from French to English and created my own dictionary. I also did all the data exploration steps	Pandas, matplotlib	Dictionary of variables and data visualization
Data preparation	Week 3-4	Handled unit differences, outlier detection, and missing values	Custom Python functions using Tuckey filter	Cleaned data set
Core modeling	Week 5-6	Clustering methods and dimensionality reduction	UMAP, PCA, k-mean and GMM	Clustered data
Additional research	Week 7	Additional research suggested by my supervisor about rimmed/weighted K-Medoids, distributional clustering (Energy distance, MMD), Wasserstein barycenter, DBSCAN/Spectral with Wasserstein.	K-Medoids variants, Wasserstein distances, Spectral clustering	Documentation, academic review, exploratory notes

Table 2 : timeline of work

In the first week, I was assigned to create a full presentation about all the aspects of unsupervised learning from clustering: K-Means, DBSCAN, Hierarchical, GMM, Spectral, Mean Shift, and OPTICS to Anomaly Detection, Isolation Forest, One-Class SVM, LOF, and Autoencoders to Dimensionality Reduction: PCA, t-SNE, UMAP, ICA, LLE, and Isomap. I went in-depth for each of them. I had to read about the mathematical background of each algorithm, how it works step by step, and where it is used.

The data preparation phase was a critical responsibility. I implemented a safe division function to handle missing ratios caused by division by zero. I also designed a method to spot unit inconsistencies by comparing year-to-year ratios and flagging suspicious jumps. Finally, I detected and capped outliers using Tukey's method within sector-size groups, ensuring comparability between companies.

In the next week, I was finally able to apply different clustering techniques and dimensionality reduction techniques and compare their results. I also tried to reduce the data using only the correlation matrix p-value and not using UMAP to see if I could find better results.

Last week was dedicated to additional research. I explored advanced clustering methods like trimmed and weighted K-Medoids, energy distance K-Medoids, MMD-based clustering, robust Wasserstein barycenter, and spectral clustering with distributional distances. These were not part of my initial tasks but came as an extension, helping me understand where the field is going and giving me ideas for BIAT's future.

4.5. Key Learnings and Observations

- **Preprocessing is crucial:** I realized that understanding, cleaning, and correcting the data is as important as modeling itself, and it is also more time-consuming.
- **Unsupervised \neq Supervised:** By working alongside Fares, I learned how profiling (unsupervised) and prediction (supervised) complement each other in risk analysis.
- **Adaptation of methods:** Financial ratios require adapted methods (safe division, sector-based outlier detection) rather than generic approaches.
- **Communication matters:** daily meetings and the final presentation improved my ability to explain technical work to non-specialists.
- **Company initiatives observed:** BIAT is clearly investing in digital transformation. The fact that both supervised and unsupervised projects are being tested in parallel shows the bank's commitment to integrating AI into its risk management framework.

5. LITERATURE REVIEW

5.1. Challenges of Credit Risk

Credit risk has always been a significant challenge for banks, as it directly impacts profitability and financial stability. Credit loss occurs when borrowers fail to meet their repayment obligations, resulting in a loss for banks. In Tunisia, the issue is particularly pressing: the IMF reported that the country's non-performing loan (NPL) ratio reached 15.8% in 2015, one of the highest in the region [1]. The Basel Committee on Banking Supervision (BCBS) provides a standardized framework for assessing credit risk. It is based on the expected loss formula.

$$EL = PD \times LGD \times EAD$$

Equation 5.1 : Expected Loss (Basel formula)

Where:

- PD: probability of default
- LGD: loss given default
- EAD: exposure at default

If PD is inaccurately estimated, the bank may misallocate capital and underestimate its exposure to risk [2].

5.2. Traditional Management of Credit Risk

Back then, Tunisian banks such as BIAT relied on experts' judgment and rule-based scorecards. But with the introduction of Basel II and III, banks shifted towards supervised learning techniques, in particular, logistic regression. This model is popular for its simplicity and interpretability.

The logistic regression model estimates the probability of default as follows:

$$P(y = 1 | X) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)})$$

Equation 5.2 Logistic Regression Probability of Default

5.3. Unsupervised Versus Supervised Learning

5.3.1. The Credit Risk Treatment Process

Unsupervised learning is when the machine learns patterns from unlabeled data. It detects clusters, anomalies, and hidden structures in the dataset. Some clustering applications are genetic research, image segmentation, medical imaging, and social network analysis. In the context of BIAT, this allows for clustering different customer groups and detecting early risk behavior without any labeled data.

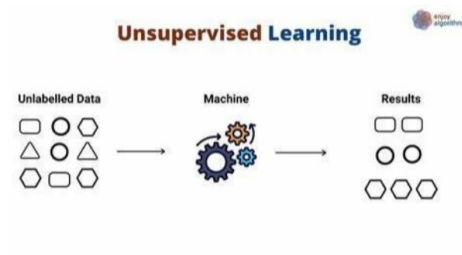


Figure 6: unsupervised learning

The primary objectives of unsupervised learning are clustering (grouping similar items), anomaly detection (identifying unusual outliers), association rule learning (market basket analysis), and dimensionality reduction (simplifying the data).

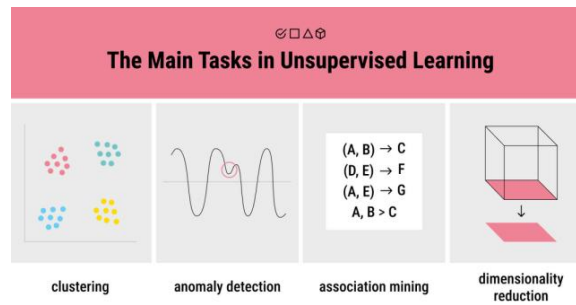


Figure 7: Main tasks of unsupervised learning

5.3.2. What is Supervised Learning?

In supervised learning, on the other hand, the algorithm learns how to classify new unseen data based on a labeled dataset. It works in a way where we give the algorithm trained data, we split it into training and testing data. In the training part, the algorithm analyzes this labeled data to find patterns and the relationship between the inputs and outputs. It then predicts unlabeled data based on its training and corrects itself against its own prediction.

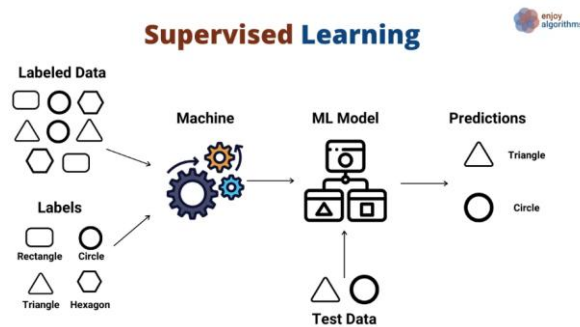


Figure 8: supervised learning

Supervised learning is mainly used in fields like credit risk scoring, fraud detection, medical diagnosis, image classification, and speech recognition. In the context of BIAT, supervised learning can be applied to predict whether a client will default, estimate loan and repayment probability based on historical labeled data.

The types of supervised learning techniques are regression and classification. The popular algorithms used are linear regression, logistic regression, decision trees, random forest, support vector machines (SVM), and neural networks.

5.3.3. Comparison

Aspect	Supervised Models	Unsupervised Models
Input	Requires labeled defaults (PD = 0/1)	Works with unlabeled financial ratios
Goal	Estimate PD, classify borrowers	Discover hidden groups and anomalies
Techniques	Logistic Regression, XGBoost, SVM	K-Means, DBSCAN, GMM, PCA, Spectral
BIAT Relevance	Basel-aligned credit scoring	Portfolio segmentation and anomaly detection

Table 3: comparison between supervised and unsupervised learning

5.4. Unsupervised Methods In Credit Risk

5.4.1. K-means Clustering

It's an algorithm for clustering data based on repeatedly assigning points to clusters and updating those clusters' centers.

$$J = \sum_{i=1}^k \sum_{x \in C_i} ||x - \mu_i||^2$$

Equation 5.3 : K-Means Objective Function

5.4.2. DBSCAN (Density-Based Clustering)

It groups points that are closely packed, in high-density areas, together. And it labels points that are in low-density areas as noise

$$Core(p) \Leftrightarrow |\{q \in D : dist(p, q) \leq \epsilon\}| \geq MinPts$$

Equation 4.4 : DBSCAN Core Condition

The minimum points are the minimum number of points together for a region to be classified as dense, and the eps (ϵ) is a distance measure that will be used to locate the points in the neighborhood of any point.

The core is a point that has at least m points within distance n from itself. Meanwhile, the border is a point that has at least one core point at a distance n . The noise, on the other hand, is a point that is neither a core nor a border. It has fewer than m points within distance n from itself.

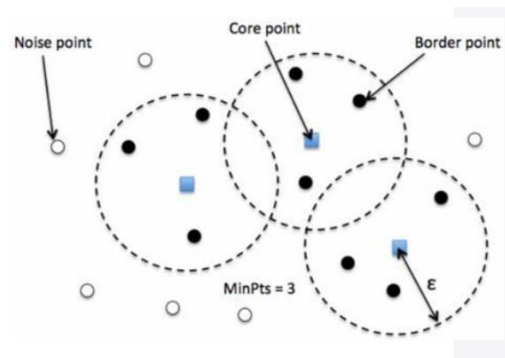


Figure 9 : DBSCAN clustering showing core, border, and noise points

5.4.3. Gaussian Mixture Models (GMM)

GMM is a soft clustering technique. It assumes that data is from a mix of Gaussians and each point has probabilities of belonging to each cluster.

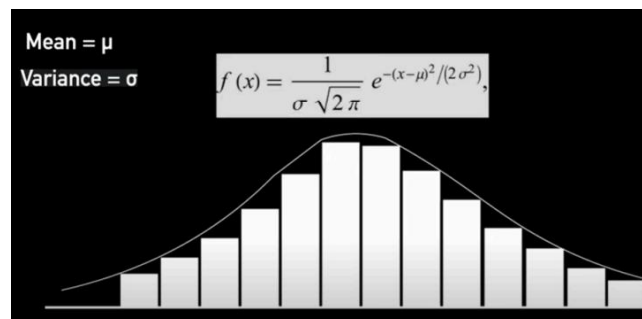


Figure 10 : Gaussian distribution

5.4.4. Advanced Methods

In addition to classical methods, my research explored more advanced methods:

5.4.4.1. Trimmed K-Medoids

This method works like regular K-Medoids but ignores a small portion of extreme points before forming clusters. By trimming those outliers, the algorithm builds groups that better reflect most of the data. In practice, this helps avoid results being distorted by companies with very unusual financial ratios.

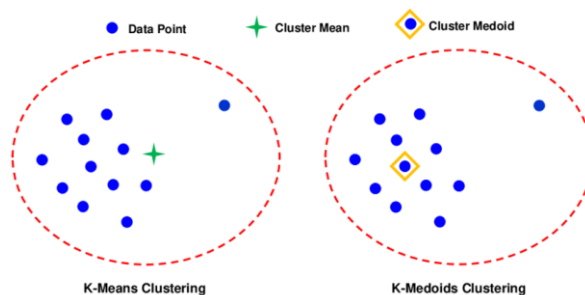


Figure 11 : K-Medoids

A medoid is the most representative object in a cluster. It is the point that has the smallest total distance to all other points in its group. This makes K-Medoids more robust to outliers and noise, as it avoids being distorted by extreme values.

5.4.4.2. Weighted k-medoids

Weighted K-Medoids gives more importance to certain points during clustering. For example, a large company or a recent observation can count more than a small one. This way, the final clusters highlight the firms that matter most for the analysis, instead of treating all data equally.

5.4.4.3. Energy distance k-medoids

It's a version of K-Medoids clustering where the distance between points is measured using energy distance instead of Euclidean distance. Energy distance is a statistical distance between probability distributions.

5.4.4.4. MMD k-medoids (Maximum Mean Discrepancy)

MMD K-Medoids uses a kernel trick to compare data in a richer space and then measures how far away their distributions are. The goal is to group firms that behave alike over time, even if their raw numbers look different. It is very helpful for financial series where hidden patterns only appear in more complex structures.

5.4.4.5. Wasserstein barycentres robustes (trimmed barycentres)

Instead of averaging simple points, this method averages entire distributions while trimming outliers that could alter the result. The outcome is a “central profile” that better represents most firms. In finance, this provides a typical borrower profile without being skewed by extreme or rare cases.

5.4.4.6. Spectral clustering

Spectral clustering builds a graph of similarities and then uses eigenvalues to split the data into groups. Unlike traditional methods, it can uncover complex shapes and hidden structures. This makes it very effective when financial profiles are diverse and not easy to separate with simple distance rules.

5.4.5. Anomaly Detection in Credit Risk

Besides clustering, there are different anomaly detection techniques such as Isolation Forest, One-Class SVM, LOF, and Autoencoders.

5.5. Dimensionality Reduction in Risk Data

Since the data is very big and we have over one hundred financial ratios, dimensionality reduction is a must. For this, there are three main methods available, which are PCA, T-SNE, and UMAP. PCA identifies the main variation in the data, helping us keep the most useful information while filtering out noise. t-SNE is especially powerful for visualization, as it shows clusters clearly by keeping nearby points close when projected into two or three dimensions. UMAP combines the strengths of both, offering speed and accuracy while preserving both local and overall details, which makes it particularly effective for complex financial datasets.

5.6. Hybrid Approaches and BIAT's Strategy

The most promising and useful technique is to use unsupervised learning to segment clients and then we apply supervised techniques within each cluster.

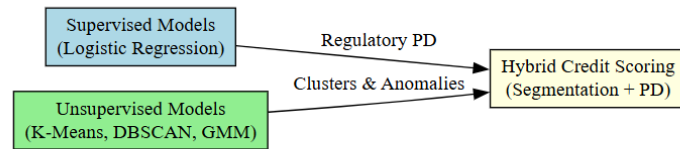


Figure 12 : Hybrid approach for BIAT

6. METHODOLOGY

This methodology section presents the comprehensive steps followed throughout the credit risk modeling project conducted at BIAT Bank. Guided by the Cross Industry Standard Process for Data Mining (CRISP-DM) framework, it details the phases from business understanding to data preparation. This part aims to address the central research question: How can an unsupervised learning model be developed and implemented to cluster different customer segments, using the internal financial data of a Tunisian commercial bank? The subsequent steps outlined here form the foundation for answering that question.

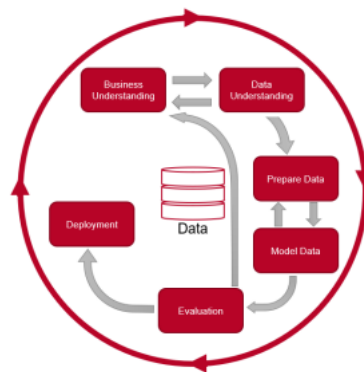


Figure 13 : CRISPDm

The research strategy chosen was quantitative. Since the project is based on data analysis, it required statistical and computational techniques rather than interviews. The raw data was given directly by my supervisor, and it was not collected; it comes from the actual database of BIAT.

6.1. Business Understanding

Credit risk assessment is an important activity in the banking sector. It has its own department and different people working on it, as it is crucial for the bank's activity. Traditionally, it mainly relies on supervised learning to estimate the probability of default (PD). However, the problem with Tunisian data is that the probability of default is scarce and imbalanced, making the prediction model unreliable.

The objective of the project is then to test whether unsupervised learning can effectively manage to segment BIAT's clients into homogenous groups and detect anomalous firms with atypical financial behavior.

6.2. Data Understanding

The dataset was provided by BIAT's Risk Department. It consisted of anonymized financial statements from corporate clients, with both raw financial items and derived ratios. It contains 109,381 rows and 157 columns with over 80 financial ratios.

6.2.1. Variables

The data was a mix of qualitative and quantitative variables.

- Qualitative variables
 - SECTEUR (sector of activity),
 - Appartenance Groupe (binary indicator of group membership).
- Quantitative variables:

More than 80 ratios, grouped into categories as shown below:

Category	Example Ratios
Liquidity	Current ratio, Quick ratio, Cash ratio
Leverage	Debt-to-equity ratio, Solvency ratio
Profitability	Return on Assets (ROA), Return on Equity (ROE), EBITDA margin
Efficiency	Asset turnover, Inventory turnover, Receivables turnover
Cash Flow	Operating cash flow ratio, Interest coverage ratio
Growth & Size	Turnover growth, Total assets, CA_GROUP

Table 4 : ratio groups

To further understand the data, I had to create my own dictionary of financial ratios, which took quite a lot of time. The variables were given in French abbreviations, so I had to look for each abbreviation's full name in French, then translate it into English to find the formula and the interpretation. After a lot of work and patience, I came up with this comprehensive table. What follows is an excerpt from the full table:

Abbreviation	Full French Name	Full English Name	Formula	Interpretation & Analysis
ID-ENT	Identifiant Entreprise	Company ID	-	Unique company identifier
ANNEE_N	Année	Fiscal Year	-	Reporting year (e.g., 2023)
CODACTIVBCT_BL	Code Activité Bancaire	Bank Activity Code	-	Regulatory classification of banking activities
SECTEUR	Secteur d'Activité	Industry Sector	-	Primary business sector (e.g., Manufacturing)
Appartenance Groupe	Appartenance à un Groupe	Group Affiliation	-	1=Yes, 0=No. Indicates corporate group membership
EXPORT	Exportations	Export Revenue	-	Revenue from foreign sales. High % = global exposure
Liq_Reducite	Liquidité Réduite	Quick Ratio	$(Act_courant - stocks) / Pass_courant$	Immediate liquidity. >1x safe
Liq_Immédiate	Liquidité Immédiate	Cash Ratio	$Liq / Pass_courant$	Strictest test. >0.2x acceptable
Part_FPN	Part des Fonds Propres	Equity Share	$FPN / (FPN + TPass)$	% of funding from equity. >30% preferred

Table 5 : dictionary of the variables

6.2.2. Data Characteristics

Exploratory analysis revealed:

- High dimensionality (80+ ratios, many correlated),
- Skewed distributions (e.g., leverage ratios with extreme tails),
- Missing values (caused by financial realities such as zero turnover or negative equity),
- Outliers (extreme values indicating potential anomalies).

6.3. Data Preparation

6.3.1. Descriptive Statistics

As part of our initial data exploration, we conducted a detailed descriptive statistical analysis of all quantitative variables. This involved calculating the mean, median, quartiles, and the 24 minimum and maximum values for each variable, allowing us to gain an overview of their distribution. We observed that for many financial ratios, there was a noticeable gap between the mean and the median, an indication of asymmetric distributions and the possible presence of extreme values. In addition, the widespread between the minimum and maximum values highlighted potential outliers, which could distort model estimates if not properly addressed during data preprocessing. We also examined the presence of missing values across the dataset. While the raw financial statement variables were complete, missing values appeared exclusively within the financial ratios. These were mostly caused by undefined calculations (such as division by zero). This confirmed the importance of applying an effective missing value handling approach that respects financial logic rather than relying on generic techniques. This descriptive analysis provided valuable insights into the dataset's structure and helped us prepare a cleaner and more reliable foundation for the subsequent modeling work.

```

Descriptive Statistics:
count      mean      std      min \
ANNEE_N    109381.0  2.014347e+03  5.812000e+00  2.003000e+03
CODACTIVBCT_BL 109381.0  4.578416e+04  2.000397e+04  0.000000e+00
EXPORT     109381.0  6.469047e+02  1.093467e+04  0.000000e+00
reserves   109381.0  -7.873025e+03  1.291061e+06  -3.700553e+08
Imm_corp_net 109381.0  1.090488e+05  1.116584e+07  -6.985800e+01
Participations 109381.0  6.084423e+03  8.260484e+05  -8.556015e+01
AI         109381.0  1.252227e+05  1.171604e+07  -4.000000e-01
AACTNC     109381.0  1.353970e+03  2.443089e+05  -1.318247e+03
ImNet      109381.0  1.265767e+05  1.177040e+07  -6.245113e+02
stocks     109381.0  1.076835e+05  1.305021e+07  -4.400000e+01
A_ACT_C    109381.0  5.258995e+04  8.636347e+06  -1.417111e+01
PLAC_A_ACTIF_FIN 109381.0  9.501015e+03  1.864836e+06  -3.418900e+04
Liq        109381.0  4.509445e+04  8.560177e+06  -4.170000e+02
Act_courant 109381.0  3.653149e+05  5.395267e+07  0.000000e+00
Tot_Bl     109381.0  4.699564e+05  5.987010e+07  0.000000e+00
Res_expl   109381.0  4.372567e+04  9.354483e+06  -1.055807e+09
PROD_PLC   109381.0  3.815998e+02  4.538904e+04  -1.509000e+03
RAIP       109381.0  3.132123e+04  8.057538e+06  -1.013287e+09
RN         109381.0  2.391560e+04  7.093808e+06  -1.015287e+09
Elem_extra 109381.0  -2.326083e+01  7.103407e+03  -2.334915e+06
div_distr  109381.0  -2.777103e+02  2.884053e+04  -9.421569e+06
REMB_EMP   109381.0  -2.327485e+04  3.546733e+06  -7.386939e+08
K_social   109381.0  3.867185e+04  5.154018e+06  -6.884526e+08
RES_REPORTE 109381.0  -2.500989e+04  1.033391e+07  -2.223656e+09
FPN        109381.0  6.366653e+04  1.683123e+07  -2.055813e+09
Fournisseurs 109381.0  1.772642e+05  2.441773e+07  -2.238900e+04
A_PASS_C   109381.0  8.242707e+04  8.565919e+06  -2.668926e+01
CONC_BANC A Passif Fin 109381.0  9.507521e+04  1.581974e+07  0.000000e+00

```

Figure 14: Descriptive Statistics for a Sample of the Variables

In the data understanding phase, I described the data using the method “. describe” to learn about the count, mean, standard deviation, minimum, maximum, and percentiles of the data.

```

Descriptive Statistics:
count      mean      std      min \
ANNEE_N    109381.0  2.014347e+03  5.812000e+00  2.003000e+03
CODACTIVBCT_BL 109381.0  4.578416e+04  2.000397e+04  0.000000e+00
EXPORT     109381.0  6.469047e+02  1.093467e+04  0.000000e+00
reserves   109381.0  -7.873025e+03  1.291061e+06  -3.700553e+08
Imm_corp_net 109381.0  1.090488e+05  1.116584e+07  -6.985800e+01
Participations 109381.0  6.084423e+03  8.260484e+05  -8.556015e+01
AI         109381.0  1.252227e+05  1.171604e+07  -4.000000e-01
AACTNC     109381.0  1.353970e+03  2.443089e+05  -1.318247e+03
ImNet      109381.0  1.265767e+05  1.177040e+07  -6.245113e+02
stocks     109381.0  1.076835e+05  1.305021e+07  -4.400000e+01
A_ACT_C    109381.0  5.258995e+04  8.636347e+06  -1.417111e+01
PLAC_A_ACTIF_FIN 109381.0  9.501015e+03  1.864836e+06  -3.418900e+04

```

Figure 15 : descriptive statistics of the dataset

I also displayed the number of missing values in each column using the method “.isnull().sum()”

```

Total Missing Values per Column:
ID-ENT                0
ANNEE_N              0
CODACTIVBCT_BL       0
SECTEUR              0
Appartenance Groupe  0
EXPORT              0
reserves             0
Imm_corp_net         0
Participations       0
AI                  0
AACTNC              0
ImNet               0
stocks              0
A_ACT_C             0
PLAC_A_ACTIF_FIN    0
Liq                 0
Act_courant         0
Tot_Bl              0
Res_expl            0

```

Figure 16 : displaying the total of missing values

6.3.2. Unit Consistency Check with K_social

Before doing any outlier treatment or clustering, I noticed that some companies report their balance sheet in different units, some years in thousands while others in dinars. This will lead to a large inconsistency during clustering, making the model unreliable. These suspicious jumps needed to be fixed. I used K_social (share capital) as a reference anchor, since it should remain relatively the same unless there is a true capital operation, and it doesn't really go from 100000 to 1000 in one year. By comparing year-to-year values of K_social within each company, I flagged for potential unit changes.

I used two rules to check for this inconsistency:

$$ratio = \max(current_k / prev_k, prev_k / current_k)$$

Equation 6.1: ratio check

If $ratio < 1/1000$ or $ratio > 1000$, the observation was flagged as inconsistent.

$$ratio_to_mode = \max(current_k / mode_k, mode_k / current_k)$$

If $ratio_to_mode > 1000$, the observation was flagged. Finally, to avoid excessive flagging, a refinement step unflags mild cases where $ratio_to_mode < 10$

Equation 6.2: mode check

The reason I used two formulas was that if, by chance, two years in a row were in a different unit from all the other years, the ratio formula wouldn't be able to detect the change because they are two successive years. For that, after checking with the ratio formula, I needed to

double-check using the mode and unflag or flag the values again. The resulting flags were merged back to the dataset (Données_financières_flagged.xlsx).

6.3.3. Correlation-based validation of flagging

To verify whether this step is crucial or not, I performed three correlation tests before and after filtering using the flag column, which are Pearson, Spearman, and Kendall. For each method, I computed correlation matrices and then measured differences. The formula used was

$$diff = |corr_full - corr_filtered|$$

Equation 6.3: absolute difference between correlations

This matrix measured how much each pair of variables changed after filtering. From it, I extracted the top 10 pairs with the largest shifts and calculated the average difference across all pairs.

	Variable 1	Variable 2	Difference
18963	CAP_PROP_CP	Part_FPN	1.054634
13413	Part_FPN	CAP_PROP_CP	1.054634
17943	DT_FPN	CAP_PROP_CP	0.986335
18993	CAP_PROP_CP	DT_FPN	0.986335
18969	CAP_PROP_CP	AE_FPN	0.982734
14319	AE_FPN	CAP_PROP_CP	0.982734
14282	AE_FPN	Part_FPN	0.956640
13382	Part_FPN	AE_FPN	0.956640
17906	DT_FPN	Part_FPN	0.950336
13406	Part_FPN	DT_FPN	0.950336

Figure 17: results of the absolute difference matrix

To go further, I tested statistical significance. For every pair of variables, I checked whether the correlation was significant (p-value ≤ 0.05) before and after filtering. This allowed me to quantify:

- the percentage of pairs that became significant after filtering,
- the percentage that lost significance,
- the percentage whose correlation strength increased.

I also generated an Excel sheet that contains all the interpretations:

Var1	Var2	earson_Fur	son_Filte	ue_Pearso	Pearson	earman_Fir	man_Filte	Spearman	Spearman	Interpretation
ANNEE_N CODACTIV		0,015913	0,015867	1,41E-07	1,6E-07	-0,01986	-0,01988	5,07E-11	5,13E-11	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N EXPORT		0,024273	0,024295	9,85E-16	1,02E-15	0,031794	0,031679	7,16E-26	1,25E-25	✅ Correlation remained significant and became stronger after filtering.
ANNEE_N reserves		-0,00765	-0,00651	0,011373	0,031647	-0,09224	-0,09195	3E-205	2,1E-203	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N Imm_corp		0,002859	0,005921	0,344308	0,050538	0,089717	0,089727	3E-194	9,1E-194	✅ No significance but the correlation strength improved after filtering.
ANNEE_N Participati		0,006991	0,007112	0,020767	0,018826	0,00518	0,005265	0,086657	0,082037	✅ Correlation remained significant and became stronger after filtering.
ANNEE_N AI		0,004812	0,008204	0,111525	0,006738	0,097683	0,097621	4,7E-230	3,9E-229	✅ Filtering revealed a significant correlation (flagging was useful).
ANNEE_N AACTNC		0,006398	-0,00613	0,034333	0,042926	-0,04449	-0,04476	4,6E-49	1,67E-49	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N ImNet		0,004922	0,00819	0,103535	0,006829	0,098055	0,098	8,3E-232	6,4E-231	✅ Filtering revealed a significant correlation (flagging was useful).
ANNEE_N stocks		0,011311	0,007388	0,000183	0,014686	0,070213	0,070199	1,4E-119	3,3E-119	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N A_ACT_C		0,008405	0,006231	0,005438	0,039598	0,181623	0,181553	0	0	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N PLAC_A_A		0,006896	0,004682	0,022564	0,12202	0,00794	0,008072	0,008644	0,007679	⚠️ Correlation lost significance after filtering (possible over-filtering).
ANNEE_N Liq		0,007129	0,005345	0,018386	0,077528	0,177018	0,177157	0	0	⚠️ Correlation lost significance after filtering (possible over-filtering).
ANNEE_N Act_coura		0,00874	0,006017	0,003843	0,046902	0,181197	0,181362	0	0	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N Tot_BI		0,00956	0,006666	0,001569	0,027697	0,175902	0,176004	0	0	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N Res_expl		0,009684	0,008888	0,00136	0,003332	0,139646	0,139496	0	0	⚠️ Correlation remained significant but got weaker after filtering.
ANNEE_N PROD_PLC		0,008052	0,006069	0,007746	0,045033	-0,00983	-0,00981	0,001144	0,001202	⚠️ Correlation remained significant but got weaker after filtering.

Figure 6: Interpretation of the tests

And a heatmap for each method before and after the flagging:

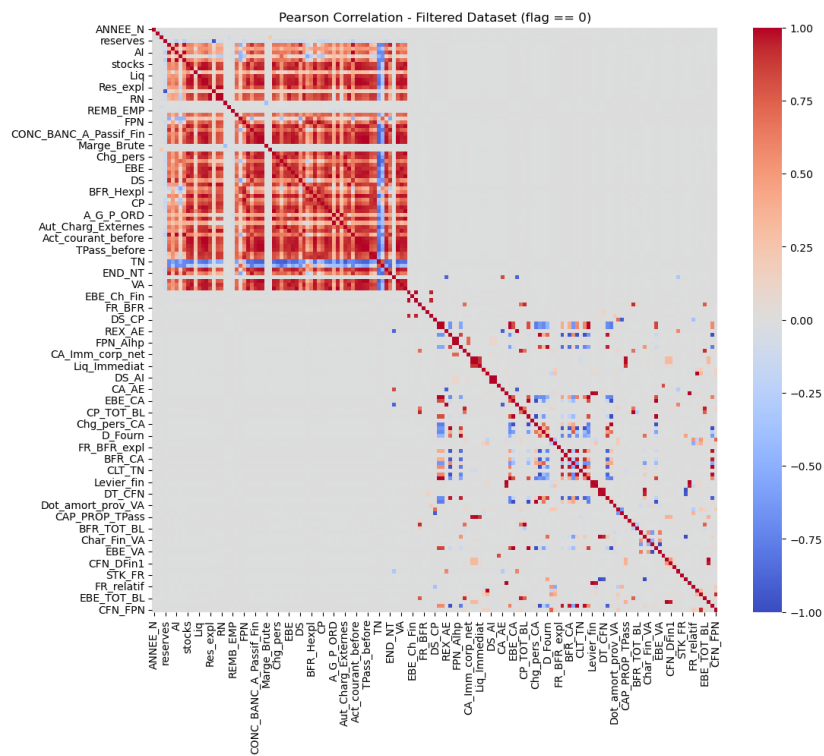


Figure 18: Pearson heatmap filtered dataset

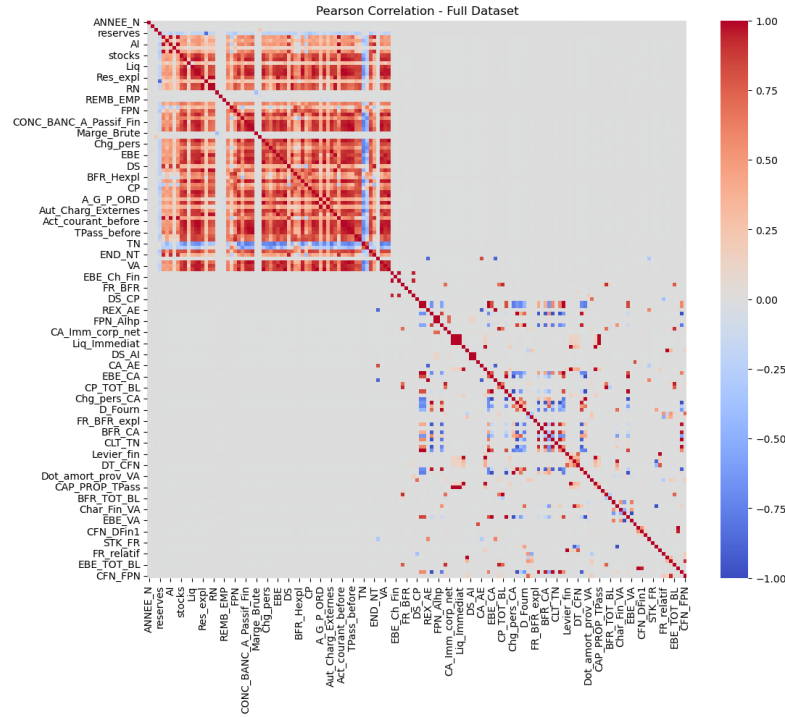


Figure 19: pearson heatmap full dataset

6.3.4. Outlier treatment

Next, I addressed outliers. Using the descriptive statistics, we can notice that there is a huge gap between the minimum and the maximum value for the same variable. I decided to group the data I grouped data by SECTEUR × CA_GROUP and applied Tukey bounds to each variable:

$$\text{Lower Bound} = Q1 - 1.5 \times IQR$$

$$\text{Upper Bound} = Q3 + 1.5 \times IQR$$

Equation 6.4: Tukey bounds

Values outside bounds were capped at Q1 or Q3 and flagged in an outlier column. This allowed me to control extreme outliers while keeping the data as real as possible. In financial situations, we never impute or delete variables; that's the golden rule for building a reliable and performant model. I also made sure to cap only the base variable because there is no need to fix ratios, as they depend on it.

6.3.5. Ratio recalculation

After capping the base variables, I made a list that contains all the financial ratios formulas and recalculated them using a safe division function to avoid division by zero problems. This was needed because the base variables were capped, so the ratios needed to match the base variables using the financial formulas.

6.3.6. Handling missing values

Missing values in financial ratios often stem from undefined operations, such as division by zero when a denominator like equity or interest expense equals zero. These missing entries are not considered random but carry meaningful financial signals, such as zero equity, indicating extreme financial distress. Instead of using standard imputation techniques (e.g., mean or median), a domain-knowledge-based approach was adopted. The logic was formalized in a custom Python function `safe_divide()`, designed to handle division cases in a way that preserves financial interpretation:

- When the nominator is negative, but the denominator is zero, we replace it with the minimum value of that ratio.
- When the nominator is positive, but the denominator is zero, we replace it with the maximum value of that ratio
- When both are zero, we put zero, since the absence of both numerator and denominator reflects a neutral financial situation.
- In all other cases, the ratio was calculated normally.

By doing so, I ensured that no ratio was lost due to missing values, while at the same time respecting the **financial interpretation** of extreme situations. The final dataset (`Données_financières_ratios_calculated.xlsx`) contained a consistent and economically meaningful set of variables, which were then used in the dimensionality reduction and clustering stages.

6.3.7. Standardization

After cleaning and recalculating all the financial ratios, the next step is to standardize the data. Some ratios, like liquidity ratios, usually stay close to 1, while others, like leverage, can reach hundreds. If I used the raw data, variables with larger magnitudes would dominate the clustering results.

To avoid this, I applied z-score standardization to all numeric variables using the StandardScaler class from scikit-learn. The formula is :

$$z = (x - \text{mean}(x)) / \text{std}(x)$$

Equation 6.5: z-score standardization

6.3.8. Feature Selection, Dimensionality Reduction and Clustering

The data set contained over 80 ratios, all highly correlated. To solve this issue and simplify data without losing information, I went through many steps.

I first removed columns that were not useful for clustering, such as identifiers (ID-ENT, ANNEE_N), sector codes, and flags (flag, outlier, CA_GROUP). This left me with a clean set of purely numeric financial ratios.

I then tried different approaches for dimensionality reduction and clustering.

6.3.9. Approach A: PCA only + K-means

The first approach consisted of using PCA to reduce the dimensionality of the data. I chose to keep the top 50 components, which captured most of the variance of the original data. The formula for PCA is:

$$Z = X \times W$$

Equation 6.6: PCA formula

I then applied k-means clustering with k=4.

6.3.9.1. Approach B: PCA + UMAP + K-Means

The second approach is to reduce using PCA and then feed UMAP the results of PCA. I then applied k-means with k=4 on the UMAP projected data.

6.3.9.2. Gaussian Mixture Model (GMM) on PCA

The third approach used a Gaussian mixture model applied to the PCA-reduced components (50 components).

6.3.9.3. Pearson p-Value Filtering + K-Means

In this method, I computed the Pearson correlation coefficient and dropped highly correlated ratios. The formula is as follows:

$$\rho(X,Y) = cov(X,Y) / (\sigma_X \times \sigma_Y)$$

Equation 6.7: Pearson correlation coefficient

After filtering, I standardized the reduced dataset and applied K-Means (k = 4) with the same objective function

7. RESULTS AND FINDINGS

7.1. Overview

The core objective of this stage was to evaluate whether unsupervised learning methods could meaningfully segment BIAT's corporate clients based on financial ratios. After preprocessing (unit correction, outlier treatment, and missing value handling), several clustering strategies were tested. Each method produced different clusters with varying level of robustness. The following section presents the result of each method and the decision-making method of choosing the best out of them.

To provide clarity, the main steps of the modeling pipeline are summarized in the figure below:

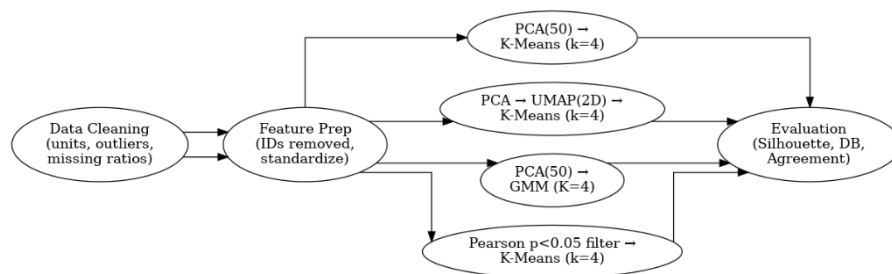


Figure 20 : Modeling Pipeline

7.2. PCA only + K-means

For the first approach applied Principal Component Analysis (PCA) was applied to reduce dimensionality before clustering with K-Means. PCA compressed the information from 157 variables into 50 principal components, capturing the major sources of variance in the dataset. I then applied k-means to group the companies into 4 different clusters. The results were as follows:

- Cluster 0: Firms with balanced performance. Profitability, liquidity, and solvency were all close to the average, and leverage remained at a reasonable level. This group reflects financially healthy companies.
- Cluster 1: Firms with extreme values. They showed extremely high profitability and solvency, combined with almost no leverage. These cases are unusual and may come from specific sectors or denominator effects.
- Cluster 2: Firms with weak profitability and negative solvency, though leverage stayed moderate. These companies are in a vulnerable position.

- Cluster 3: Firms under financial stress. Profitability was close to zero or negative, and leverage was extremely high. These firms rely heavily on debt while generating little to no returns, giving them the highest credit risk.

Indicator	Cluster 0	Cluster 1	Cluster 2	Cluster 3
ROA	0.26	904.02	0.00	-0.00
EBE/CA	0.10	0.20	-0.03	0.07
Liquidity	1.27	0.00	0.68	1.02
FR/BFR	0.78	-3.45	1.09	0.28
Solvency	0.30	762.53	-0.36	0.27
Leverage	8.38	0.01	7.11	43.74

Table 6: CA-only Cluster Profiles (Median Values)

7.3. UMAP + PCA + K-Means

In the second approach, I combined Uniform Manifold Approximation and Projection (UMAP) with PCA before clustering. PCA first reduced the dataset into principal components, while UMAP projected them into a two-dimensional space. Afterward, I applied K-Means algorithm. The results provided clearer visual separation of firms, although the quality metrics were weaker than PCA-only.

- Cluster 0: Firms with average performance but weaker client turnover, meaning they take longer to collect from clients.
- Cluster 1: Firms with stable profitability and solvency, representing low credit risk.
- Cluster 2: Companies with liquidity difficulties but higher value-added relative to sales, showing mixed financial health.
- Cluster 3: Firms with above-average profitability and stable solvency, also in a healthy position.

Indicator	Cluster 0	Cluster 1	Cluster 2	Cluster 3
ROA	0.23	0.17	0.12	0.62
EBE/CA	0.10	0.09	0.12	0.10
Liquidity	1.26	1.24	0.93	1.46
FR/BFR	0.72	0.65	0.98	0.86
Solvency	0.28	0.31	0.27	0.38
Leverage	8.89	8.21	6.98	8.16

Table 7: UMAP + PCA Cluster Profiles (Median Values)

7.4. Pearson P-Value Filtering + K-Means

The third approach focused on reducing redundancy in the data. I removed variables that were highly correlated. It was based on p-value ($p < 0.05$). This produced a smaller set of features, which was then clustered using K-Means. This generated 4 clusters, 3 of which were similar.

- Clusters 1, 2, and 3: All showed financially healthy firms, with positive ROA, good solvency, and manageable leverage.
- Cluster 0: Firms with weaker performance, showing lower profitability and liquidity, making them more vulnerable.

Indicator	Cluster 0	Cluster 1	Cluster 2	Cluster 3
ROA	0.27	0.08	0.08	0.08
EBE/CA	0.10	0.12	0.12	0.13
Liquidity	1.27	1.35	1.31	1.32
FR/BFR	0.79	0.42	0.48	0.43
Solvency	0.30	0.39	0.41	0.39
Leverage	8.46	5.72	6.50	4.47

Table 8: Pearson-Filtered Cluster Profiles (Median Values)

7.5. Gaussian Mixture Model (GMM) on PCA

Finally, I tested a Gaussian Mixture Model (GMM) on PCA-reduced data. Unlike K-Means, GMM allows clusters to have different shapes and variances.

This method gave us the below clusters:

- Cluster 0 and 1: Firms with relatively good profitability and solvency, considered low risk.
- Cluster 2: Firms with average indicators but slightly higher leverage, indicating moderate financial health.
- Cluster 3: Firms with higher risk, although still less extreme than in K-Means clustering.

Indicator	Cluster 0	Cluster 1	Cluster 2	Cluster 3
ROA	0.07	904.02	0.32	0.33
EBE/CA	0.09	0.20	0.11	0.11
Liquidity	1.19	0.00	1.14	1.31
FR/BFR	0.72	-3.45	0.92	0.75
Solvency	0.32	762.53	0.26	0.31
Leverage	5.94	0.01	10.60	8.82

Table 9: GMM Cluster Profiles (Median Values)

7.6. Comparison Across Methods

To test the stability of results, I compared the cluster assignments from the three main approaches: PCA-only, UMAP + PCA, and Pearson filtering. The comparison showed that PCA and Pearson were highly consistent, while UMAP introduced more variability.

- About 98.83% of firms were assigned to the same cluster in both PCA-only and Pearson approaches.
- Between PCA and UMAP, agreement dropped to 44.78%, showing that UMAP reshuffled firms more strongly.
- Between Pearson and UMAP, agreement was similar (45.24%).
- Only 44.77% of firms stayed in the same cluster across all three methods.

Comparison	Agreement (%)
Same across all methods	44.77%
PCA vs. UMAP	44.78%
UMAP vs. Pearson	45.24%
PCA vs. Pearson	98.83%

Table 10: Agreement Between Clustering Methods

This shows that PCA and Pearson filtering produce stable, overlapping clusters, while UMAP generates more variability. In practice, PCA and Pearson can be trusted for consistent segmentation, while UMAP is better suited for visualization.

7.7. Cluster Quality Metrics

I also measured the quality of the clusters using two well-known metrics: Silhouette score (higher is better) and Davies–Bouldin index (lower is better).

Method	Silhouette	Davies–Bouldin
PCA-only + KMeans	0.865	0.588
UMAP + PCA + KMeans	0.486	0.746
Pearson Filtering + KMeans	0.992	0.418
GMM on PCA	0.299	3.225

Table 11: Cluster Quality Metrics

The Pearson method achieved the highest Silhouette score (0.992), but its clusters were unrealistic because of extreme outliers. PCA-only offered the best balance between quality and interpretability, while UMAP was moderate and mainly useful for visualization. GMM performed poorly and was not suitable for this dataset.

7.8. Credit Decision Mapping

To make the results actionable for BIAT's risk department, I mapped the PCA-only clusters into credit decision categories. This step shows how clustering can directly support lending credit decisions.

- Cluster 0 and Cluster 1: Firms with solid profitability, liquidity, and solvency. Classified as Strong candidates – approve.
- Cluster 2: Firms with weak profitability and negative solvency, but leverage under control. Classified as Moderate risk – requires further review.

- Cluster 3: Firms with very weak or negative profitability and extremely high leverage. Classified as High risk – likely reject.

Cluster	Profile	Credit Decision
0	Balanced financial indicators	Strong candidate – Approve
1	Extremely high returns, low debt	Strong candidate – Approve (with caution)
2	Weak profitability, low solvency	Moderate risk – Further review
3	Negative returns, high leverage	High risk – Reject

Table 12: Credit Decision Mapping for PCA-only Clusters

7.9. Summary

The experiments demonstrated that unsupervised learning could provide valuable insights into BIAT's corporate client portfolio.

- PCA-only clustering was the most reliable method: interpretable, stable, and consistent with financial logic.
- Pearson filtering produced excellent metrics but unrealistic clusters due to outliers.
- UMAP offered clear visualization but lacked stability.
- GMM performed poorly and is not recommended for this dataset.

By mapping clusters to credit decisions, the analysis showed how unsupervised learning can complement traditional risk assessment. It can help identify healthy firms for faster approval, highlight moderate-risk cases for closer monitoring, and flag stressed firms that pose high credit risk.

This integration of machine learning into BIAT's risk framework has strong potential to improve portfolio monitoring, detect early warning signs, and support data-driven decision-making in credit risk management.

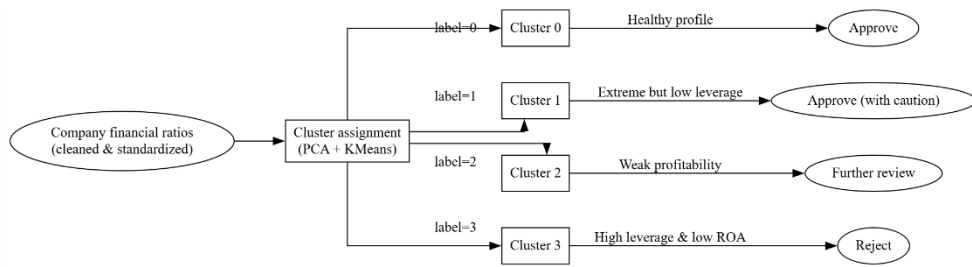


Figure 21: Credit decision mapping based on PCA + KMeans cluster label.

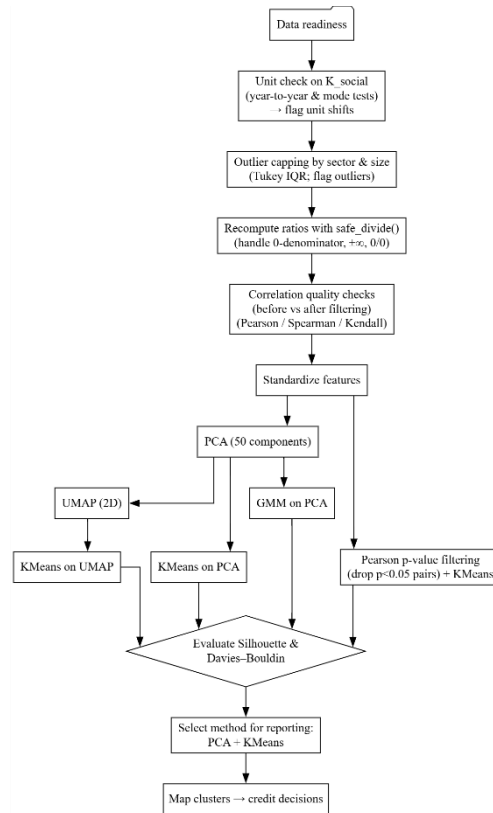


Figure 22: End-to-end unsupervised risk segmentation pipeline used in this study

8. RECOMMENDATIONS

Although the modeling approach presented in this study provides satisfactory performance in predicting corporate credit risk, several areas for improvement have been identified, both in terms of data quality and model design. These recommendations aim to enhance the predictive power, robustness, and applicability of the model in a real-world banking context.

First, particular attention should be given to the quality and integrity of the financial data used in model construction. Throughout the data preparation process, we encountered an enormous number of entry anomalies, such as implausible ratio values or missing denominators, which may distort the estimation of default probabilities. Ensuring rigorous data validation protocols, both automated and manual before modeling, is essential to limit noise and prevent misclassification caused by faulty inputs. Future implementations should prioritize the verification of financial ratios at the source and consider internal audit flags to better isolate abnormal entries.

Second, while this project focused exclusively on financial ratios derived from accounting statements, it is important to acknowledge that creditworthiness is influenced by a broader set of company characteristics. Factors such as the age of the enterprise, the number of employees, and client-specific behavioral data such as payment history and the frequency of account movements, can provide complementary signals that improve prediction accuracy. Integrating such indicators into future modeling efforts would offer a more performant assessment of clustering, especially for companies with atypical financial structures.

Third, while the segmentation based on revenue thresholds (TPE, PME, and Corporate) has proven useful in improving model discrimination, this dimension alone may not capture the full spectrum of risk heterogeneity across firms. Deeper segmentation strategies, particularly those that account for sector-specific dynamics, could better reflect the economic environment in which a company operates. Incorporating indicators related to the competitiveness, asset structure, and market positioning of each sector would likely refine risk classification and improve the relevance of credit decisions.

Finally, the inclusion of macroeconomic and contextual variables should be considered in future iterations of the model. Changes in interest rates, inflation, or other systemic economic shifts can significantly impact a firm's probability of default. Incorporating macro-financial indicators would make the model more adaptive to shifts in economic cycles and improve predictive robustness in times of volatility or crisis.

9. CONCLUSIONS

This internship has been an enriching experience. It allowed me to get exposure to the financial world, while combining my two areas of interest which are IT and finance. The journey started with a major challenge: how to handle a complex and large dataset? But with the guidance of my supervisor, I was able to overcome it by following the steps he provided. I spent quite a lot of time understanding the dataset I was given, I went through every variable diving them into qualitative and quantitative. I then had to create my own dictionary that contains the variable name as listed in the dataset, translate it into English and look for its meaning, interpretation and formula in the case of ratios.

The next step was to check unit consistency, for this, I developed a custom python algorithm that compares k-social year to year and flag if there is an abnormal jump. I then compared them to the mode for a second check. Then, I had to deal with outliers, for this, I also had to get use of financial sense, since ready techniques like KNN imputation don't provide accurate results on financial data. I also had to come up with an idea that caps outliers without affecting data accuracy. Next, I had to deal with missing values which were a result of incomputable ratios like division by zero problems or division by missing values.

Once the data was ready, I moved to the clustering techniques where I applied different ones. I had to first reduce the dimensionality of the data once using PCA and UMAP and once using p-value of the correlation matrix and pick the best one. Second, I applied K-mean, GMM and DBSCAN. The analysis revealed important insights: Pearson filtering gave the best scores numerically but produced unrealistic clusters; UMAP helped with visualization but lacked stability; GMM proved less effective for this dataset. In contrast, PCA-only clustering struck the right balance, delivering interpretable clusters that clearly separated financially healthy firms from those under stress.

The method that I approved which is PCA + k-means provided 4 clusters: the first one has balanced financial indicators which make It a strong candidate , the second one has extremely high returns, low debt which makes it a strong candidate, the third one has weak profitability, low solvency which would classify as moderate risk, and the fourth one has negative returns, high leverage making it fall into the high risk category and thus get rejected.

Beyond the technical part, this report provided a comprehensive literature review that contains in-depth research into the techniques used in this project as well as more advanced techniques to be integrated. I also provided a recommendation part where I insisted on improving data gathering and data quality as well as integrating other indicators besides the financial ratios such as client behavioral data, inflation rates and interest rates.

REFERENCES

- [1] Banque Internationale Arabe de Tunisie (BIAT), *Rapport sur la gestion de la banque 2024*. Tunis, Tunisia. Accessed: Sept. 6, 2025. [Online]. Available : <https://www.biat.com.tn/communication-financiere/rapports-annuels>
- [2] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.
- [3] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
- [4] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for dimension reduction," *arXiv preprint*, arXiv:1802.03426, 2018.
- [5] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, Berkeley, CA, USA, 1967, pp. 281–297.
- [6] D. Reynolds, "Gaussian Mixture Models," in *Encyclopedia of Biometrics*, S. Li and A. Jain, Eds. Boston, MA, USA: Springer, 2009, pp. 659–663.
- [7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
- [8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
- [9] N. Ng, "A Guide to Supervised Learning," *Medium*, 2019. Accessed: Sept. 6, 2025. [Online]. Available: <https://medium.com/@ngneha090/a-guide-to-supervised-learning-f2ddf1018ee0>
- [10] I. A. Glover and P. M. Grant, *Digital Communications*, 3rd ed. Harlow, U.K.: Prentice Hall, 2009.