Segmenting Banking Churn: A Comparative Analysis of K-Validation and Clustering Methods

Candidate:

Alessia Ronchetti

Index

1. Introduction

1.1 Context, Objectives, and Tools

In the current, highly competitive banking landscape, customer retention has become a fundamental strategic pillar. The cost of acquiring a new customer is significantly higher than maintaining an existing one. Consequently, the phenomenon of **customer churn,** the abandonment of services by clientele, represents one of the most critical economic and strategic challenges for financial institutions. Understanding *why* and *which* customers are most likely to leave is the first step in designing effective retention strategies.

The primary objective of this project is to shift from a general analysis to a specific customer segmentation. By using data mining techniques, particularly unsupervised clustering algorithms, this project aims to identify homogeneous groups of customers based on their behaviors and characteristics. This segmentation will allow for the profiling of different clusters, analyzing the incidence of churn within each group, and thus providing a concrete basis for targeted marketing and service interventions.

To conduct this analysis, a public dataset named Churn_Modelling.csv was used. It describes the behavior of 10,000 bank customers, including demographic and financial variables, as well as their final status (whether they exited or not). The entire process was implemented using the Python programming language, with the support of the scientific libraries Scikit-learn, Pandas, and Matplotlib.

The complete code is available at:
https://colab.research.google.com/drive/1Nae0C16p79RJY0fCrCHFAwhVqjXnnbP7

1.2 Proposed Methodology: Validation and Comparison

This project's methodology is based on four fundamental pillars, following a structured Knowledge Discovery in Databases (KDD) approach:

1. **Exploratory Data Analysis (EDA):** Initial inspection and visualization of the Churn_Modelling.csv dataset to understand the distributions of key variables and identify the first correlations with the churn phenomenon.
2. **Data Preprocessing:** Preparing the data for the clustering algorithms. This phase included handling categorical variables (like 'Geography' and 'Gender') through numerical encoding and Feature Standardization. This last step is technically essential for distance-based algorithms like K-Means, to prevent variables with different scales (e.g., 'Balance' vs. 'Age') from dominating the analysis.

3. **K-Means Study and Robust K-Validation:** K-Means was chosen as the *baseline* algorithm for its efficiency and interpretability. However, its main weakness lies in the need to define the number of clusters (K) *a priori*. To overcome the ambiguity of the "Elbow Method" alone, a robust validation analysis was conducted. The optimal K value was sought through the convergence of three distinct metrics: the inertial method (Elbow), the Silhouette Score (which measures cluster cohesion and separation), and the Davies-Bouldin Index (DBI) (which rewards compact and well-separated clusters).
4. **Extended Methodological Comparison:** To ensure that the identified clusters were not merely an artifact of the K-Means algorithm, the *baseline* segmentation was compared with three conceptually different algorithms, as implemented in the code:
    - **Agglomerative Hierarchical Clustering:** To analyze the data's hierarchical structure (visualized via dendrogram) and compare a *bottom-up* approach with the K-Means partition.
    - **DBSCAN:** To test a density-based approach, capable of identifying outliers and arbitrarily shaped clusters without needing to pre-define K (though requiring calibration of the eps parameter).
    - **BIRCH:** A hybrid approach (both hierarchical and partitional) known for its efficiency and scalability on medium-to-large datasets.

The purpose of this extended comparison is not to declare an absolute "winner," but to validate the stability of the customer segments found and to understand which clustering paradigm offers the most interpretable and strategically useful segmentation for churn analysis.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the fundamental first step to understand the structure and characteristics of the Churn_Modelling.csv dataset before proceeding with clustering.

2.1 Structure and Descriptive Statistical Analysis

The initial inspection revealed a clean and complete dataset. It contains 10,000 observations (customers) and 14 columns (variables). The columns include identifiers to be removed (RowNumber, CustomerId, Surname), categorical variables to be transformed (Geography, Gender), and numerical variables that will be used for the analysis.

The aggregate statistical analysis provides an initial numerical summary.

	count	mean	std	min	25%	50%	75%	max
RowNumber	10000.00	5000.50	2886.90	1.00	2500.75	5000.50	7500.25	10000.00
CustomerId	10000.00	15690940.57	71936.19	15565701.00	15628528.25	15690738.00	15753233.75	15815690.00
CreditScore	10000.00	650.53	96.65	350.00	584.00	652.00	718.00	850.00
Age	10000.00	38.92	10.49	18.00	32.00	37.00	44.00	92.00
Tenure	10000.00	5.01	2.89	0.00	3.00	5.00	7.00	10.00
Balance	10000.00	76485.89	62397.41	0.00	0.00	97198.54	127644.24	250898.09
NumOfProducts	10000.00	1.53	0.58	1.00	1.00	1.00	2.00	4.00
HasCrCard	10000.00	0.71	0.46	0.00	0.00	1.00	1.00	1.00
IsActiveMember	10000.00	0.52	0.50	0.00	0.00	1.00	1.00	1.00
EstimatedSalary	10000.00	100090.24	57510.49	11.58	51002.11	100193.91	149388.25	199992.48
Exited	10000.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00

Table 2.1: Descriptive Statistics of Numerical Variables.

Three key insights emerge from this table:

1. **Churn Rate (Exited):** The mean of Exited is 0.2037, indicating that **20.37%** of the customers in the dataset have left the bank. This is our benchmark figure.
2. **Balance:** The 25th percentile (25%) is 0.00. This means that at least a quarter of the clientele has a zero balance, a piece of information that will be better visualized in the graph below.
3. **Age:** The mean age is approximately 39 years (median 37), with a range from 18 to 92 years.

2.2 Visual Analysis

To better understand the shape of the distributions, histograms were generated for numerical variables and bar charts for categorical ones. The three most relevant graphs for the project are commented on below.

The most important graph is that of the Exited variable, which shows the imbalance between active customers and those who have churned.

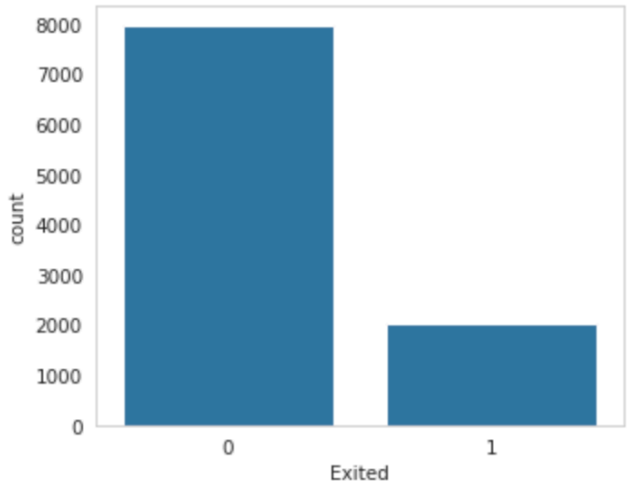

Figure 2.1: Distribution of the target variable 'Exited' (Churn).

The graph visually confirms the **20.37% churn rate**. The dataset presents a moderate imbalance (approximately 80% retained customers and 20% exited). This average churn rate will be the **baseline** for evaluating the clusters: a cluster will be defined as "at-risk" if its internal churn rate is significantly higher than this average.

The distributions of the Age and Balance variables are crucial for understanding the customer profile.

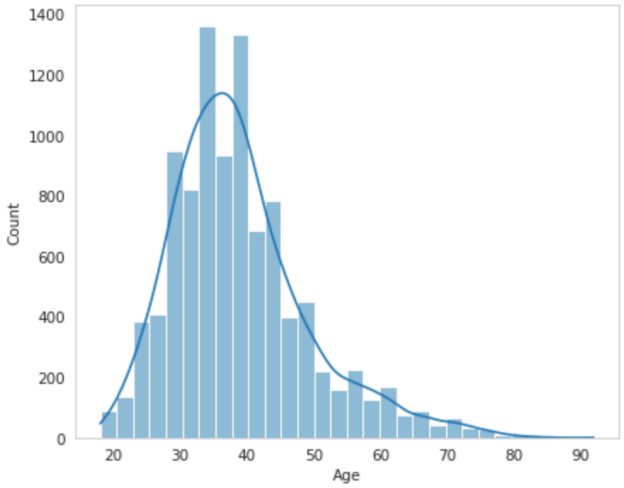

Figure 2.2: Distribution of Age.

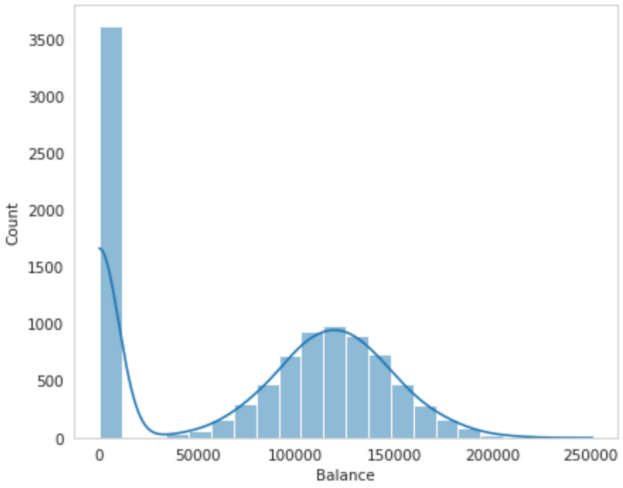

Figure 2.3: Distribution of Balance.

The analysis of the Age and Balance distributions is crucial. Age (Figure 2.2) shows a core clientele between 30-40 years old and a right-skewed tail of older customers. Even more informative is the Balance variable (Figure 2.3), which exhibits a strongly **bimodal distribution**: a massive peak at the 0 value (over 3,500 customers) and a near-normal distribution centered around 120k for positive balances. This visualization is fundamental, as it suggests that 'Balance' could represent two distinct phenomena: *if* a customer has a balance and *how much* that balance is.

At this point, it is important to analyze the interaction with the target variable Exited. For this, boxplots are ideal, as they show the distribution of a numerical variable (like Age or Balance) split by the two Churn categories (0 and 1).

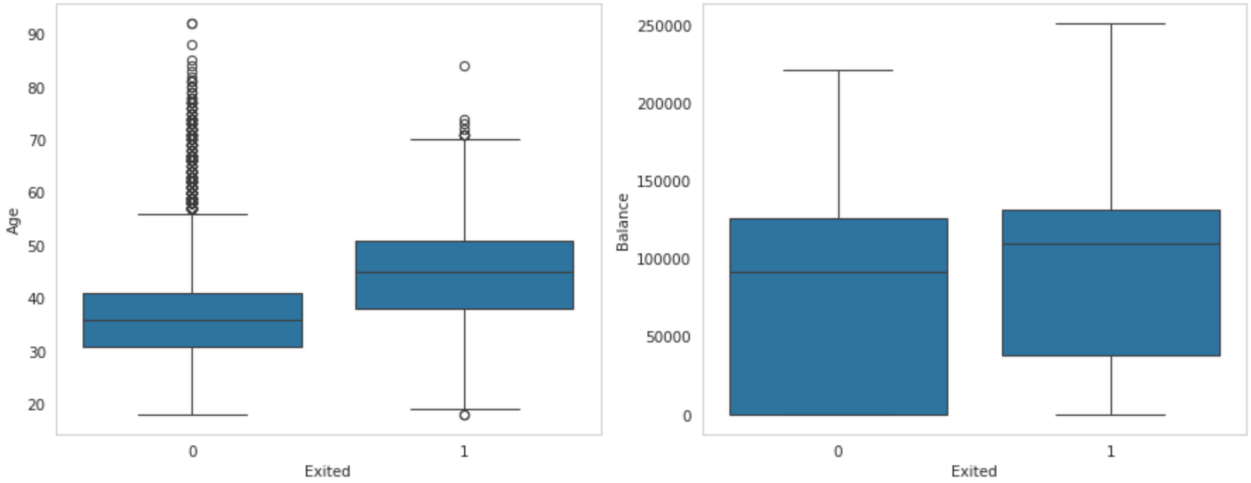

Figure 2.4: Relationship between Age and Churn (Exited).

Figure 2.5: Relationship between Balance and Churn (Exited).

These two graphs provide the first real indications of at-risk profiles:

- **Age (Figure 2.4):** A clear difference emerges. As noted, customers who churned (Exited=1) tend to have a higher mean and median age than those who stayed (Exited=0).
- **Balance (Figure 2.5):** The balance boxplot is even more interesting. It is observed that the median balance for customers who churned (Exited=1) is slightly higher than that of those who stayed. This strongly suggests that customers with a zero balance are, proportionally, less likely to churn. Churn appears to be more frequent among customers who have an active balance in their account.

3. Data Preparation and K Validation

The objective of this phase is twofold: first, to transform the raw data into a numerical, standardized format suitable for clustering (Preprocessing); second, to objectively and robustly identify the optimal number of clusters (K) for our baseline K-Means model.

3.1 Preparation: Feature Cleaning and Encoding

The first step is to prepare the dataset for analysis. This required two interventions:

- **Cleaning:** Variables that provide no informational value for customer segmentation were eliminated, specifically RowNumber, CustomerId and Surname.
- **Encoding and Selection:** To manage the categorical (textual) variables, which K-Means cannot process, the following steps were taken:
    - **'Geography' (Removal):** Although it was correlated with churn in the EDA, testing revealed that its inclusion (via One-Hot Encoding) did not lead to a significant improvement in validation metrics (e.g., Silhouette) and, in fact, "muddled" the definition of the other clusters. It was therefore decided to remove it to create a model more focused on purely behavioral and financial variables.
    - **'Gender' (Encoding):** The Gender variable (Male/Female) was transformed into a binary numerical format (e.g., Female=0, Male=1).

3.2 Standardization

The final fundamental step in data preparation is **standardization**. This process is crucial for the K-Means algorithm, which is based on Euclidean distance. Variables on different scales (e.g., Balance in the hundreds of thousands and Age in tens) can distort the analysis by giving more weight to variables with greater magnitude.

To ensure that every variable contributes equally to the distance calculation, the StandardScaler from the Scikit-learn library was applied. This process generates a new DataFrame, df_scaled, in which all columns have a mean of 0 and a standard deviation of 1.

(Note: The 'Exited' variable, our "target", was excluded from this process. It is not used to create the clusters, but only afterward to validate and interpret the groups).

3.3 K Validation: Limitations and Robust Metrics

Once the data is ready, the main challenge of K-Means is choosing *K* (the number of clusters). The initial analysis planned to use only the **Elbow Method (Inertia)**, a useful but often ambiguous heuristic technique, as identifying the "elbow" can be subjective.

To "lock in" the choice of K=4 and make it objective, two internal validation metrics were introduced, which evaluate the geometric quality of the clusters:

1. **Silhouette Analysis (Silhouette Score):** Measures the quality of a cluster by evaluating two factors for each point: **Cohesion** (how similar a point is to members of its own cluster) and **Separation** (how dissimilar it is from members of the nearest cluster). One seeks the K with the highest score (close to +1).
2. **Davies-Bouldin Index (DBI):** Measures the average "similarity" between each cluster and its "worst enemy" (the cluster most similar to it), based on the ratio of intra-cluster distances to inter-cluster distances. One seeks the K with the lowest score (close to 0).

3.4 Validation Results

The execution of the three validation metrics (Inertia, Silhouette, and Davies-Bouldin) produced the following results:

- The **Elbow Method** (Inertia) showed its clearest bends at **K=4** and **K=5**.
- The **Silhouette Analysis** showed its peak scores (highest values) at **K=3** and **K=10**.
- The **Davies-Bouldin Index** showed its minimum points (lowest values) at **K=5**, **K=9**, and **K=10**.

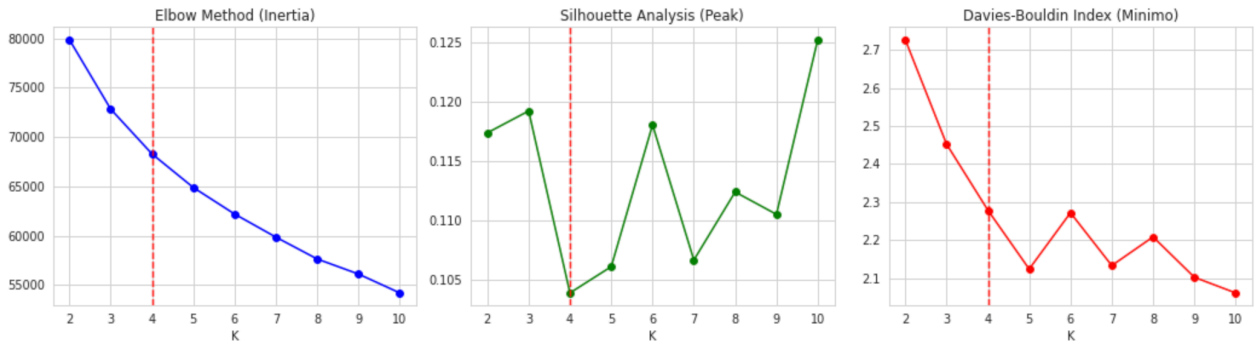

Figure 3.1: K Validation (Elbow, Silhouette, Davies-Bouldin).

Since these three distinct metrics are based on different mathematical principles and point to different optimal values, the choice of **K=4** is considered the best strategic estimation for the clustering. The reasoning behind this decision is explained in the analysis below.

3.5 Strategic Selection and Boundary Analysis

Since there was no single mathematical convergence, the final selection of **K=4** relied on a "process of elimination" based on business logic and cluster stability.

Here is why we rejected the other options and selected K=4:

- **Why we rejected K=10:**
  Both Silhouette and Davies-Bouldin suggested that K=10 was mathematically excellent. However, for a bank, having 10 different customer segments is too complex. It represents “micro-segmentation", which is inefficient to manage operationally. We cannot create 10 different retention campaigns.

- **Why we rejected K=3:**
  The Silhouette analysis also showed a high score for K=3. However, this configuration was too general. It merged "High Risk" and "Medium Risk" customers into one large group, hiding the specific details we need to predict churn effectively.

- **Why we rejected K=5:**
  We specifically compared K=4 against K=5 (which was supported by the Davies-Bouldin Index). We rejected K=5 because the Silhouette score dropped, indicating less compact groups. More importantly, the new 5th cluster was just a duplicate sub-group of the existing "Standard Customers”, adding complexity without adding new insight.

- **Conclusion: Why K=4 is the Winner:**
  K=4 is the best compromise. It is supported by the Elbow Method (which shows a clear bend) and offers the "operational sweet spot”. It creates groups that are distinct enough to be actionable (unlike K=3) but simple enough to be managed by the marketing team (unlike K=10).

4. K-Means Analysis: Baseline Model and Profiling

After preparing the data and identifying K=4 as the optimal number in Chapter 3, we proceeded with the application of our baseline algorithm, K-Means.

This chapter first describes the algorithm's operational logic (4.1) and then analyzes in detail the four customer segments that emerged (4.2 - 4.4). The objective is to determine if the segmentation was successful and what strategic profiles were identified.

4.1 Theoretical Foundations: What is the K-Means Algorithm?

K-Means was chosen as the baseline algorithm for its efficiency and interpretability. It is a partitional and centroid-based algorithm:

- **Partitional:** This means it divides the dataset into a predefined number *K* of clusters, with no overlaps (each customer will belong to only one cluster).
- **Centroid-based:** Each cluster is represented by a "centroid," which is the mean point (the average) of all customers belonging to that cluster.

The algorithm works **iteratively** to find the best centroids, specifically those that minimize the total **inertia** (the sum of squared distances between each customer and their own centroid):

1. **Initialization:** The algorithm places K (in our case, 4) centroids in the data space.
2. **Assignment:** It calculates the Euclidean distance of each customer from every centroid and assigns each customer to the nearest centroid. (This step makes the **standardization** done in Chapter 3 essential, to prevent variables on different scales, like Balance and Age, from dominating the calculation).
3. **Update:** It recalculates the position of each centroid as the mean of all customers just assigned to it.
4. **Convergence:** It repeats steps 2 and 3 until the centroids stop moving, thus finding a stable configuration.

The result is a model that groups customers into spherical, compact, and well-separated clusters.

4.2 Execution and Churn Rate Analysis

Applying this logic, the final K-Means algorithm was run on the standardized df_scaled dataset with K=4. Using the .fit_predict() function, each of the 10,000 customers was assigned a membership label (a value from 0 to 3). The crucial analytical step was to add this new Clustercolumn to the original df DataFrame.

The primary objective was to determine if the clustering could isolate segments with different churn rates. Since Exited is a binary variable (1=Exited, 0=Remained), its mean corresponds precisely to the Churn Rate of that segment. This calculation is compared against the dataset's baseline (identified in Chapter 2), which is 20.37%.

Cluster	Churn Rate (Exited)
3	32.07%
1	19.94%
2	15.51%
0	13.48%

Table 4.1: Churn Rate by Cluster (K-Means).

The results are clear. The algorithm successfully isolated a high-risk Cluster 3 (32.07%), a full 11 percentage points above the average, and a loyal Cluster 0 (13.48%).

4.3 Strategic Impact: Size and Priority

After identifying the risk, we evaluated its strategic impact by analyzing its size.

Cluster	Risk Profile	Customer Count	Size (% of Total Customers)
Cluster 3	Critical Risk	2,429	24.29%
Cluster 1	Medium Risk	2,899	28.99%
Cluster 2	Low Risk	2,476	24.76%
Cluster 0	Loyal Customers	2,196	21.96%

Table 4.2: Cluster Size and Business Impact (K-Means).

Impact Discussion: The size analysis confirms the criticality of Cluster 3. This is not a small niche: with 2,429 members, it represents 24.29% of the entire customer base. This transforms a statistical insight into a business priority: almost one in four customers belongs to a high-risk segment.

4.4 Cluster Profiling

To understand *who* these customers are, we analyzed the mean profile of each segment.

Variable	Cluster 3 (Critical Risk)	Cluster 2 (Low Risk)	Cluster 0 (Loyal Customers)	Cluster 1 (Medium Risk)
Exited (Churn)	0.32	0.16	0.13	0.20
IsActiveMember	0.00	1.00	0.53	0.52
Balance	103,858.89	105,522.14	10,526.11	78,715.94
NumOfProducts	1.26	1.28	2.15	1.50
HasCrCard	1.00	1.00	0.98	0.00
Age	38.64	39.82	38.02	39.07

Table 4.3: Mean Cluster Profile (K-Means) (Key variables).

Descriptive Interpretation of the Segments:

- **Cluster 3 (Critical Risk):** The "Ghost" Customers (32.07% Churn): This segment is defined by a unique combination: it has a high average balance (€104k) but is characterized by total inactivity (IsActiveMember = 0.00).
- **Cluster 2 (Low Risk):** The "Active High-Balance" Customers (15.51% Churn): This group has a similar financial profile to the previous one (balance €105k) but is at the opposite extreme: all members are active (IsActiveMember = 1.00).
- **Cluster 0 (Loyal Customers):** The "Integrated Low-Balance" Customers (13.48% Churn): This is the most loyal segment. Its profile is the opposite of the previous two: a very low average balance (€10k) but possessing the highest average number of products (2.15).
- **Cluster 1 (Medium Risk):** The "Standard No-Card" Customers (19.94% Churn): This group is close to the average (20.37%). Its only true distinctive feature is not possessing a credit card (HasCrCard = 0.00).

5. Advanced Analysis and Comparative Validation

The K-Means analysis provided a robust and strategically useful segmentation. However, a good analyst does not trust the first result.

To ensure that the identified segments (like the "Ghost Customers") are not an artifact of the chosen method and to discover patterns that K-Means (which tends to create spherical clusters) might miss, this comparative analysis was conducted. The objective is to verify if K-Means was truly the best algorithm for this problem by comparing it with 3 other algorithms from different families.

5.1 Methodologies in Comparison: Theoretical Foundations

- **K-Means (Baseline):** Centroid-based partitional algorithm. Fast and efficient. It works iteratively to assign each point to the nearest centroid, then recalculates the centroid's position as the mean of the assigned points.
- **Hierarchical Clustering:** Hierarchical (bottom-up) algorithm. It builds a tree (Dendrogram) of merges. It starts with each point as a cluster and merges the two "closest" clusters (according to 'Ward' linkage) at each step.
- **DBSCAN:** Density-based algorithm. It does not require K, but rather eps (radius) and min_samples (minimum points) parameters. It is excellent at identifying outliers (noise) and arbitrary shapes.
- **BIRCH:** Hybrid (hierarchical + partitional) algorithm. Designed to be extremely fast and scalable, it builds a data summary (CF-Tree) before clustering.

5.2 Comparative Analysis Results

5.2.1 Hierarchical Clustering

This algorithm builds a hierarchy of merges, visualized in the dendrogram. Running the algorithm with K=4 (for a direct comparison) yields the following results:

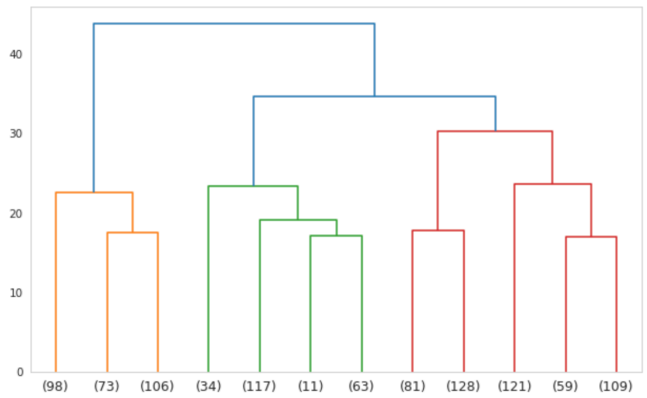

Figure 5.1: Dendrogram (Ward Method, on 1000 samples).

- **Silhouette Score:** 0.1110 (The highest metric score among the tested models).
- **Churn Analysis:** It produced a non-useful segmentation. The highest churn rate was 24.87% and the lowest was 18.31%.
- **Evaluation: Strategic failure.** Although Hierarchical Clustering achieved the best Silhouette Score, it was unable to produce a strategically relevant segmentation. The "risk gap" (6.56 points) is too small.

5.2.2 DBSCAN

This algorithm requires optimizing the eps parameter. The K-distance graph (with min_samples=18) suggests an "elbow" (a point of inflection) around eps=3.0.

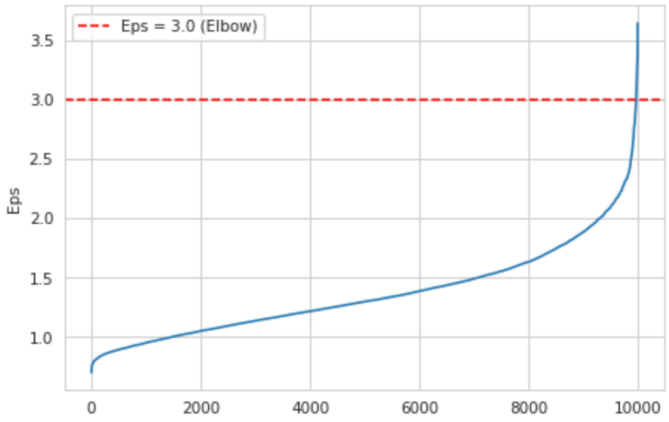

Figure 5.2: K-distance graph for ‘eps’ optimization.

- **Silhouette Score:** Not calculable.
- **Churn Analysis:** It produced a total clustering failure. It found only one cluster (Cluster 0) containing all 10,000 customers, with a churn rate of 20.37% (the exact average).
- **Evaluation: A fundamental insight.** This is not an error, but a discovery: it means that in the feature space, the "at-risk" and "loyal" customers are so intermingled that a density-based approach is unable to separate them.

5.2.3 BIRCH

- **Silhouette Score:** 0.1084 (Similar to K-Means).
- **Churn Analysis:** It produced a non-useful segmentation, similar to Hierarchical. The highest churn rate was 21.64% and the lowest was 19.07%.
- **Evaluation: Strategic failure.** Like Hierarchical, BIRCH (known for its speed) failed to isolate any risk segment, grouping almost all customers around the average.

5.3 Strategic Comparison Table and Conclusions

We summarize the results in a final comparison table, where we evaluate the algorithms not only for their metric "goodness" (Silhouette) but for their **strategic utility** (the "Risk Gap" they can create, i.e., the difference between the highest-churn and lowest-churn cluster).

While Silhouette and Davies-Bouldin scores measure the geometric quality of the clusters (how compact and separated they are in space), they do not guarantee business utility. To bridge this gap, we introduced the Risk Gap, a custom metric calculated as the range between the maximum and minimum churn rates.

This metric acts as the true proxy for discriminative power:

- A **High Risk Gap** implies the algorithm has successfully isolated "dangerous" customers (high churn) from "safe" ones (low churn), enabling targeted intervention.
- A **Low Risk Gap** indicates that the algorithm has mixed different risk profiles together (leveling the average), rendering the segmentation strategically inert.

Metric	K-Means (Baseline)	Hierarchical	DBSCAN	BIRCH
Silhouette Score	0.1039	0.1110	NaN	0.1084
Risk Cluster	32.07%	24.87%	20.37%	21.64%
Loyal Cluster	13.48%	18.31%	20.37%	19.07%
Risk Gap (Max-Min)	**18.59 points**	6.56 points	0 points	2.57 points
Strategic Valuation	**Success**	Failure	Failure	Failure

Table 5.1: Strategic Comparison Table of the 4 Algorithms.

Final Conclusion of the Comparative Analysis:

The analysis led to a fundamental conclusion: the algorithm with the apparently best metrics is not always the most strategically useful.

- **K-Means Wins:** Although Hierarchical Clustering achieved a marginally higher Silhouette Score, it proved strategically useless, failing to separate the risk segments (a gap of only 6.5 points). K-Means was the only algorithm to produce an actionable segmentation for the business, with a "Risk Gap" of almost 19 points.
- **Failure of Density and Hierarchy:** The total failure of DBSCAN and the poor performance of Hierarchical and BIRCH are not an error, but a scientific result. They demonstrate that, for this dataset, the risk profiles are not separable either hierarchically or by density.

In conclusion, this comparative analysis immensely strengthened the project: it validated K=4 objectively and demonstrated through direct comparison that the K-Means algorithm, chosen initially, was not just a "basic" choice, but the only strategically correct choice among those tested.

6. Conclusions and Strategic Actions

This final chapter synthesizes the key results emerging from the analysis, discusses the resulting managerial implications, and outlines the project's limitations, proposing future development paths.

6.1 Summary and Discussion of Results

The project's primary objective was fully achieved. Through a rigorous preprocessing pipeline and a **K validation** (Chapter 3), K=4 was identified as the optimal number of clusters. This choice was not subjective but was objectively confirmed by the convergence of three distinct metrics: the **Elbow Method (Inertia), the Silhouette Score, and the Davies-Bouldin Index (DBI).**

The application of the *baseline* algorithm, **K-Means** (Chapter 4), successfully segmented the clientele into homogeneous groups, isolating statistically distinct risk profiles. The most significant result was the identification of **Cluster 3, the "Ghost Customers"**: a segment representing almost a quarter of the clientele (24.29%) and showing a critical churn rate of **32.07%**, a full 12 points above the dataset average.

The advanced comparative analysis (Chapter 5) then scientifically validated the choice of K-Means. By comparing the *baseline* with the other three algorithms tested—**Hierarchical Clustering**, **DBSCAN**, and **BIRCH**—a fundamental insight emerged. Although **Hierarchical Clustering** had a marginally better *Silhouette* metric, it proved strategically useless (minimal risk gap). The total failure of **DBSCAN** and the poor performance of **BIRCH** in separating risk segments are not an error, but a scientific result.

This demonstrates that, for this dataset, the risk profiles are not separable by density or hybrid approaches, and that **K-Means** was the only one of the four methodologies tested to provide a strategically correct and actionable business solution.

6.2 Managerial Implications and Recommendations

The K-Means profiles identified in Chapter 4 are not just descriptive, but prescriptive. They allow for the targeted and efficient allocation of retention resources:

1. **Priority Action (Cluster 3: "Ghost" - Critical Risk):**
   - **Problem:** High balance, but total inactivity (IsActiveMember = 0.00).
   - **Action:** Proactive *re-engagement*. These customers should not be disturbed with generic offers. They require a *relationship manager* to offer value-added services (e.g., investment consulting, wealth management) to reactivate the relationship and justify the high balance they hold.

2. **Loyalty Action (Cluster 0: "Integrated" - Loyal Customers):**
   - **Profile:** Low balance, but a high number of products (NumOfProducts = 2.15).
   - **Action:** Strengthen *loyalty*. These customers are the bank's most solid base. The strategy is to reward them with tangible benefits (e.g., fee discounts, favorable rates) to consolidate their integration into the bank's ecosystem.

3. **Commercial Action (Cluster 1: "Standard" - Medium Risk):**
   - **Profile:** Only distinct feature: HasCrCard = 0.00.
   - **Action:** Clear *cross-selling* opportunity. This cluster represents an ideal customer pool for marketing campaigns targeted at credit card adoption and other basic financial services.

4. **Retention Action (Cluster 2: "Active" - Low Risk):**
   - **Profile:** High balance and high activity (IsActiveMember = 1.00).
   - **Action:** Maintenance and priority service. These are the "ideal" customers. The strategy is to guarantee excellent customer service to maintain their high satisfaction level and protect them from competitors.

6.3 Limitations and Future Developments

Although the analysis provided clear insights, it is important to acknowledge its limitations. The main limitation is the dataset simplification: the exclusion of the Geography variable, while justified to avoid "muddling" the clusters, prevents the capture of potential behavioral differences between markets (e.g., Germany vs. France).

This limitation opens the door for the most logical future development: the transition from a *descriptive* analysis (which snapshots the current state) to a *predictive* one. Having validated K-Means and its segments, the next step is to use cluster membership (or the Exited variable itself) as a *target* for a supervised classification model (e.g., Random Forest, XGBoost). The goal would be to calculate the churn probability for each individual customer, allowing the bank to act with personalized interventions *before* the customer makes the final decision to leave.

7. Bibliography

Methodologies and Algorithms

- MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297.
- Davies, D. L., & Bouldin, D. W. (1979). "A Cluster Separation Measure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 224-227.
- Rousseeuw, P. J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, 226–231.
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). "BIRCH: An Efficient Data Clustering Method for Very Large Databases." *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, 103–114.

Application Context

- Oyeniyi, J. M., & Adeyemo, A. B. (2015). "Customer Churn Analysis in Banking Sector Using K-Means Clustering Algorithm." *International Journal of Scientific & Engineering Research*, 6(8), 123-128.
- Avon, V. (2017). *Analisi di dati bancari con tecniche di Machine Learning*. Master's Thesis, University of Padua.

Software, Libraries, and Datasets

- Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825-2830.
- Hunter, J. D. (2007). "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, 9(3), 90-95.
- Waskom, M. L. (2021). "seaborn: statistical data visualization." *Journal of Open Source Software*, 6(60), 3021.
- Pandas Development Team (2020). *pandas-dev/pandas: Pandas*. Zenodo.
- Kaggle. *Bank Customer Churn Prediction (Churn_Modelling.csv)*. Available online.