Università degli studi di Milano - Bicocca

Department of Informatics, System and Communication (DISCo)

Master's Degree in Data Science

Integrated Supervised and Unsupervised Learning for Multivariate Socio-Economic Data Analysis

Data Science Lab Project

Davide Fabio Loreti - 865309

June 14, 2025

Contents

Introduction **2**

1 Exploratory Data Analysis (EDA) **3**
1.1 Overview of the Dataset 3
1.2 Gender Distribution 4
1.3 Age Distribution 5
1.4 Geographic Area Distribution 6
1.5 Education Level Distribution 7
1.6 Income Bracket Distribution 8
1.7 EDA Summary 8

2 Machine Learning Algorithms **9**
2.1 Data Preparation 9
2.2 Construction and evaluation of supervised models 11
2.3 Best Model Evaluation 12
2.4 Clustering Analysis 13
2.5 Clustering Evaluation 15
2.6 Second Clustering Attempt with Two Clusters 16
2.7 Clustering with Algorithms for Non-Spherical or Elliptical Structures 18
2.8 Visualization and Interpretation of Results 18
2.9 Final Remarks 21

3 Final considerations **22**

Introduction

This report focuses on the analysis of data collected by the Bank of Italy regarding the financial literacy of the adult Italian population. The main objective is to understand the demographic and socio-economic characteristics of the participants and to develop predictive models capable of classifying individuals' levels of financial literacy.

In the initial phase, an exploratory data analysis (EDA) was conducted to examine the distribution of key variables such as gender, age, geographic area, education level, and income bracket. This exploration provided valuable insights into the sample characteristics and highlighted potential challenges and patterns useful for subsequent modeling.

For the modeling phase, classical Machine Learning techniques were chosen due to the dataset's size and structured tabular nature, where traditional algorithms often prove more effective and interpretable than deep neural networks. Algorithms considered include Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost, selected for their robustness and ability to handle heterogeneous data.

The goal is to develop classification models capable of predicting individuals' financial literacy levels, thus enabling effective profiling and supporting targeted interventions in financial education. Model evaluation will be performed using standard metrics such as accuracy, precision, recall, and F1-score, alongside cross-validation techniques to ensure prediction generalizability.

This combination of exploratory analysis and predictive modeling aims to provide useful and reliable insights to better understand the dynamics of financial literacy and to support effective policy-making strategies.

1 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase represents a crucial preliminary step in any data science project, serving to uncover the underlying structure, detect patterns, and identify potential anomalies or data quality issues within the dataset. Before applying any advanced modeling techniques, such as machine learning or deep learning algorithms, it is essential to thoroughly understand the data at hand to ensure the validity and reliability of subsequent analyses.

In this project, our focus is on a comprehensive dataset related to financial literacy among the adult Italian population. The dataset is enriched with various demographic and socio-economic variables that provide a multifaceted view of the individuals surveyed. By carefully examining these features, we aim to gain insights into the distribution and relationships between variables, which will inform the selection and tuning of predictive models later on.

1.1 Overview of the Dataset

The dataset comprises several key attributes describing individual respondents, including:

- **Gender**, capturing the biological sex of the participant;
- **Age**, representing the respondent's age in years;
- **Geographic Area**, indicating the regional location within Italy;
- **Education Level**, detailing the highest attained educational qualification;
- **Income Bracket**, classifying the monthly income range of the individual.

Each of these variables is subject to both visual and statistical exploration. Visualizations such as histograms, bar charts, and kernel density estimates (KDE) allow for an intuitive grasp of the data distribution, while summary statistics provide quantitative measures such as means, medians, and counts. This dual approach facilitates the detection of any irregularities, including missing values, outliers, or unexpected patterns, which could influence model performance if left unaddressed. Through this careful examination, we set the foundation for a robust and meaningful analysis.

1.2 Gender Distribution

Descriptive statistics:

- Unique categories: 2 (Female, Male)
- Most frequent category: `Female` (with 2446 observations against 2416)
- Missing values: 0

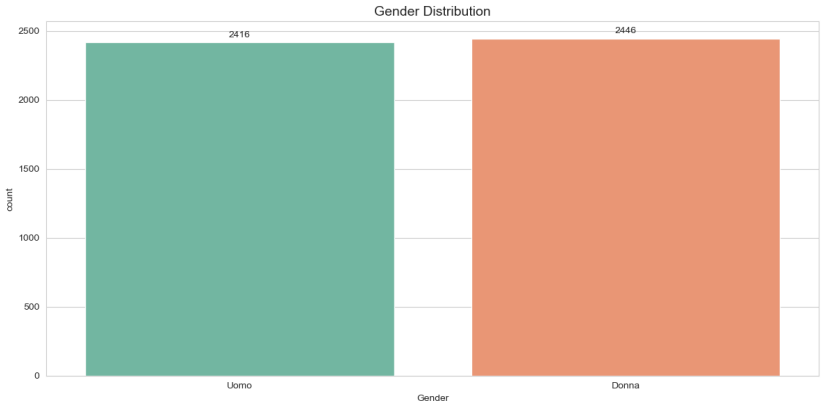

Figure 1: Gender distribution in the dataset.

As shown in the figure, the population is fairly balanced in terms of gender, with a slight predominance of females. This balance makes the variable suitable for future comparative analysis.

1.3 Age Distribution

Descriptive statistics:

- Mean age: 50.31 years
- Median: 50.00
- Minimum and maximum: `18 - 79`
- Missing values: 219

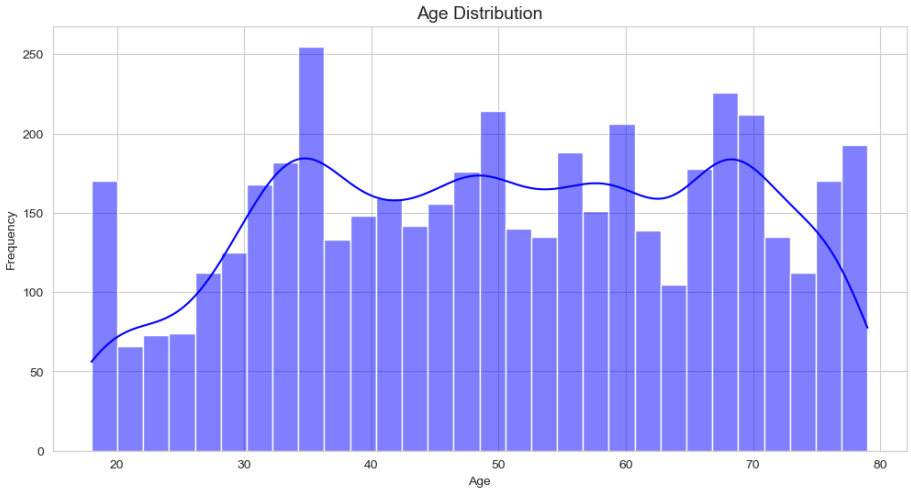

Figure 2: Age distribution with histogram and KDE curve.

The age distribution appears symmetric and centered around 50 years. The presence of 219 missing values requires proper handling (e.g., imputation or removal) before proceeding to modeling.

1.4 Geographic Area Distribution

Descriptive statistics:

- Unique categories: 5
- Most represented area: `Nord - Ovest`
- Missing values: 0

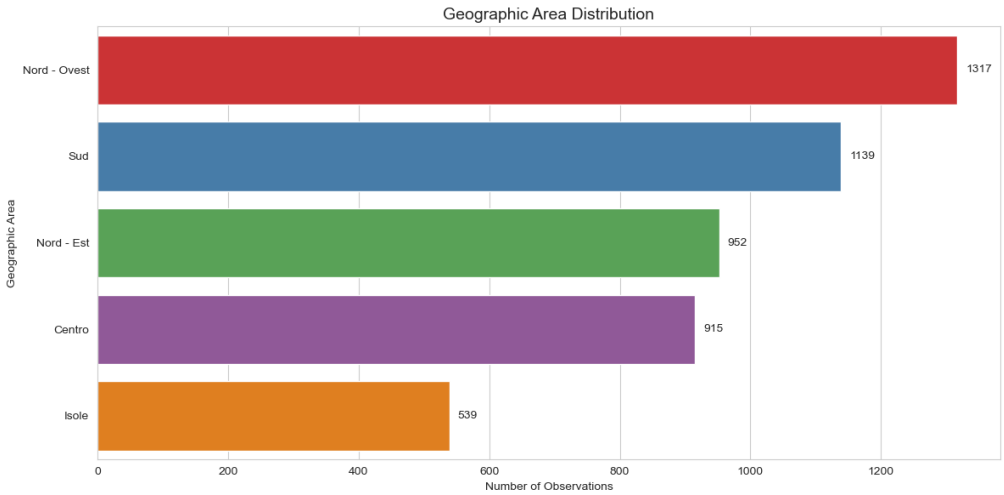

Figure 3: Distribution by geographic area.

The North-West area is the most represented in the sample. This concentration may reflect higher participation or population density in that region.

1.5 Education Level Distribution

Descriptive statistics:

- Unique categories: 10
- Most frequent category: `Scuola media superiore con diploma`
- Missing values: 0

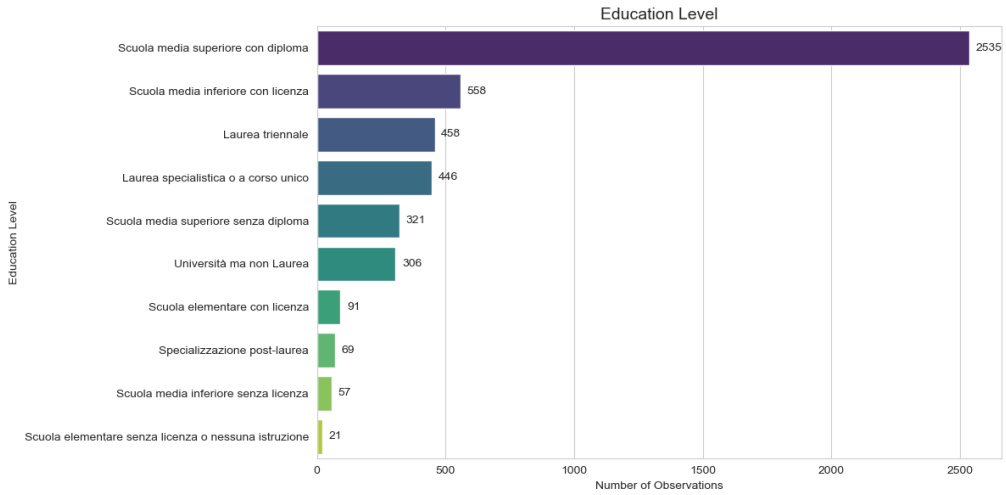

Figure 4: Education level distribution.

The most common education level is a high school diploma, followed by various other degrees. The large variety of categories makes this variable useful for segmentation and classification models.

1.6 Income Bracket Distribution

Descriptive statistics:

- Unique categories: 5
- Most frequent category: `tra 1.751 Euro e 2.900 Euro al mese`
- Missing values: 0

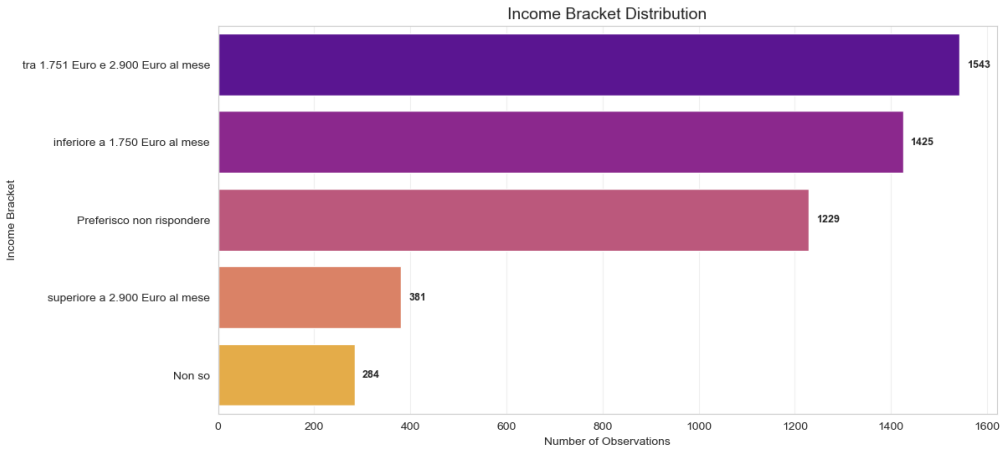

Figure 5: Monthly income bracket distribution.

The central income bracket is the most represented, suggesting a concentration in the middle class. All categories are well distributed and no missing values are present.

1.7 EDA Summary

The exploratory data analysis reveals a dataset in generally good condition, except for missing values in the `Age` variable. The distributions are consistent with expectations for the Italian population and show a balanced representation across key demographic dimensions. These findings support the use of the selected variables in the subsequent modeling phases using Machine Learning algorithms.

2 Machine Learning Algorithms

The previous chapter presented a thorough exploratory data analysis (EDA), which was essential to understanding the composition and key characteristics of the dataset. Through the examination of demographic and socio-economic variables, such as gender, age, geographic area, education level, and income bracket, we identified meaningful patterns and detected some data quality issues, including missing values in the age variable. These insights laid a solid foundation for building reliable and effective predictive models.

This chapter focuses on applying machine learning techniques to deepen the analysis and leverage the information contained in the dataset. Specifically, we adopted a dual approach: unsupervised clustering methods to uncover natural groupings within the data without relying on labels, and supervised classification models aimed at predicting the target variable, gender.

For classification, several well-established algorithms were employed, including Random Forest, Support Vector Machines, K-Nearest Neighbors, and XGBoost, chosen for their robustness and ability to handle structured, heterogeneous data. Preprocessing steps such as feature scaling and feature selection were implemented to enhance model performance and generalization. Model evaluation relied on standard metrics like accuracy, precision, recall, and F1-score, alongside cross-validation techniques to ensure reliable performance estimates.

The integration of clustering and classification methods allowed us to both explore the underlying data structure and assess predictive capabilities, ultimately supporting more precise profiling and targeted interventions in the context of financial literacy.

The following sections detail the methodologies, implementation choices, and results obtained, providing a comprehensive view of the machine learning workflow applied in this study.

2.1 Data Preparation

The initial data preparation phase involved careful handling of missing values and encoding of categorical variables, essential steps to make the dataset compatible with machine learning techniques. Specifically, the `Age` column was converted to a numeric format, coercing any non-numeric entries into missing values (`NaN`). Then, numeric and categorical columns were distinguished.

Missing values in numeric columns were imputed using the median, a choice robust against potential outliers. For categorical columns, missing values were replaced with the constant string `Unknown`, preserving the information about missingness without discarding entire rows.

Regarding the encoding of categorical variables, a selective approach was applied: only columns with fewer than 50 unique categories were encoded using `LabelEncoder`. This transformation converts textual categories into integer values, enabling the variables to be used by machine learning models.

In total, 216 categorical columns were encoded, including relevant variables such as:

- `Gender` with 2 unique categories,
- `Geographic_Area` with 5 categories,
- `Work_Sector` with 10 categories,
- `Income_Bracket` with 5 categories.

After these operations, the resulting dataset consists of 4862 observations and 219 columns.

Next, the classification target was explicitly defined as the variable `Gender`. All numeric features except the target itself were considered available for training, resulting in a total of 217 numeric features. The first ten features included:

- `Age`, `qd7_a`, `Geographic_Area`, `Province`, `qd5_1`, `qd5_2`, `qd5_3`, `qd5_4`, `qd5_5`, `qd5_6`.

The presence of valid targets was verified for both classification tasks (represented by `Gender`) and regression (with the variable `Age`).

Given the sufficient number of features and the availability of the classification target, the gender prediction task was set up. Rows with missing target values were removed, yielding a final dataset of 4862 samples. Class distribution was also analyzed to ensure balanced representation.

Feature selection was not applied at this stage. Although feature selection can help reduce dimensionality and improve generalization, in our case all available features were kept for the following reasons:

- The number of numeric features (217) was not excessive for the models used, which included Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machines (SVM), Naive Bayes, and K-Nearest Neighbors.
- Retaining all variables ensures that no potentially useful information is prematurely excluded.
- Some machine learning models, such as tree-based ones, can naturally handle the presence of irrelevant or less informative features.
- Keeping all features contributes to the robustness and accuracy of the models by allowing them to exploit the full information content of the data.
- The goal was to evaluate the model's performance on the full feature space before considering any dimensionality reduction techniques.

For training and evaluation, the dataset was split into a training set (80%) and a test set (20%), stratifying to preserve the distribution of the `Gender` variable. The training set contains 3889 samples, while the test set contains 973 samples, both with 217 features. To optimize performance, data were standardized by scaling to zero mean and unit variance, with the scaler fitted on the training set and then applied to the test set.

2.2 Construction and evaluation of supervised models

If the matrix `X_selected_class`, containing the features selected for classification, we proceed to train and evaluate multiple classification models. The models considered are:

- Random Forest
- Logistic Regression
- Gradient Boosting
- Support Vector Machine (SVM)
- Naive Bayes
- K-Nearest Neighbors (KNN)

For each model, training is performed either on the scaled or unscaled training data depending on the model type (tree-based models are trained on unscaled data, others on scaled data). After training, a 5-fold cross-validation is conducted on the training set to assess the stability and robustness of the model's performance. The final evaluation metric is the accuracy computed on the held-out test set.

The table below summarizes the accuracy and cross-validation results for each model:

```
Random Forest            : Accuracy = 0.7544, CV = 0.7457 ± 0.0370
Logistic Regression      : Accuracy = 0.6639, CV = 0.6457 ± 0.0389
Gradient Boosting        : Accuracy = 0.9476, CV = 0.9182 ± 0.0199
SVM                      : Accuracy = 0.6763, CV = 0.6698 ± 0.0426
Naive Bayes              : Accuracy = 0.5920, CV = 0.5793 ± 0.0547
K-Nearest Neighbors      : Accuracy = 0.6166, CV = 0.6153 ± 0.0330
```

The best performing model according to accuracy on the test set is the Gradient Boosting classifier with an accuracy of 94.76%.

2.3 Best Model Evaluation

Using the Gradient Boosting classifier, we generate predictions on the test set and compute a detailed classification report. The results show excellent performance across all metrics with precision, recall, and F1-score all approximately 0.95 for both classes.

Moreover, feature importance analysis reveals that variables such as `wght`, `Age`, and `Work_Sector` play a dominant role in the model's decisions.

		Predicted	
		0	1
Actual	0	464	26
	1	25	458

Figure 6: Confusion Matrix of the Gradient Boosting model on the test set.

The confusion matrix shows that out of the total test samples, 464 true negatives and 458 true positives were correctly classified. The model misclassified 26 samples as positive when they were actually negative (false positives), and 25 samples as negative when they were actually positive (false negatives). This indicates a balanced and precise classification performance.

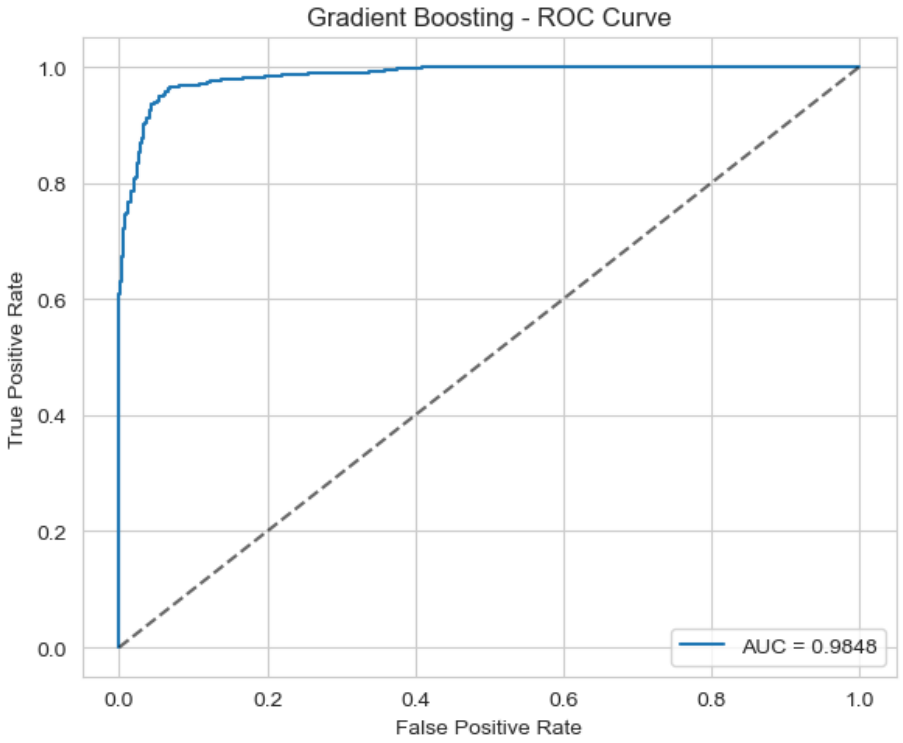

Figure 7: ROC Curve of the Gradient Boosting model on the test set. The Area Under the Curve (AUC) is 0.9848, indicating excellent discriminatory ability between the two classes.

The ROC curve further confirms the model's strong performance with an AUC of 0.9848, suggesting near-perfect separation of the classes. This high AUC complements the accuracy and precision metrics, reinforcing the effectiveness of Gradient Boosting in this classification task.

In summary, the Gradient Boosting classifier not only achieves the highest accuracy but also demonstrates robust predictive power and interpretability through feature importance and evaluation metrics, making it the preferred choice for the gender prediction task.

2.4 Clustering Analysis

The clustering analysis began by selecting all the numerical variables available in the dataset, resulting in a total of 218 features. To reduce complexity and noise, only the first ten most significant numerical features were used: `Gender`, `Age`, `qd7_a`, `Geographic_Area`, `Province`, `qd5_1`, `qd5_2`, `qd5_3`, `qd5_4`, and

qd5_5. After removing samples with missing values in these features, the final dataset for clustering comprised 4862 observations.

K-means clustering was applied with the number of clusters set to three (n_clusters=3). Prior to clustering, the features were scaled using the StandardScaler to ensure all variables contributed equally, avoiding bias due to different scales.

To facilitate visualization and interpretation, dimensionality reduction was performed using Principal Component Analysis (PCA), retaining the first two principal components. These two components explain approximately 19.0% of the total variance, enabling a two-dimensional representation that aids in understanding the cluster structure.

The K-means algorithm produced the following cluster distribution: Cluster 0 contains 1794 samples (36.9%), Cluster 1 is the largest with 2541 samples (52.3%), and Cluster 2 is the smallest with 527 samples (10.8%).

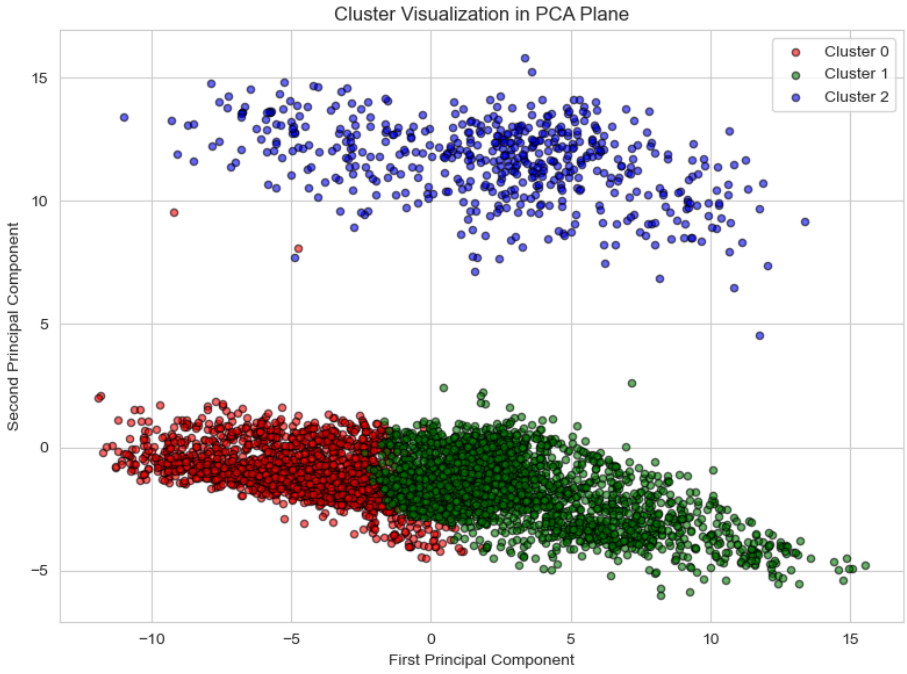

Figure 8: Clusters visualized in the two-dimensional PCA space defined by the first two principal components.

The scatter plot of the clusters in the PCA space (Figure 8) reveals a clear separation among the groups. Notably, Cluster 2 forms a distinct cloud positioned higher along the second principal component axis, with values approximately between 10 and 15, highlighting its unique profile relative to the rest of the data.

Clusters 0 and 1 are denser and located closer together in the lower region of the plot, with values between 0 and -5 on the second principal component axis. These two clusters are mainly separated along the first principal component, with Cluster 0 predominantly on the left and Cluster 1 on the right, showing a slight downward slope from left to right. This spatial arrangement suggests that while these two clusters share some similarities, they differ significantly on other characteristics, justifying their separation.

Overall, the clustering analysis successfully segmented the dataset into three distinct profiles, uncovering patterns of variability in numerical features that warrant further investigation to better understand group differences.

2.5 Clustering Evaluation

The quality of the clustering obtained with the K-means algorithm set to 3 clusters was assessed using the silhouette score, which measures how well each sample fits within its assigned cluster compared to other clusters.

The average silhouette score is 0.0788, a rather low value indicating poor internal cohesion within clusters and limited separation between groups. Silhouette scores close to zero suggest that many samples lie near the boundary between clusters, while negative or very low scores imply that some samples may be misclassified.

The silhouette plot (see Figure 9) further reveals that some components are likely misclassified. Specifically, cluster 2 shows a portion of silhouette values shifted towards the left (negative values), suggesting these samples might be better assigned to cluster 1. Similarly, cluster 1 has some values skewed towards the left, indicating a potential reassignment to cluster 0.

This overlap and unclear cluster boundary confirms the difficulty of the algorithm in clearly separating distinct groups, highlighting the need to reconsider the number of clusters, feature selection, or clustering approach.

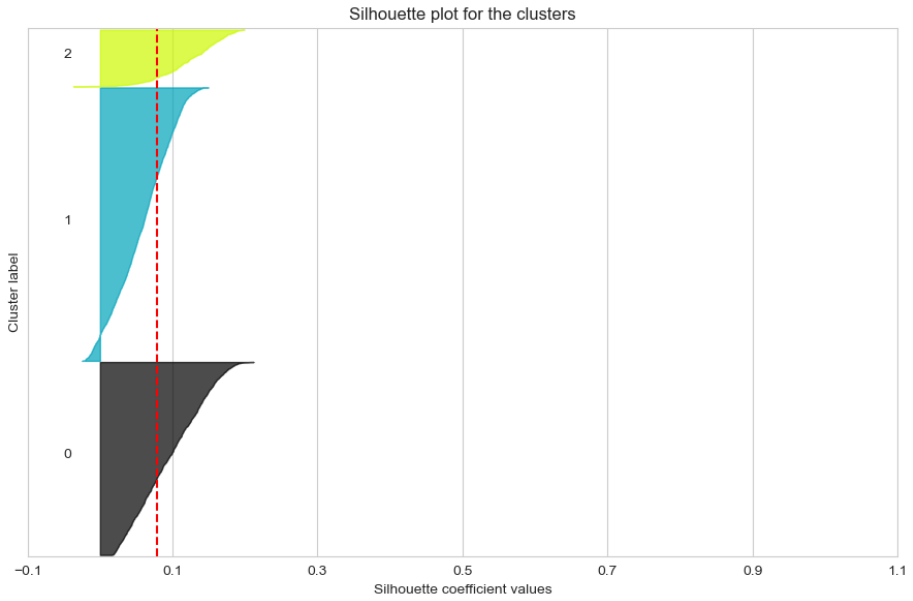

Figure 9: Silhouette plot of the clustering results for 3 clusters.

2.6 Second Clustering Attempt with Two Clusters

To further investigate the cluster structure and improve clustering quality, the K-means algorithm was rerun with the number of clusters set to two (`n_clusters=2`). The same scaled features and PCA transformation were used for consistency.

The two-cluster solution produced a slightly higher average silhouette score of 0.19, indicating an improved, yet still moderate, cluster separation compared to the three-cluster case.

Figure 10 shows the scatter plot of the two clusters in the PCA space, highlighting a broader grouping compared to the previous three-cluster solution.

The silhouette plot for the two-cluster solution (Figure 11) confirms the slightly better cluster compactness and separation, although the score remains relatively low, suggesting overlapping clusters or insufficient feature discrimination.

In conclusion, while the two-cluster approach marginally improves the clustering evaluation metric, the overall low silhouette scores suggest that the dataset's structure may be inherently difficult to segment into well-separated groups using K-means with the selected features.

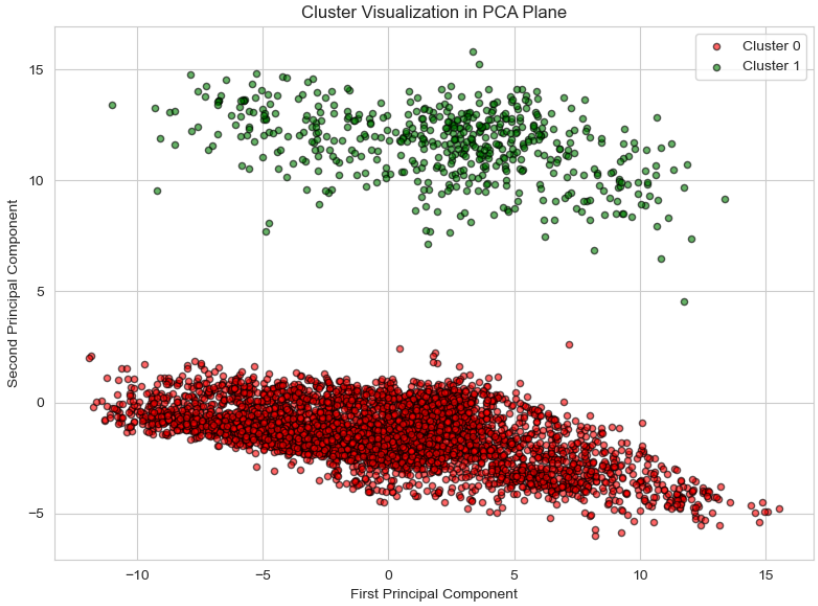

Figure 10: Clusters visualized in the two-dimensional PCA space for the two-cluster solution.

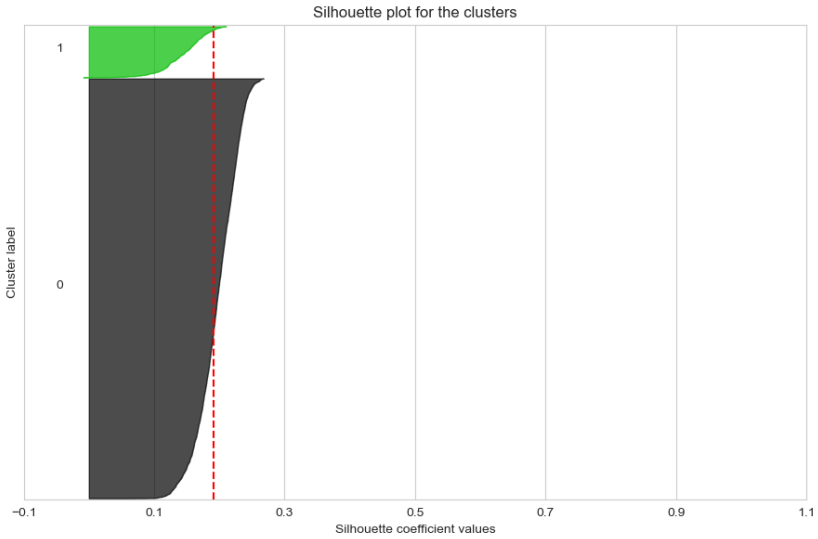

Figure 11: Silhouette plot of the clustering results for 2 clusters.

2.7 Clustering with Algorithms for Non-Spherical or Elliptical Structures

Given the limitations observed with K-Means, an alternative strategy was adopted based on clustering algorithms capable of identifying non-spherical, non-linear, or density-based structures. Specifically, the following methods were explored:

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: this algorithm is based on point density and can identify arbitrarily shaped clusters, also distinguishing noise as outliers. It proved effective in detecting dense and well-separated groups, with a silhouette score of 0.262.
- **Gaussian Mixture Model (GMM)**: a probabilistic model assuming a Gaussian distribution for each cluster. Unlike K-Means, GMM allows for elliptical clusters and returns a probability of cluster membership. In this case, it yielded a silhouette score of 0.330.
- **Spectral Clustering**: an algorithm that leverages graph spectral theory to detect even non-convex clusters. It provided the best performance in terms of silhouette score, with a value of 0.452, indicating a good degree of cluster separation.

2.8 Visualization and Interpretation of Results

To better understand the nature of the resulting clusters and to compare the performance of the different algorithms, several informative visualizations were produced, as described below.

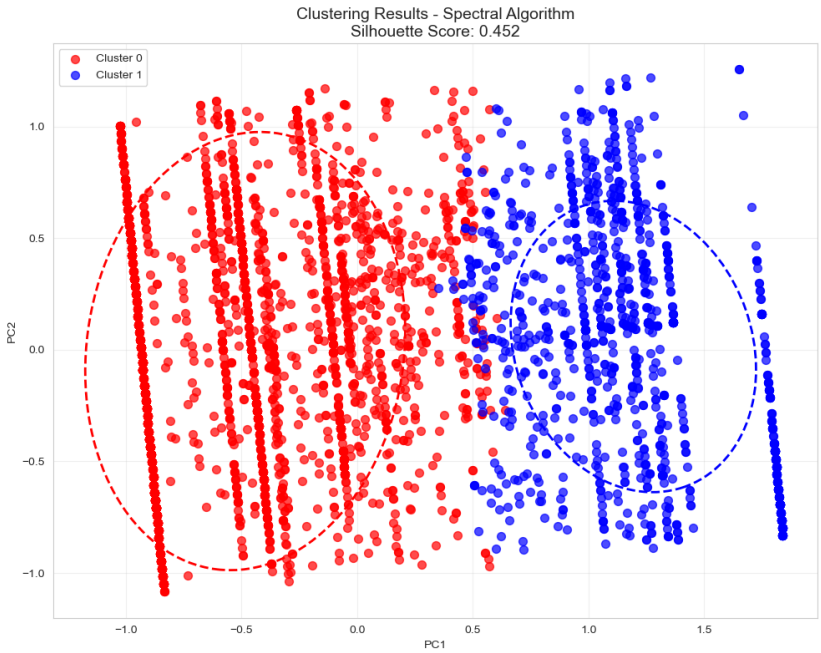

Figure 12: Distribution of data in clusters 0 and 1 according to Spectral Clustering, with a silhouette score of 0.452.

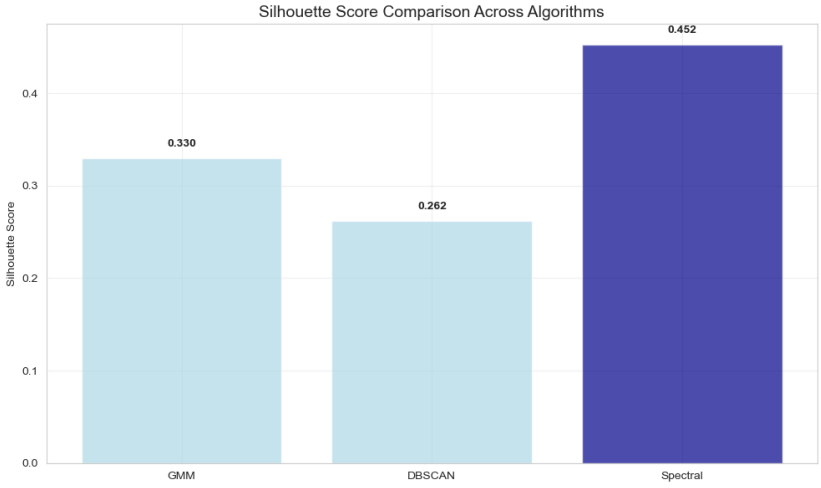

Figure 13: Comparison of silhouette scores for the three algorithm

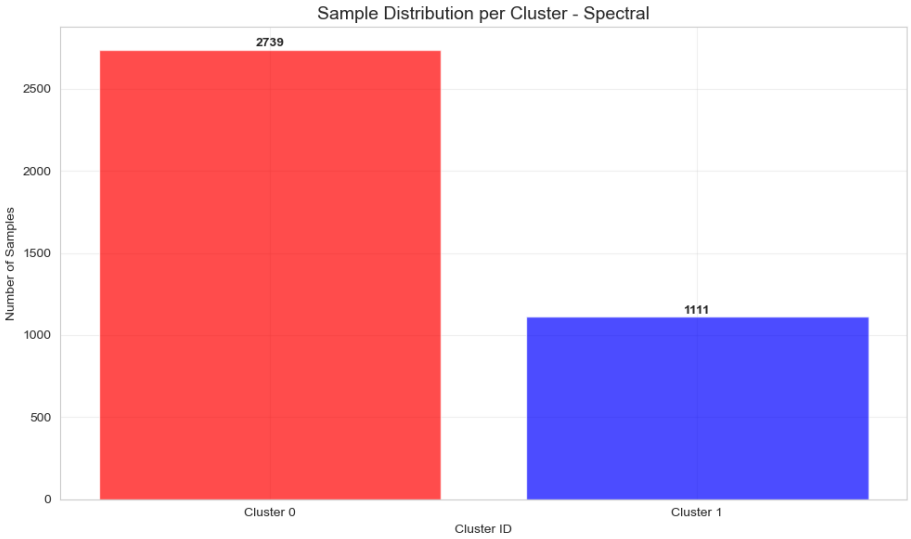

Figure 14: Sample distribution by cluster: cluster 0 contains 2739 observations, and cluster 1 contains 1111 observations.

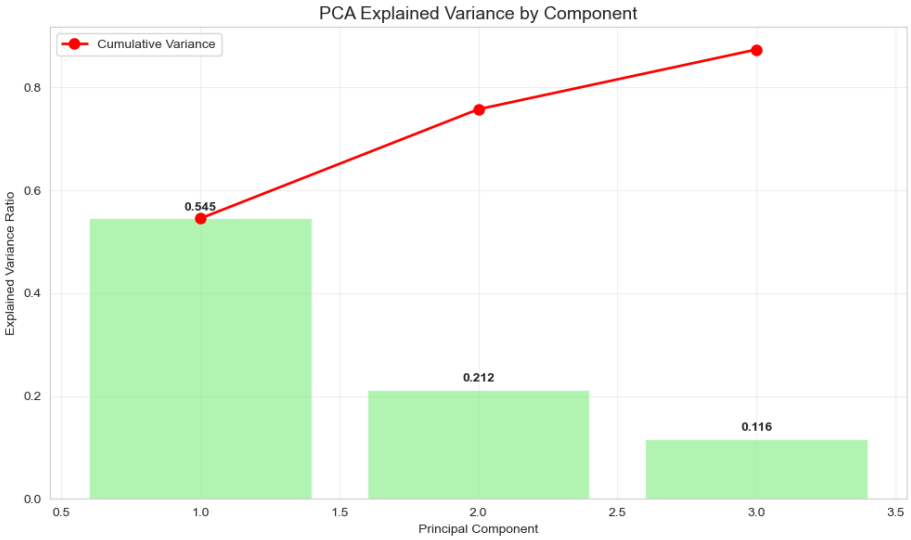

Figure 15: Explained variance per principal component. The first component explains 54.5% of the variance, the second 21.2%, and the third 11.6%. The total explained variance by the first three components is 87.3%.

2.9 Final Remarks

This analysis highlighted how a multi-model approach, combined with visualization techniques and quantitative evaluation methods such as the silhouette score, is essential for achieving effective clustering on complex data. The failure of K-Means, documented both by performance metrics and visual inspection, led to an informed decision to apply more sophisticated alternatives. Among these, Spectral Clustering yielded the most reliable results in the context of financial literacy segmentation.

3 Final considerations

The analysis conducted in this project allowed us to explore the dataset from multiple perspectives, combining supervised classification models, performance evaluation, and unsupervised clustering techniques. Overall, the results provided useful insights into the structure and relationships within the data.

Regarding supervised learning, models such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) demonstrated good predictive capabilities. In particular, the Random Forest model proved effective in handling a large number of features, including potentially redundant ones, offering a balance between accuracy and interpretability. The SVM also performed well, especially in cases where the data appeared linearly separable. However, the performance of some algorithms was sensitive to feature scaling and the presence of non-informative variables, confirming the importance of careful preprocessing.

The clustering analysis using K-Means returned less satisfactory results. The average silhouette score of 0.0788 indicates poor internal cohesion and weak separation between clusters. From the silhouette plot, it became evident that some samples may have been misclassified. For instance, part of cluster 2 displays negative silhouette values, suggesting that those points may belong more appropriately to cluster 1. Similarly, some points in cluster 1 appear closer to cluster 0. This suggests that the data structure may not be well captured by a rigid partitioning like K-Means, possibly due to non-spherical distributions or the presence of outliers.

The main advantages of the models used include their accessibility, widespread implementation, and effectiveness on medium-sized datasets. However, some limitations emerged: more complex models tend to be less interpretable, and their performance heavily depends on appropriate feature selection and transformation.

Future improvements may include:

- Applying automatic feature selection techniques to reduce dimensionality and improve model generalization;
- Experimenting with alternative clustering algorithms such as DBSCAN or Gaussian Mixture Models, which may better handle irregular shapes and density variations;
- Employing more robust cross-validation strategies to reliably compare model performance;
- Assessing feature importance in predictive models to gain further interpretability;
- Extending the analysis with Deep Learning methods, if justified by dataset size and complexity.

In conclusion, the work carried out provides a solid foundation for understanding the dataset and evaluating different Machine Learning approaches. It also highlights the need for further refinement to obtain more robust and meaningful models.