ISSN 2723-6471

Applied Density-Based Clustering Techniques for Classifying High-Risk Customers: A Case Study of Commercial Banks in Vietnam

Nguyen Minh Nhat[1,*]

[1]*Faculty of Banking, Ho Chi Minh University of Banking (HUB), Ho Chi Minh City, Vietnam*

(Received: July 19, 2024; Revised: July 29, 2024; Accepted: August 12, 2024; Available online: October 15, 2024)

Abstract

Understanding and effectively engaging with customers is paramount in today's rapidly evolving business landscape. With rapid technological advances, banks have unprecedented opportunities to improve their approach to customer segmentation. This change is driven by integrating resource planning systems and digital tools, enabling a more comprehensive and data-driven understanding of customer behavior. Therefore, identifying high-risk customers is critical for banks aiming to mitigate credit risk. This study evaluates the performance of both density-based and non-density-based clustering algorithms in classifying customers at risk of default. Specifically, it assesses K-Means (a non-density-based method) alongside three density-based algorithms: DBSCAN, HDBSCAN, and Birch. Each algorithm brings a unique approach to handling complex customer data. The study uses a dataset of 77,272 customers from Vietnamese commercial banks (2010–2022) and evaluates these algorithms using seven key performance metrics: Davies-Bouldin Index, Silhouette Score, Adjusted Rand Index, Homogeneity, Completeness, V-Measure, and Accuracy. These metrics were selected to assess various aspects of clustering quality, including cluster compactness, separation, and alignment with true data distributions. By focusing on both internal (cluster quality) and external (alignment with ground truth) metrics, the study ensures a comprehensive evaluation of the algorithms' performance in addressing the goal of identifying high-risk customers. The study's key findings are: (1) DBSCAN and HDBSCAN excel in identifying high-risk clusters, even in noisy data environments, though they face challenges in cluster separation. (2) Birch offers strong cluster compactness and separation but requires optimization for accuracy. (3) K-Means struggles to capture the complexity of customer behavior in this context, limiting its effectiveness for credit risk classification. The contributions of this research demonstrate the value of using density-based clustering methods in credit risk management frameworks. The findings suggest that these techniques significantly enhance the accuracy of identifying high-risk customers, offering actionable insights for banks to reduce financial risk and improve operational efficiency. The study also contributes to the broader understanding of clustering techniques by thoroughly evaluating both density-based and non-density-based algorithms, offering valuable guidance for practitioners in the banking sector.

Keywords: Credit Risk Management, Clustering Algorithms, Customer Risk Assessment, Irregular Clusters, Density-Based Clustering

1. Introduction

In the context of an ever-fluctuating global economy, financial institutions and banks encounter significant challenges in managing credit risk and predicting personal bankruptcy (PB) among individual customers. Accurate PB prediction is vital as it enables these institutions to estimate appropriate interest rates, set lending conditions, and effectively manage investment portfolios. This meticulous process not only minimizes potential losses but also enhances operational efficiency. As data technology and computer science continue to advance, employing sophisticated methods for analyzing and forecasting PB has become increasingly critical. These cutting-edge techniques offer enhanced precision, allowing financial institutions to navigate the complexities of credit risk management with greater efficacy and resilience [1], [2].

In the dynamic realm of credit risk management, machine learning approaches, particularly clustering algorithms, have revolutionized the prediction of defaults by unveiling intricate data patterns that traditional methods often miss. K-Means Clustering shines for its straightforwardness and efficiency, ideal when the number of clusters is predefined, and data is uniformly distributed. However, it demands precise tuning for optimal results [3]. In contrast, DBSCAN excels in identifying clusters of arbitrary shapes and effectively managing noise, though it requires careful parameter calibration to avoid inconsistencies [4]. HDBSCAN builds on this by offering a hierarchical structure, adept at handling clusters of varying densities, thereby enhancing robustness and reducing noise sensitivity. Lastly, Birch, tailored for large-scale datasets, employs an incremental approach, making it perfect for hierarchical data frameworks but

***Corresponding author: Nguyen Minh Nhat (nhatnm@hub.edu.vn)**
DOI: https://doi.org/10.47738/jads.v5i4.344

somewhat less effective with high-dimensional data [5]. These clustering techniques, each with unique strengths, are pivotal in enhancing predictive accuracy and operational efficiency in credit risk management.

However, new research has sparked intense debate over the efficacy of different clustering models in predicting personal loan defaults. Discussions often center around the relative strengths of algorithms like K-Means, DBSCAN, HDBSCAN, and Birch. Some studies suggest that HDBSCAN is superior due to its ability to handle clusters of varying densities more effectively and its hierarchical clustering structure, which helps in identifying complex cluster formations. HDBSCAN reduces the noise sensitivity issue present in DBSCAN and offers better performance for datasets with complex structures [6], [7]. On the other hand, other research highlights the advantages of Birch in handling large datasets through incremental clustering, which allows it to efficiently manage massive data. Birch is praised for its scalability and ability to handle data that fits well into a hierarchical framework, making it particularly useful for large-scale applications [8]. However, Birch may not perform as well with high-dimensional data compared to other methods [9]. Comparative studies reveal these nuances, with some researchers advocating for the use of HDBSCAN in scenarios with complex data distributions and varying densities, while others emphasize Birch's efficiency and scalability for large datasets [10]. K-Means is often recommended for its simplicity and efficiency in partitioning data into distinct clusters when the number of clusters is known a priori, but it struggles with clusters of varying densities and irregular shapes. DBSCAN excels in identifying clusters of arbitrary shapes and effectively managing noise, but its sensitivity to parameter settings can lead to inconsistent results if not properly tuned [11]. While the study evaluates multiple clustering techniques - K-Means, DBSCAN, HDBSCAN, and Birch - it is essential to discuss the specific conditions under which each method performs best. Different clustering algorithms have varying strengths, and understanding the context in which each technique is most effective can help practitioners select the right approach for their needs.

Therefore, this study undertakes a comprehensive comparative analysis of clustering algorithms, including K-Means, DBSCAN, HDBSCAN, and Birch, to identify the optimal predictive model and the key factors influencing personal default risk. The performance of these models is rigorously evaluated using seven critical metrics: Davies-Bouldin Index, Silhouette Score, Adjusted Rand Index, Homogeneity, Completeness, V-Measure, and Accuracy. Utilizing a dataset comprising 77,272 individual customers from Vietnamese commercial banks spanning from 2010 to 2022, this evaluation aims to provide financial institutions with profound insights into the efficacy of various clustering algorithms [12]. The anticipated outcomes will empower these institutions to make well-informed decisions and implement effective strategies for enhancing their personal default risk forecasting and management processes. To mitigate these weaknesses, Birch can be combined with other methods better suited for handling noise or non-hierarchical data, ensuring that it benefits from its efficiency while compensating for its limitations in more complex scenarios.

The study paper's format contains a theoretical introduction with a thorough analysis of boosting algorithms and their applications in credit risk assessment. The research methodology section details the data sources, preprocessing stages, feature selection approaches, and boosting algorithms employed in the study. The results and analysis section includes the empirical data, which compares the performance of various boosting models. Finally, the conclusion summarizes the study's important findings, emphasizing the benefits of improving algorithms in personal default prediction and recommending future research topics.

2. Literature Review

The rapid development in the field of credit risk management has garnered significant attention from both the research community and practitioners. Numerous studies have compared advanced technologies with traditional statistical techniques, emphasizing the importance of integrating modern credit assessment tools into practice [13], [14]. These findings provide evidence for the need to update credit evaluation methods in the banking and finance industry. However, there is a specific gap in the research focusing on the comparative effectiveness of clustering methods in accurately predicting personal bankruptcy (PB).

Clustering techniques have been widely applied to identify and predict clusters of personal bankruptcy cases. Methods such as K-Means Clustering, DBSCAN, HDBSCAN, and Birch have shown promise in this regard. For instance, K-

Means Clustering to profile personal bankruptcy members, identifying five distinct clusters based on demographic, financial, debt, and social stigma indicators [15]. This study demonstrated the utility of clustering in understanding the characteristics of individuals at risk of bankruptcy.

K-Means Clustering is a partition-based, non-hierarchical method that has been effective in producing good clustering results for various practical applications [16]. It involves selecting initial seeds and evaluating each item by its distance to the nearest seed, iteratively updating the centroids until convergence [17], [18]. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters based on density, allowing for the discovery of arbitrarily shaped clusters and the identification of noise points [19]. HDBSCAN extends DBSCAN by creating a hierarchy of clusters and selecting the optimal flat clustering based on stability. Birch (Balanced Iterative Reducing and Clustering using Hierarchies) handles large datasets efficiently by building a tree structure (CF tree) incrementally and clustering the leaves. This method is particularly effective for large-scale data due to its ability to perform clustering incrementally and dynamically [20].

Despite the growing application of clustering techniques in personal bankruptcy prediction, there is still a need for comprehensive comparative studies that evaluate the accuracy and effectiveness of different clustering methods. This research aims to fill this gap by systematically comparing K-Means Clustering, DBSCAN, HDBSCAN, and Birch in predicting personal bankruptcy clusters, focusing on their ability to identify high-risk individuals accurately [21]. This study is grounded in the theory of cluster analysis, which is used to identify groups of entities that share certain common characteristics and to understand behaviors by identifying homogeneous groups. Clustering techniques are critical in credit risk management as they help in segmenting the borrower population based on risk characteristics, thereby enabling more targeted and effective risk mitigation strategies [22]. K-means clustering has been widely applied in various contexts. Using K-Means to profile personal bankruptcy members, finding that K-Means was effective in identifying distinct clusters of individuals based on demographic and financial indicators.

The study revealed that clusters identified through K-Means could help in understanding the characteristics of high-risk individuals, thereby aiding in risk management. DBSCAN is known for its ability to identify clusters of varying shapes and sizes, making it suitable for complex datasets [23]. HDBSCAN improves upon DBSCAN by providing a more robust clustering approach that accounts for hierarchical structures within the data. These techniques have been applied to various domains, including credit risk management, to identify high-risk clusters. Birch is designed for large datasets and has been applied in various contexts to efficiently identify clusters [24]. Its ability to handle incremental data makes it suitable for applications in dynamic environments such as credit risk management. Profiling distressed borrowers using clustering techniques provides valuable insights for credit organizations in identifying defaulters and individuals at high risk of personal bankruptcy. The profiles generated from clustering analysis help in tailoring risk management strategies to specific segments, thereby improving the overall effectiveness of credit risk management.

3. Research Methodology

3.1. Research Design

This research aims to develop and evaluate clustering models for predicting personal default, leveraging advanced machine learning techniques. The process is structured into distinct phases: data preprocessing, encoding, model training, evaluation, and deployment. Each stage is methodically designed to assess the effectiveness of these models comprehensively. To ensure the developed clustering models are robust and accurate, the process is structured into distinct phases: data preprocessing, encoding, model training, evaluation, and deployment, as shown in figure 1. This figure illustrates the clustering techniques training process for predicting personal defaults, providing a visual representation of each phase in the workflow.

The initial stage centers on thorough data preprocessing to ensure the models receive high-quality inputs. The dataset, containing personal financial details indicative of financial distress, is meticulously reviewed to understand its structure, features, and potential issues. To maintain the uniqueness of each record and prevent data redundancy, duplicate entries are identified and removed. Missing values are addressed using appropriate imputation methods, such as replacing them with the mean, the most frequent value, or the median. This step is crucial to ensure that the dataset is complete and ready for analysis [25].

In the encoding phase, features are scaled to normalize the range of values, ensuring that all features contribute equally to the distance measurements used in clustering. Principal Component Analysis (PCA) is applied to reduce the dimensionality of the dataset, retaining the most significant components. This step helps simplify the dataset while preserving its variance, making the clustering process more efficient and manageable.

During the model training phase, a variety of clustering models are set up and trained on the encoded dataset. K-Means is initialized to partition the data into a predefined number of clusters, aiming to minimize the variance within each cluster. DBSCAN is used to identify clusters of varying shapes and sizes, particularly effective for datasets with noise and outliers. HDBSCAN (Hierarchical DBSCAN), an extension of DBSCAN, is employed to handle clusters of varying densities more effectively. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is utilized for large-scale clustering, particularly useful for its efficiency in handling large datasets incrementally [25].

Once the training phase is completed, the models are evaluated using a separate test dataset to assess their predictive ability. Performance metrics such as the Silhouette Score, Davies-Bouldin Index, and Adjusted Rand Index are used to evaluate the quality of the clusters formed. These metrics provide insights into the cohesion within clusters, the separation between clusters, and the alignment with the true data labels. By comparing these metrics across different models, we can identify which clustering method or combination of methods delivers the most accurate and reliable results. To further refine prediction accuracy, an ensemble model is developed by merging the outputs of the individual clustering models. This ensemble method utilizes techniques such as majority voting or weighted averaging to combine predictions, effectively harnessing the strengths of each model. The ensemble model is then rigorously tested and validated, ensuring it consistently outperforms the individual models.

Finally, the trained and evaluated models are saved and prepared for deployment. The models are serialized and stored in a format that facilitates easy reloading and usage in future predictions, including saving the model architecture, weights, and any preprocessing steps required [26]. These models are then deployed in a real-world environment where they can predict personal default based on new data, with the necessary infrastructure set up to handle model inference requests efficiently. This structured approach ensures that the developed clustering models are robust, accurate, and ready for practical use in predicting personal default in figure 1.

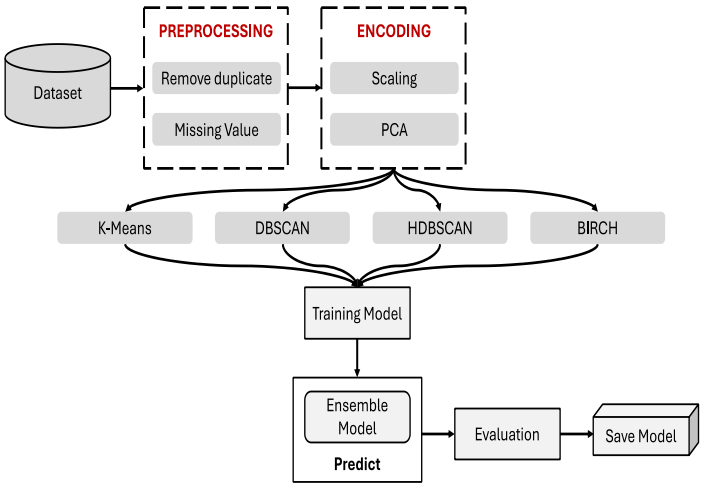

Figure 1. Clustering techniques training process for predicting personal default

3.2. Clustering Techniques

Clustering is an unsupervised learning approach in machine learning that organizes data points into groups, where each group contains points that are more alike to one another than to those in other groups. This technique is especially valuable for exploratory data analysis as it helps uncover the natural structure within the data without the need for predefined labels. In the context of predicting personal bankruptcy, clustering helps identify patterns and profiles of individuals who are at risk, thus aiding financial institutions in targeted risk management strategies [27].

K-Means Clustering is a method used to partition data into k distinct clusters in a non-hierarchical manner. Each cluster is represented by a centroid, which is the average of the data points in that cluster. The algorithm operates through

 ISSN 2723-6471

several steps: initialization, assignment, update, and iteration. Initially, k centroids are either selected randomly or determined using a specific heuristic. Each data point is then assigned to the nearest centroid based on Euclidean distance. The centroids are subsequently updated by calculating the mean of the data points assigned to them. This process is repeated until the centroids stabilize and no longer change significantly. The goal of K-Means is to minimize the variance within each cluster, as expressed by the following formula:

$$J = \sum_{i=1}^{k} \sum_{x \epsilon C_i} \|x - \mu_i\|^2 \quad (1)$$

Where J is the objective function, k denotes the number of clusters, C_i is the collection of points in cluster i, x is a data point, and μ_i is the centroid of cluster i. This technique works well for clusters that are spherical in shape and is computationally efficient, making it well-suited for handling large datasets. Nonetheless, it necessitates specifying the number of clusters (k) in advance and is sensitive to the initial placement of centroids [27].

DBSCAN is a clustering algorithm that operates on the principle of density. It identifies clusters as regions of high data point density, separated by regions of low density. The primary parameters for this algorithm are epsilon (ϵ), which is the maximum distance between two points for them to be considered neighbors, and MinPts, the minimum number of points required to form a dense region (core point). The process involves identifying core points that have at least MinPts neighbors within the epsilon radius. These core points are then connected to form clusters, while points that do not connect to any core point are considered noise. DBSCAN is particularly useful for identifying clusters of arbitrary shapes and is robust against outliers. However, it requires careful adjustment of the epsilon and MinPts parameters to perform effectively. The formal definition of a cluster in DBSCAN relies on the concepts of density reachability and density connectivity, which are outlined by the following criteria:

A point p is directly density-reachable from a point q if p is within distance ϵ from q and q has at least MinPts neighbors within ϵ.

A point p is density-reachable from a point q if there is a chain of points $p1, p2, \ldots, pn$ where $p_1 = q$ and $p_n = p$ such that $p_i + 1$ is directly density-reachable from p_i.

A point p is considered density-connected to a point q if there exists a point o such that both p and q are density-reachable from o.

HDBSCAN extends DBSCAN by introducing a hierarchical clustering approach. It constructs a hierarchy of clusters and uses stability measures to extract the most stable clusters. The steps involved are hierarchy construction, condensation, and stability analysis. A minimum spanning tree of the data points is built with edge weights representing the distance between points. The tree is then condensed by removing edges with weights greater than a threshold. The stability of clusters at different levels of the hierarchy is analyzed to identify the most stable clusters. HDBSCAN does not require the specification of eps and MinPts parameters but instead determines the appropriate clustering structure from the data itself, making it more adaptive to different data distributions. The stability of a cluster is quantified using the concept of excess mass and the lifetime of a cluster, providing a robust measure for cluster selection.

Birch is designed for large datasets and performs clustering incrementally. It builds a tree structure called the Clustering Feature (CF) tree, which summarizes the data distribution. The process involves CF tree construction, where data points are inserted into the CF tree, and each node represents a cluster. Optionally, a global clustering algorithm is applied to refine the clusters obtained from the CF tree. Birch is particularly effective for large-scale data due to its ability to perform clustering incrementally and dynamically. It is also efficient in terms of memory and computational requirements [27]. The CF tree uses Clustering Feature vectors, which are triplets $CF = (N, LS, SS)$ where N is the number of points, LS is the linear sum of the points, and SS is the squared sum of the points. These vectors are used to efficiently summarize and merge clusters. This research follows a structured process comprising data preprocessing, encoding, model training, evaluation, and deployment. Each stage ensures the models built are accurate and suitable for predicting personal defaults. Missing values were encountered in various columns of the dataset. To address this issue, we applied different imputation methods based on the nature of the missing data: Numerical variables were imputed using the mean or median value, depending on the distribution (mean for normally distributed data and median

for skewed distributions). Categorical variables were imputed using the most frequent value to maintain consistency in the dataset. Outliers in the dataset, particularly in financial variables like loan amounts and credit history, were detected using Z-scores and IQR (Interquartile Range) methods. Data points with Z-scores greater than 3 or those falling outside 1.5 times the IQR were flagged as outliers. The following strategies were employed: Winsorization was used to limit extreme values by capping them at a predefined percentile. In cases where outliers were deemed critical for the model, they were retained but normalized during feature scaling to prevent skewing the results.

3.3. Criteria to Measure the Model's Ability to Predict Default Risk

To evaluate the effectiveness of each clustering method, we compare the clusters generated by each algorithm against the original target labels. These metrics provide different perspectives on the quality of the clusters formed by each algorithm and are essential for comparing their performance accurately. The metrics used include the Davies-Bouldin Index, Silhouette Score, Adjusted Rand Index, Homogeneity, Completeness, V-Measure, and Accuracy. By comparing these metrics, we aim to identify the most accurate clustering method for predicting personal bankruptcy clusters. Below is a detailed description of these indices:

The Davies-Bouldin Index (DBI) is an internal metric used to evaluate clustering quality by measuring the average similarity between each cluster and the cluster most similar to it. This similarity is calculated as the ratio of within-cluster distances to between-cluster distances. A lower DBI value indicates superior clustering performance, as it reflects well-separated and compact clusters. The DBI is computed using the following formula:

$$\mathrm{DBI} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \max_{\mathrm{j}\neq\mathrm{i}}\left(\frac{\sigma_\mathrm{i}+\sigma_\mathrm{j}}{\mathrm{d_{ij}}}\right) \tag{2}$$

Where k is the number of clusters, σ_i is the average distance of all points in cluster i to the centroid of cluster i, and d_{ij} is the distance between the centroids of clusters i and j. The Davies-Bouldin Index provides a measure of both cluster compactness (how tightly the points in a cluster are grouped together) and separation (how distinct or separate a cluster is from other clusters). It penalizes the clustering solution if clusters are too close to each other or if the clusters themselves are dispersed [27].

The Silhouette Score is an internal evaluation metric that assesses the similarity of an object to its own cluster relative to other clusters. The score ranges between -1 and 1, with higher values indicating that data points are well-aligned with their own cluster and poorly aligned with neighboring clusters. The Silhouette Score for an individual data point is calculated as follows:

$$\mathrm{s(i)} = \frac{\mathrm{b(i)}-\mathrm{a(i)}}{\max\{\mathrm{a(i)},\mathrm{b(i)}\}} \tag{3}$$

In this context, $a(i)$ represents the average distance from data point i to all other points within the same cluster, while $b(i)$ denotes the smallest average distance from data point i to points in any other cluster. The overall Silhouette Score is derived by averaging the silhouette scores of all individual data points. This metric effectively captures both the cohesion within clusters and the separation between different clusters, providing a thorough assessment of clustering quality [28].

The Adjusted Rand Index (ARI) is an external evaluation metric used to assess the similarity between clustering outcomes and the true labels. Unlike the Rand Index, the ARI corrects for the possibility of random grouping. The ARI score ranges from -1 to 1, with 1 signifying perfect correspondence between the clustering and the actual labels, 0 representing random labeling, and negative values indicating worse than random agreement. The ARI is calculated using the following formula:

$$\mathrm{ARI} = \frac{\mathrm{RI}-\mathrm{E[RI]}}{\max(\mathrm{RI})-\mathrm{E[RI]}} \tag{4}$$

Where RI is the Rand Index and $E[RI]$ is the expected Rand Index of random clustering. The Rand Index measures the percentage of decisions that are correct, but ARI adjusts this by considering the expected similarity of all pairwise

combinations under random labeling. This makes ARI a more reliable measure for comparing clustering results with ground truth labels, especially when the number of clusters or cluster sizes vary significantly [28].

Homogeneity is a metric used to evaluate whether each cluster consists exclusively of members from a single class. Clustering results are considered homogeneous if every cluster contains only data points from one class. The homogeneity score ranges from 0 to 1, with 1 indicating perfectly homogeneous clusters. The formula to calculate homogeneity is:

$$h = 1 - \frac{H(C/K)}{H(C)} \tag{5}$$

Where $H(C/K)$ represents the conditional entropy of the class distribution given the cluster assignments, and H(C) denotes the entropy of the class distribution. This metric ensures that the clusters do not mix different classes, thus providing a measure of purity for each cluster.

(v) Completeness: Completeness is a metric that evaluates whether all data points of a particular class are grouped within the same cluster. A clustering outcome achieves completeness if every data point belonging to a specific class is assigned to a single cluster. This metric ranges from 0 to 1, with 1 indicating clusters that are perfectly complete. The formula for calculating completeness is:

$$c = 1 - \frac{H(K/C)}{H(K)} \tag{6}$$

Where $H(K/C)$ is the conditional entropy of the cluster distribution given the class assignments, and H(K) is the entropy of the cluster distribution. This metric ensures that all data points of the same class are grouped together, providing a measure of coverage for each class within the clusters.

(vi) V-Measure: V-Measure is the harmonic mean of homogeneity and completeness, providing a balanced measure of the two. It is defined as:

$$v = 2 \times \frac{h \times c}{h + c} \tag{7}$$

Where h is homogeneity and c is completeness. A high V-Measure indicates that the clustering algorithm produces results that are both homogeneous and complete. This combined metric ensures that clusters are both pure and complete, providing a comprehensive evaluation of clustering quality [28].

Accuracy is a simple and intuitive metric that measures the proportion of correctly classified instances. In the context of clustering, it compares the cluster labels assigned by the algorithm to the ground truth labels, treating the clustering problem as a classification problem. Accuracy ranges from 0 to 1, where 1 indicates perfect accuracy. The formula for accuracy is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Where TP represents the number of true positives, TN denotes the number of true negatives, FP indicates the number of false positives, and FN stands for the number of false negatives. This metric offers a clear assessment of how effectively the clustering algorithm aligns with the actual labels.

4. Result and Discussion

4.1. Research results

The dataset employed in this study to predict personal defaults encompasses data from 77.272 customers of various commercial banks in Vietnam, spanning the years 2010 to 2022. To ensure strict adherence to data protection regulations, all customer information was encrypted during the data collection process. This comprehensive dataset includes 14 explanatory variables along with one target variable, thoroughly detailed in table 1.

Table 1. Dataset variables

	Variables	Description
Explanatory variables	X1	Indicates the current status of the loan
	X2	The total amount of money currently owed on the loan
	X3	The duration of the loan
	X4	The borrower's annual income
	X5	The number of years the borrower has been in their current job
	X6	Indicates the home ownership status of the borrower
	X7	The purpose of the loan
	X8	The total amount of monthly debt payments
	X9	The number of years the borrower has had credit
	X10	The number of months since the borrower was last delinquent on a payment
	X11	The total number of open credit accounts
	X12	The number of recorded credit problems
	X13	The current balance across all credit accounts
	X14	The highest amount of credit ever extended to the borrower
Target variable	Bankruptcies	Current status of whether the applicant is bankrupt or non-bankrupt

Table 1 showed the target variable in this study is Bankruptcy, classified into five levels: 0 (non-bankruptcy) and 1 to 4 indicating increasing levels of financial distress. This classification provides a comprehensive view of the severity of financial difficulties faced by borrowers, essential for developing precise predictive models and risk assessment strategies. The Bankruptcy variable categorizes historical occurrences from 0 (no bankruptcy) to 4 (severe financial distress), with an average of 0.1 and a standard deviation of 0.3, indicating most consumers have not declared bankruptcy. Addtionally, the statistical descriptions of explanatory variables in the dataset are presented in detail in table 2.

Table 2. Description of data statistics

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	Bankruptcies
count	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272	77272
mean	0.2	13.3	0.3	14.0	5.9	0.9	1.6	9.6	17.9	48.1	11.2	0.2	12.2	13.1	0.1
std	0.4	2.17	0.5	0.5	3.6	0.9	1.9	0.7	6.8	19.6	5.0	0.5	1.1	1.1	0.3
min	0.0	9.6	0.0	11.2	0.5	0.0	0.0	2.0	3.7	0.0	1.0	0.0	2.9	8.4	0.0
25%	0.0	12.2	0.0	13.7	3.0	0.0	1.0	9.3	13.4	34.0	8.0	0.0	11.7	12.5	0.0
50%	0.0	12.7	0.0	13.9	6.0	1.0	1.0	9.7	16.8	60.0	10.0	0.0	12.3	13.1	0.0
75%	0.0	13.2	1.0	14.3	10.0	2.0	1.0	10.1	21.5	60.0	14.0	0.0	12.8	13.6	0.0
max	1.0	18.4	1.0	18.9	10.0	3.0	15.0	12.9	70.5	176.0	76.0	15.0	17.3	21.2	1.0

Table 2 showed that the mean of 0.2 and a maximum value of 1.0 indicate that the majority of loans are likely in a particular status (e.g., non-default), with some in another status (e.g., default). The standard deviation of 0.4 suggests some variability. To analyze the statistical description of the dataset variables, we can break down the statistics for each variable provided in table 2. This will give us insights into the distribution, central tendency, and variability of the data. The dataset seems well-rounded with a good mix of explanatory variables. Most variables have a range that indicates diversity in borrower characteristics. The target variable, "Bankruptcies," shows a low mean, indicating few bankruptcies in the dataset.

4.2. Comparison Results on the Predictive Ability of the Models

The out-of-sample test results of various clustering methods, including K-Means, DBSCAN, HDBSCAN, Birch are presented in detail in table 3.

Table 3. Clustering performance metrics for predicting personal bankruptcy

	Metric	Davies-Bouldin Index	Silhouette Score	Adjusted Rand Index	Homogeneity	Completeness	V-Measure	Accuracy
1	K-Means	0.925	0.32	0.017	0.138	0.032	0.052	0.325
2	DBSCAN	2.459	0.4	-0.003	0.001	0.008	0.002	0.885
3	HDBSCAN	4.429	0.57	-0.001	0.001	0.004	0.001	0.884
4	Birch	0.73	0.298	0.034	0.069	0.025	0.037	0.506

Table 3 showed that K-Means Clustering is a popular method due to its simplicity and efficiency. The results show a Davies-Bouldin Index of 0.925, indicating relatively good cluster separation and compactness. The Silhouette Score of 0.32 suggests moderate cohesion within clusters. The Adjusted Rand Index is 0.017, which is low, indicating that the clusters formed are not very similar to the ground truth labels. Homogeneity and Completeness scores are 0.138 and 0.032, respectively, reflecting that while the clusters are somewhat homogeneous, they are not very complete. The V-Measure, which combines homogeneity and completeness, is 0.052. Overall, K-Means achieved an Accuracy of 0.325, suggesting it performs reasonably well but is not the most effective method for this dataset [29].

DBSCAN, a density-based clustering method, performed differently. It achieved a higher Silhouette Score of 0.4, indicating better cohesion within clusters compared to K-Means. However, the Davies-Bouldin Index is 2.459, suggesting that the clusters are not as well separated. The Adjusted Rand Index is -0.003, which is slightly negative, indicating poor similarity with the ground truth. Both Homogeneity and Completeness are very low at 0.001 and 0.008, respectively, resulting in a V-Measure of 0.002. Despite these results, DBSCAN achieved the highest accuracy of 0.885, indicating that it effectively identified the clusters relevant to predicting bankruptcy, despite poor performance in other metrics [30].

HDBSCAN, an extension of DBSCAN, also performed well in certain aspects. It achieved the highest Silhouette Score of 0.57, suggesting very good cohesion within clusters. However, the Davies-Bouldin Index is 4.429, indicating that the clusters are not well separated. The Adjusted Rand Index is -0.001, similar to DBSCAN, indicating poor similarity with the ground truth. Homogeneity and Completeness scores are both very low at 0.001 and 0.004, respectively, resulting in a V-Measure of 0.001. Despite this, HDBSCAN achieved an Accuracy of 0.884, almost as high as DBSCAN, indicating it is also effective in identifying the relevant clusters for predicting bankruptcy.

Birch is particularly effective for large datasets due to its incremental and dynamic clustering capabilities. The results show a Davies-Bouldin Index of 0.73, the lowest among the methods, indicating very good cluster separation and compactness. The Silhouette Score is 0.298, slightly lower than K-Means but still reasonable. The Adjusted Rand Index is 0.034, the highest among the methods, suggesting better similarity with the ground truth. Homogeneity and Completeness scores are 0.069 and 0.025, respectively, resulting in a V-Measure of 0.037. Birch achieved an Accuracy of 0.506, indicating it performs better than K-Means but not as well as DBSCAN or HDBSCAN in this context.

Numerous studies have explored the effectiveness of various clustering methods in predicting personal bankruptcy, leading to diverse and often conflicting conclusions. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is frequently highlighted for its robustness in identifying high-risk clusters. Emphasizing DBSCAN's strength in managing noise and detecting clusters of arbitrary shapes, making it highly effective in practical applications. Despite some challenges in cluster separation and alignment with true data labels, DBSCAN remains a robust choice for accurately identifying high-risk groups [31].

Conversely, K-Means delivers balanced performance across various metrics, making it a versatile and practical choice for many scenarios. Its straightforward implementation and interpretability make K-Means particularly suitable for less complex datasets, ensuring reliable clustering results. Studies provide a comparative analysis, emphasizing that the choice between K-Means and DBSCAN often depends on specific application needs and dataset characteristics. They highlight that while DBSCAN is superior in handling noise and arbitrary shapes, K-Means is more efficient for large-scale data and predefined cluster numbers [30], [31].

HDBSCAN achieves high accuracy in identifying relevant clusters, which is particularly beneficial for financial applications. However, they also noted some challenges with cluster separation and alignment, similar to DBSCAN.

Commending Birch for its robust cluster separation and alignment with actual data structures. However, they also observed that Birch's overall accuracy might be lower compared to DBSCAN and HDBSCAN, indicating a need for further optimization in some cases.

Recent studies have expanded the analysis of clustering methods. The results demonstrated the effectiveness of convolutional neural networks (CNNs) in bankruptcy prediction, suggesting that combining CNNs with clustering methods like K-Means can enhance performance. Results reviewed various machine learning and deep learning techniques, highlighting that ensemble methods combining clustering with predictive models often yield superior results.

Additionally, a survey by showed that hybrid ensemble models, which integrate clustering techniques with other classifiers, improve accuracy in credit scoring and bankruptcy prediction. This reinforces the notion that the effectiveness of clustering methods can be significantly enhanced when combined with other machine learning approaches [31].

Comparing these clustering methods provides valuable insights into their effectiveness in predicting personal bankruptcy. DBSCAN and HDBSCAN achieved the highest accuracy, suggesting that density-based methods are highly effective in identifying high-risk clusters. However, their poor performance in homogeneity, completeness, and adjusted Rand Index indicates that while they can identify relevant clusters, these clusters may not align well with the actual distribution of the data [32]. K-Means, although simpler, provides a balanced performance across different metrics but falls short in accuracy compared to DBSCAN and HDBSCAN. Its moderate Silhouette Score and relatively low Davies-Bouldin Index indicate reasonable cluster quality, but the method struggles to match the true cluster distribution as indicated by the low Adjusted Rand Index. Birch stands out for its low Davies-Bouldin Index and highest Adjusted Rand Index, suggesting it forms well-separated and relevant clusters. However, its accuracy, while better than K-Means, is lower than that of DBSCAN and HDBSCAN. Birch's ability to handle large datasets efficiently makes it a viable option, but its performance metrics suggest it may require further tuning or combination with other methods for optimal results.

4.3. Results Discussions of Analyzing the Clustering

Scatter plots are a powerful visualization tool to understand how different clustering algorithms partition the data. By projecting high-dimensional data onto a two-dimensional plane using principal component analysis, we can visually inspect the effectiveness of clustering methods like K-Means, DBSCAN, HDBSCAN, and Birch. The scatter plots presented below illustrate how these algorithms cluster a dataset of individual customers, aiming to predict the likelihood of personal bankruptcy in figure 2.

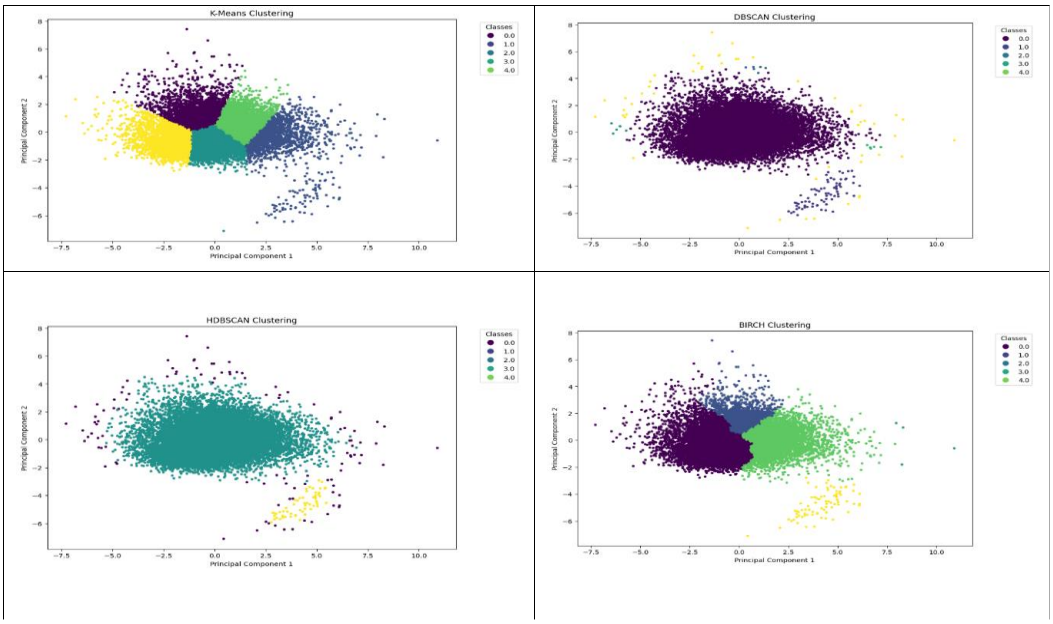

Figure 2. Estimation results of clustering

In the K-Means clustering scatter plot, the data points are divided into distinct, spherical clusters. The algorithm has partitioned the data into five clusters (classes 0 to 4), with each cluster color-coded differently. The clusters appear to be relatively well-separated, indicating that K-Means has identified distinct groups within the data. However, some overlap between clusters suggests that while K-Means is effective, it may struggle with complex boundaries. The centroids of each cluster are well-positioned, showing that the algorithm has effectively minimized within-cluster variance. This clustering method excels at creating clear and distinct boundaries when the data is relatively well-behaved and does not contain too many anomalies or irregular shapes. Despite this, the presence of some overlap indicates that the assumptions of K-Means may not hold perfectly, especially in the context of more complex financial data with nuanced patterns.

The DBSCAN scatter plot shows a different clustering pattern compared to K-Means. DBSCAN identifies clusters based on the density of data points, resulting in a more natural grouping. Here, we see a large, dense cluster (class 0) surrounded by smaller, sparser clusters (classes 1 to 4). DBSCAN is particularly effective at identifying noise and outliers, which are the scattered points not belonging to any major cluster. This method works well for identifying irregularly shaped clusters and handling noise, as evidenced by the presence of distinct small clusters and outliers. The large central cluster suggests that DBSCAN effectively groups dense regions, but the algorithm also identifies points that do not fit into these dense areas, marking them as noise. This approach is highly beneficial for datasets where the cluster shapes are not spherical and where outliers can significantly affect the analysis. However, the larger Davies-Bouldin Index indicates that the clusters are not as well-separated compared to those formed by K-Means.

HDBSCAN extends DBSCAN by introducing a hierarchical approach to clustering. The scatter plot for HDBSCAN shows a primary, dense cluster (class 2) with smaller, less dense clusters (classes 0, 1, 3, and 4) scattered around. The hierarchical nature of HDBSCAN allows it to identify both large and small clusters effectively. The presence of more distinct clusters and fewer outliers compared to DBSCAN indicates HDBSCAN's ability to capture a broader range of cluster densities, making it well-suited for datasets with hierarchical structures. The highest Silhouette Score among the methods suggests that HDBSCAN clusters are very cohesive. The hierarchical approach enables HDBSCAN to manage the data's complexity better, capturing fine details in cluster formation that DBSCAN might miss. However, the high Davies-Bouldin Index points to some challenges in achieving well-separated clusters, possibly due to the algorithm's sensitivity to the density parameters.

The Birch scatter plot shows a clustering pattern similar to K-Means, but with more nuanced cluster boundaries. Birch is designed to handle large datasets incrementally and efficiently, forming clusters dynamically. The plot shows well-separated clusters (classes 0 to 4), with clear boundaries and minimal overlap. The clusters appear more elongated compared to K-Means, reflecting Birch's flexibility in cluster shape. This method effectively balances the need for well-defined clusters and computational efficiency, making it suitable for large datasets with diverse structures. Birch's ability to form flexible, non-spherical clusters makes it advantageous for data with varying densities and distributions. The lowest Davies-Bouldin Index among the methods suggests that Birch forms the most compact and well-separated clusters. The higher Adjusted Rand Index also indicates that Birch clusters align more closely with the actual data distribution, providing a more accurate representation of the underlying structure.

Comparing these clustering methods provides valuable insights into their effectiveness in predicting personal bankruptcy. DBSCAN and HDBSCAN achieved the highest accuracy, suggesting that density-based methods are highly effective in identifying high-risk clusters. However, their poor performance in homogeneity, completeness, and Adjusted Rand Index indicates that while they can identify relevant clusters, these clusters may not align well with the actual distribution of the data. The high accuracy of DBSCAN and HDBSCAN demonstrates their capability in correctly classifying the majority of data points, but the low values in other metrics suggest that these clusters might not be as pure or complete as desired. This disparity indicates that while these methods are excellent at forming clusters, the resulting clusters might mix different classes or miss capturing all members of a class within a single cluster.

K-Means, although simpler, provides a balanced performance across different metrics but falls short in accuracy compared to DBSCAN and HDBSCAN. Its moderate Silhouette Score and relatively low Davies-Bouldin Index indicate reasonable cluster quality, but the method struggles to match the true cluster distribution as indicated by the low Adjusted Rand Index. K-Means is particularly effective when the data follows assumptions of spherical clusters

with similar variance. However, its limitations become apparent in more complex datasets, where the simplicity of the model can lead to suboptimal clustering results. Birch stands out for its low Davies-Bouldin Index and highest Adjusted Rand Index, suggesting it forms well-separated and relevant clusters [33]. However, its accuracy, while better than K-Means, is lower than that of DBSCAN and HDBSCAN. Birch's ability to handle large datasets efficiently makes it a viable option, but its performance metrics suggest it may require further tuning or combination with other methods for optimal results. The flexibility of Birch in forming clusters of various shapes and sizes makes it particularly useful in datasets with diverse structures, and its incremental approach ensures that it can scale well with larger datasets.

5. Conclusions and Recommendations

Accurately predicting personal bankruptcy is a critical challenge for financial institutions. This study aims to evaluate and compare the effectiveness of four clustering algorithms - K-Means Clustering, DBSCAN, HDBSCAN, and Birch - in identifying clusters of individuals at high risk of bankruptcy. By employing various evaluation metrics, including Davies-Bouldin Index, Silhouette Score, Adjusted Rand Index, Homogeneity, Completeness, V-Measure, and Accuracy, this research provides a comprehensive analysis of which clustering method best predicts personal bankruptcy clusters. The insights gained from this comparison can significantly enhance credit risk management and decision-making processes in financial institutions.

Personal bankruptcy remains a significant concern for financial institutions, affecting their portfolios and overall financial health. Individuals declaring bankruptcy often exhibit distinct financial behaviors and characteristics, such as high loan amounts, low credit scores, substantial monthly debts, and multiple credit problems. These features, when analyzed correctly, can provide insights into the likelihood of an individual defaulting on their financial obligations. Understanding these patterns and predicting potential bankruptcies can help banks and financial institutions take preemptive actions, such as adjusting credit limits or interest rates, to mitigate risks.

The results of this study demonstrate the varying effectiveness of different clustering methods in predicting personal bankruptcy. Density-based methods like DBSCAN and HDBSCAN achieve high accuracy but struggle with alignment to true data distributions. K-Means provides balanced performance but lacks the accuracy of density-based methods. Birch shows promise in forming well-separated clusters but requires further refinement for optimal accuracy. These insights underscore the potential of these clustering methods to significantly enhance predictive accuracy and alignment with true class distributions.

Building on this research, the next step will focus on understanding how clustering algorithms can accurately predict personal bankruptcy based on specific features. This approach will not only enhance the effectiveness of clustering algorithms but also provide deeper insights into the factors contributing to the risk of personal bankruptcy. Firstly, we will identify key financial features such as credit scores, monthly debt amounts, number of loans, payment history, monthly income, and debt-to-income ratio. Next, techniques like PCA or t-SNE will be applied to retain important information. Then, we will use evaluation methods such as SHAP and LIME to determine the importance of each feature. Finally, we will optimize the clustering algorithms and integrate them into the risk management systems of financial institutions. This research direction will significantly contribute to improving risk management and decision-making processes, helping financial institutions predict and prevent personal bankruptcy risks more effectively.

In summary, this comprehensive analysis highlights the importance of choosing the right algorithm based on the dataset's characteristics and the analysis goals. The integration of clustering methods can significantly enhance predictive accuracy and risk management strategies, contributing to more effective credit risk management in financial institutions.

The study limitations and future research: The study relies on a dataset from Vietnamese commercial banks spanning from 2010 to 2022, which may not be representative of other regions or periods. The generalizability of the findings might be limited due to regional economic conditions and banking practices. Some clustering algorithms, such as DBSCAN and HDBSCAN, are sensitive to parameter settings, which can lead to inconsistent results if not properly tuned. This sensitivity might affect the robustness of the clustering outcomes. While the clustering algorithms identified high-risk clusters effectively, the interpretation of these clusters in practical terms remains challenging. The clusters may not align perfectly with real-world categories used by financial institutions. Therefore, future research directions

could focus on identifying and incorporating additional financial and non-financial features that could improve the predictive power of clustering algorithms. This might include behavioral data, macroeconomic indicators, and more detailed credit history. Finally, combining clustering methods with machine learning techniques, such as ensemble learning, could enhance prediction accuracy and robustness. Future studies could explore hybrid models that integrate the strengths of different algorithms. By addressing these ethical concerns, the study will provide a more comprehensive view of the potential risks associated with using customer data and algorithmic decision-making in the financial sector. This section highlights the importance of privacy, fairness, transparency, and bias mitigation, helping ensure that predictive models are practical and ethically sound.

6. Declarations

6.1. Author Contributions

Conceptualization: N.M.N.; Methodology: N.M.N.; Software: N.M.N.; Validation: N.M.N.; Formal Analysis: N.M.N.; Investigation: N.M.N.; Resources: N.M.N.; Data Curation: N.M.N.; Writing Original Draft Preparation: N.M.N.; Writing Review and Editing: N.M.N.; Visualization: N.M.N. All author has read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The author received no financial support for the research by Ho Chi Minh University of Banking.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The author declares that the author has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation", *Electronics*, vol. 9, no. 8, pp. 1-12, 2020.

[2] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach", *Expert Systems with Applications*, vol. 64, no. 1, pp. 36-55, 2016.

[3] M. Ala'raj and M. F. Abbod, "Classifiers consensus system approach for credit scoring", *Knowledge-Based Systems*, vol. 104, no. 7, pp. 89-105, 2016.

[4] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints", *Information Retrieval*, vol. 12, no. 7, pp. 461-486, 2009.

[5] J. N. Crook, D. B. Edelman, and L. C. Thomas, "Recent developments in consumer credit risk assessment", *European Journal of Operational Research*, vol. 183, no. 3, pp. 1447-1465, 2007.

[6] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques", *Journal of Intelligent Information Systems*, vol. 17, no. 1, pp. 107-145, 2001.

[7] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks", *Expert Systems with Applications*, vol. 117, no. 3, pp. 287-299, 2019.

[8] G. Kou, Y. Peng, and G. Wang, "Evaluation of clustering algorithms for financial risk analysis using MCDM methods", *Information Sciences*, vol. 275, no. 8, pp. 1-12, 2014.

[9] P. R. Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review", *European Journal of Operational Research*, vol. 180, no. 1, pp. 1-28, 2007.

[10] X. Li, P. Zhang, and G. Zhu, "DBSCAN clustering algorithms for non-uniform density data and its application in urban rail passenger aggregation distribution", *Energies*, vol. 12, no. 19, pp. 1-22, 2019.

[11] M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano, "Corporate default forecasting with machine learning", *Expert Systems with Applications*, vol. 161, no. 1, pp. 1-17, 2020.

[12] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions", *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007-3057, 2020.

[13] D. L. Olson, D. Delen, and Y. Meng, "Comparative analysis of data mining methods for bankruptcy prediction", *Decision Support Systems*, vol. 52, no. 2, pp. 464-473, 2012.

[14] Y. Qu, P. Quan, M. Lei, and Y. Shi, "Review of bankruptcy prediction using machine learning and deep learning techniques", *Procedia Computer Science*, vol. 162, no. 2019, pp. 895-899, 2019.

[15] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, no. 1987, pp. 53-65, 1987.

[16] S. D. Vrontos, J. Galakis, and I. D. Vrontos, "Modeling and predicting US recessions using machine learning techniques", *International Journal of Forecasting*, vol. 37, no. 2, pp. 647-671, 2021.

[17] L. Wang, P. Chen, L. Chen, and J. Mou, "Ship AIS trajectory clustering: An HDBSCAN-based approach", *Journal of Marine Science and Engineering*, vol. 9, no. 6, pp. 1-20, 2021.

[18] R. Xu and D. Wunsch, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[19] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models", *Expert Systems with Applications*, vol. 38, no. 10, pp. 13274-13283, 2011.

[20] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases", A*CM Sigmod Record*, vol. 25, no. 2, pp. 103-114, 1996.

[21] F. Yang and S. Gu, "Industry 4.0, a revolution that requires technology and national strategies", *Complex and Intelligent Systems*, vol. 7, no. 3, pp. 1311-1325, 2021.

[22] M. Westerdijk, J. Zuurbier, M. Ludwig, and S. Prins, "Defining care products to finance health care in the Netherlands", *The European Journal of Health Economics*, vol. 13, no. 2, pp. 203-221, 2012.

[23] D. Tingley, J. Ásmundsson, E. Borodzicz, A. Conides, B. Drakeford, I. R. Eðvarðsson, D. Holm, K. Kapiris, S. Kuikka, and B. Mortensen, "Risk identification and perception in the fisheries sector: Comparisons between the Faroes, Greece, Iceland and UK", *Marine Policy*, vol. 34, no. 6, pp. 1249-1260, 2010.

[24] Y. Thakare and S. Bagal, "Performance evaluation of K-means clustering algorithm with various distance metrics", *International Journal of Computer Applications*, vol. 110, no. 11, pp. 12-16, 2015.

[25] V. Singhal, A. B. Singh, V. Ahuja, and R. Gera, "Consumer segmentation in the fashion industry using social media: An empirical analysis", *Journal of Information and Organizational Sciences*, vol. 47, no. 2, pp. 399-419, 2023.

[26] A. Sheikh, T. Ghanbarpour, and D. Gholamiangonabadi, "A preliminary study of fintech industry: A two-stage clustering analysis for customer segmentation in the B2B setting", *Journal of Business-to-Business Marketing*, vol. 26, no. 2, pp. 197-207, 2019.

[27] L. Rivera, D. Gligor, and Y. Sheffi, "The benefits of logistics clustering", I*nternational Journal of Physical Distribution and Logistics Management*, vol. 46, no. 3, pp. 242-268, 2016.

[28] M. Argüelles, C. Benavides, and I. Fernández, "A new approach to the identification of regional clusters: Hierarchical clustering on principal components", *Applied Economics*, vol. 46, no. 21, pp. 1-19, 2014.

[29] M. Bagherzadeh, M. Ghaderi, and A. S. Fernandez, "Coopetition for innovation – The more, the better? An empirical study based on preference disaggregation analysis", *European Journal of Operational Research*, vol. 297, no. 2, pp. 695-708, 2022.

[30] G. G. Dureti, M. P. Tabe-Ojong, and E. Owusu-Sekyere, "The new normal? Cluster farming and smallholder commercialization in Ethiopia", *Agricultural Economics*, vol. 54, no. 6, pp. 900-920, 2023.

[31] W. Hadhri, R. Arvanitis, and H. M'Henni, "Determinants of innovation activities in small and open economies: The Lebanese business sector", *Journal of Innovation Economics and Management*, vol. 21, no. 3, pp. 77-107, 2016.

[32] W. S. Hwang and H. S. Kim, "Does the adoption of emerging technologies improve technical efficiency? Evidence from Korean manufacturing SMEs", *Small Business Economics*, vol. 59, no. 2, pp. 627-643, 2022.

[33] G. J. Oyewole and G. A. Thopil, "Data clustering: Application and trends", *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6439-6475, 2023.