IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁎ Impact Factor 8.102 ⁎ Peer-reviewed / Refereed journal ⁎ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12805

Clustering Approach to High-Dimensional Data for Banking Customer Segmentation

Arif Farooq Shaikh [1], S.A Khedkar [2]

M.TECH, Computer Engineering and IT, Veermata Jijabai Technological Institute, Mumbai, India[1]

Assistant Professor, Computer Engineering and IT, Veermata Jijabai Technological Institute, Mumbai, India[2]

Abstract: This paper presents a comprehensive study and analysis on customer segmentation using banking data, aiming to enhance the accuracy and effectiveness of segmentation techniques. The rationale for conducting this research lies in the growing need for personalized services in the banking sector, where understanding customer behavior and preferences is crucial for strategic decision-making. The problem addressed in this study is the challenge of accurate and meaningful customer segmentation, considering the intricate patterns and complexities inherent in banking data. Conventional segmentation methods like k-mean, improve k-mean, and fuzzy c have been widely applied; however, their limitations in handling non-linear and complex data structures necessitate the exploration of more advanced techniques. The methodology employed involves a multi-faceted approach to address the segmentation challenge. Initially, conventional methods such as k-means, improved k-means, and fuzzy c-means are applied to the banking data to establish a benchmark for comparison. These methods are effective for relatively simple data distributions but may fall short in capturing intricate patterns. To address this, a novel approach utilizing spectral clustering is proposed. The proposed method, spectral clustering, leverages the spectral properties of the data to capture underlying structures and relationships. Unlike traditional methods, spectral clustering can effectively identify non-linear and complex patterns in the data, making it suitable for the nuances of banking customer behavior. Through experimentation and analysis, the proposed method's performance is evaluated against the established benchmarks, showcasing its potential to yield more accurate and meaningful customer segments. This research contributes to the field of customer segmentation in the banking sector by highlighting the limitations of traditional methods and introducing a novel spectral clustering approach. The customer segmentation using Neural Network and Spectral Clustering performs well compared to the previous research our proposed system gives an accuracy of 99.54 and also gives the best Gini obtained.

Keywords: Customer Segment, K–Means, Machine Learning, Banking Profiling, Spectral Clustering.

I. INTRODUCTION

Banks currently have access to enormous databases that contain data about their clients and the history of their transactions[1]. Banks must segment these large datasets into minute clusters in order to analyse consumer behaviour and suggest the best course of action to maximise benefits, customer satisfaction, and increased profitability. So we can easily detect the need of the customer and we can provide banking service as per need. To the customer, big data can be divided into small parts with the help of the machine learning algorithm. The provider now needs to increase client loyalty as a result of increased provider competition and the organization uses the client segment as a means of boosting customer satisfaction. It categorizes potential consumers into different categories based on demographic or user views. Most financial agencies now require customer segmentation to target offers to the right clients. In the modern banking industry, banks maintain extensive records including information about their clients and a history of transactions.

To analyse and optimize these consumer behaviors, banks can split these enormous data collections into smaller clusters. We must be able to offer recommendations for improved profit and contentment. To do this, client profiles or customer segmentation are employed. A client profile is created by profiling, giving the bank a detailed account of the consumer based on a variety of variables. The process of categorising different categories of clients based on their behaviours or other characteristics (such as location, age, or income for demographic segmentation or behavioural segmentation) is known as client segmentation. But "client segmentation" and "profiling" are seen as two sides of the same coin.. Default forecasting, risk management, customer retention, and client profiling are just a few of the difficulties that banks must overcome to increase profitability and lower risk. Thus, it's important to accurately identify clients to tackle these issues. The concept of machine learning is what enables computers to run without even being programmed.

One could be utilizing machine learning inadvertently thousands of times each day since it is so prevalent now. In addition to the examples already stored in the computer, machine learning may also be used for tasks that are yet to be performed. It imparts equation and function discovery to machines. Machine learning is essential for forecasting

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12805

consumer behaviors based on certain patterns or sequencing of occurrences, which not only enhances the level of connection with current customers i.e. clients, it the way banks re currently used by emphasizing the client[2]. (K-means, fuzzy c-means, ANN, etc) are some of the machine learning techniques in this research.

RFM model in a market studied on consumer segmentation. This model combines novelty, frequency, and money, where novelty refers to a customer's most recent activity, such as a visit or a purchase. Money represents the customer's desire to spend on purchases, whereas frequency seems to be the customer's transactions or visit [2]. For instance, the Obama re-election campaign in 2012 used weather mom to gather all of his demographic and regional data from the image and process the agenda appropriately, which resulted in success.

When analyzing consumer behavioral data to create important business choices, customer analytics is an essential component. According to our study statistics, 67% of customers list a negative experience as one of the main reasons they departed, with many choosing to go back to the same establishment and use the service as just a result of their negative experience 85% urge caution in others deal with the corporation commercially. Customer analytics is essential for understanding customer satisfaction and trust behaviors since they have a big impact on company scenarios.

Results using ML with Customer Analytics are precise and dependable. Large businesses may better target both current and potential customers by using customer segmentation. Wherein regular consumers receive awards and limited families receive extra benefits. Big Data keeps track of the top clients based on their monthly use. The customer gets an invitation to join this program through a message [3].

II. RELATED WORK

There has been a variety of research been done related to Customer segmentation using machine learning as follows

In 2015, Sharahi and Aligholi [4] published a classification model utilizing two stages and the k-means clustering technique for the dataset of Sepah Bank Branches in Tehran. Segmenting the 60 businesses that made up Sepah Bank's clientele was a form of behavioral and demographic segmentation that aided in locating the most devoted clients.

In 2016, Ayoubi [5] described a two-step algorithm and Kohonen’s neural network-based client segmentation model in 2016. Using effective factors based on Customer Lifetime Value, segment customers (CLV). In this study, a dataset including information on 56000 "Taavon bank" clients was employed. First, the ideal number of clusters was established using a two-step methodology. Kohonen neural network was then used. Each cluster's value was determined using the WRFM (weight of recency, frequency, and monetary) model.

A five-year profile model for clients of a Portuguese retail bank was reported by Palaniappan et al. in 2017 [6] (2008 to 2013). This study aimed on identifying a group of consumers who were very likely to sign up for a long-term deposit and on assisting banks in improving the accuracy of their customer profiles through categorization. Naive Bayes, Random Forest, and Decision Tree were the three classification methods employed

Bansal et al. modified the k-means method in a clustering model in 2017 [7]. This adjustment based on normalisation. The researcher used the Cancer Set of data to make her discoveries. Despite the fact that the initial data were extremely multidimensional, just five attributes were finally considered based on necessity. This analysis showed that the total accuracy for such a modified strategy was 92.86% while the accuracy rate for the current method was 57.14%.

In 2017, Patil and Dharwadkar [8] built a predicting and classifying model for two datasets of data from bank customers. They used an Artificial Neural Network (ANN) in this model, then they assessed the results. It is shown that the ANN approach works well for the two datasets by combining the recommended model and the ANN algorithm. The accuracy rate of this approach was 72% and 98% for datasets 1 and 2, respectively.In 2018, Yang and Zhang [9] developed a classification method using five clustering algorithms again for Taiwanese banks’ credit card default data set. Using 10-fold cross- validation, the total under-the-curve area (AUC) and the model's accuracy rate were determined. Using Light GBM, an elevated Gradient Boosting framework developed by the Microsoft Company, the best performance rate was attained. For the Light GBM model, the efficiency ratio using the F1 measure was 89.34%.

A Taiwanese company was given a categorization algorithm for its debit or credit card failure set of data. by Niloy and Navid in 2018 [10]. To determine whether or not the user is the default credit cardholder, Naive Bayes Classifiers and Decisions Trees were utilized as classification methods. The outcome of this study demonstrated that Naive Bayesian had the highest accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ∺ Impact Factor 8.102 ∺ Peer-reviewed / Refereed journal ∺ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12805

For 18 datasets from the UCI collection, A multi-class clustering algorithm was released by Arshad et al [11] in 2019. This study employed the Deep Fuzzy C-Mean with Semi-Supervision (DFCM-MC) to cluster semi-supervised data. They applied a new label to data sets by utilizing fuzzy c-means. With the use of the new label, they combined labeled (supervised data) and unlabeled (unsupervised data) data to extract the discriminating information needed for classification. The f-measure was 78.16%, while the accuracy rate of the DFCM-MC was 80.82%.

In the study of [12], customer turnover was predicted using decision trees and logistic regression. According to their findings, Decision trees fared better on their datasets than Logistic Regression. And utilizing a Decision Tree, they achieved the maximum accuracy of 99.67% on the sizable dataset.

This [13] utilized KNN, Random Forest, and XGBoost to forecast customer attrition. They found that XGBoost had the best accuracy (79.8%), as well as the highest AUC (58.2%).

In the study by [14], the effectiveness of more than 100 classifiers was evaluated for the telecom industry's churn prediction problem. They discovered that Bagging Random Forest outperformed them all in regards to AUC, getting 67.20%, while Regularized Random Forest outperformed them all in terms of accuracy, achieving 73.04%.

Idris et al [15] is investigation of particle swarm optimization (PSO) based undersampling techniques combined with various feature reduction techniques included the usage of random forest & KNN classifiers According to experimental results, the PSO, mRMR, and RF-based approach (Chr-PmRF) had the greatest AUC, at 75.11%.

III. PROPOSE MODEL

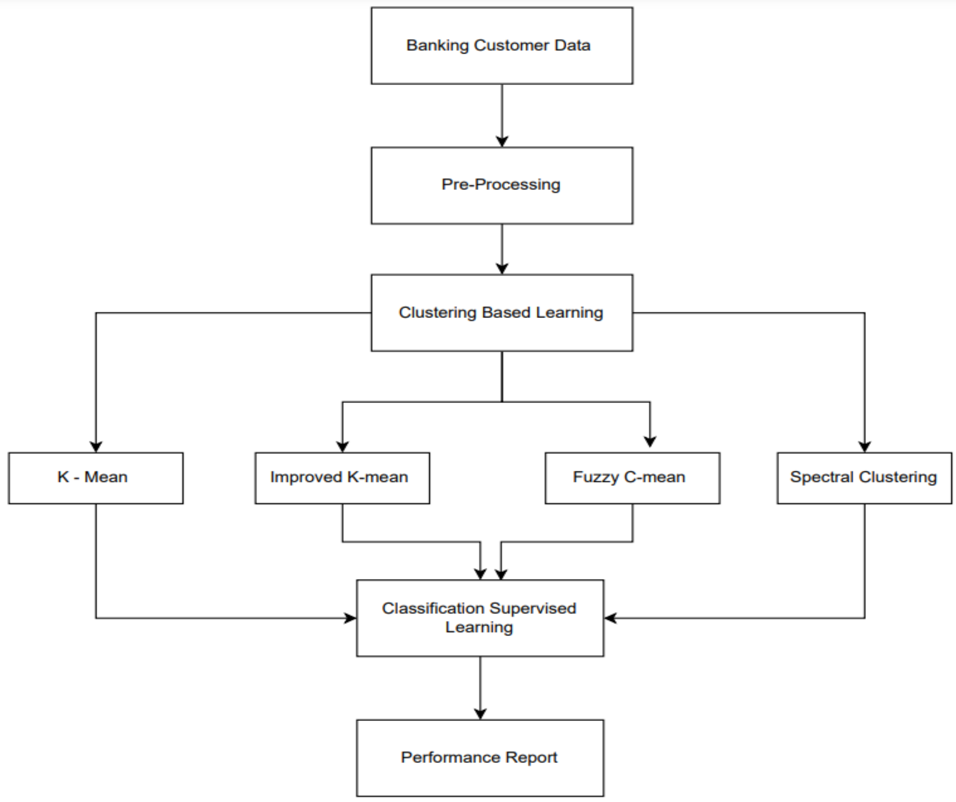


Figure 1: Spectral Clustering Based Banking Customer Segmentation Framework

K-mean Clustering

The Mac Queen's suggested simplicity and by using K-mean clustering approach's stability have made it one of the most frequently used techniques for years. The K-Means clustering method was used to create a partition-based clustering approach in 1967. By using K-means, items are divided into groups that are "similar" to one another and "dissimilar" to objects in other clusters. Cluster analysis frequently uses the K-means approach because of its greater efficiency,

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023
DOI: 10.17148/IJARCCE.2023.12805

scalability, and ability to quickly converge when working with large data sets. Unsupervised learning techniques like K-means clustering are employed when there are no labels on the data, or when there are no clearly defined categories or groupings/ groups.

$j = \sum i = 1n \sum j = 1k||xi - Cj||2$ (1)

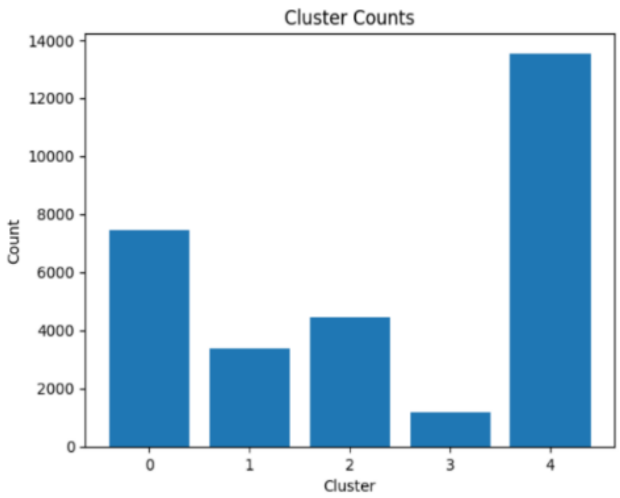


Figure 2: K-mean Clustering to identify customer numbers in each cluster

Improved k-mean

Since it can automatically calculate the necessary number of clusters and assign the necessary cluster to unclustered points, the k-means clustering approach has been improved. The recommended adjustment makes the dissimilarity-based k-means clustering approach more efficient and provides good accuracy while cutting the clustering time for the cluster member.

The Huffman tree, which makes use of the dissimilarity matrix, is used to select the initial centroids. Numerous experiments demonstrate the effectiveness of the new approach, which maintains the same algorithm time complexity while offering improved clustering accuracy [19]. Improved k-mean is just an extended version of the k-mean.

So this is how we have discussed many different types of clustering /grouping approaches in this study we will see how we can use this algorithm in the banking sector or how we can apply the banking data set so we can get use full information from that data so we can use in future to improve the banking profile and help the banking sector to gain more profit from that.

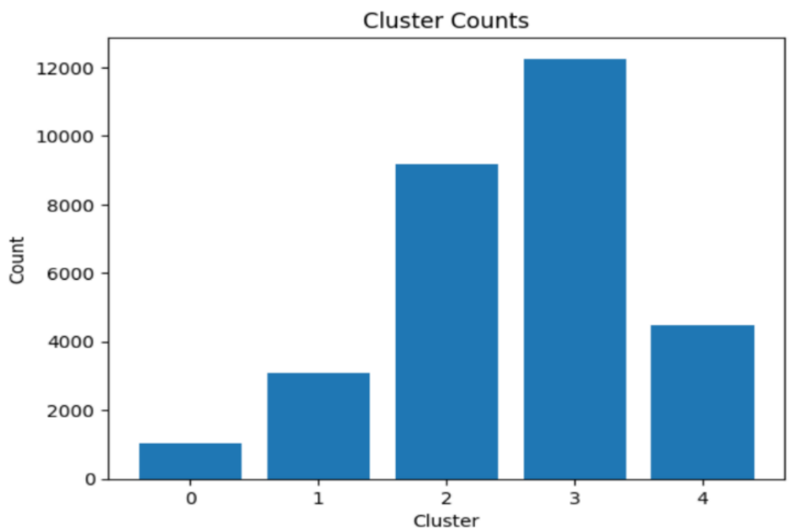


Figure 3. Improved K -K-mean cluster Graph representation Clustering to identify customer numbers in each cluster

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023
DOI: 10.17148/IJARCCE.2023.12805

Fuzzy C-mean

In fuzzy clustering, sometimes referred to as soft clustering, any data point can be a member of multiple groups. Data points may be a part of many clusters when using fuzzy clustering. One of the most well-known fuzzy clustering techniques is fuzzy C-means clustering (FCM). J.C. Bezdek revised J.C. Dunn's 1973 design for clustering utilising (FCM). The approach disregards noise and outliers in favour of centroid or clustering calculation optimisation.

Calculate the Fuzzy membership μij using the :$\mu ij = 1/ \sum_{k=1}^{c}$ (2)

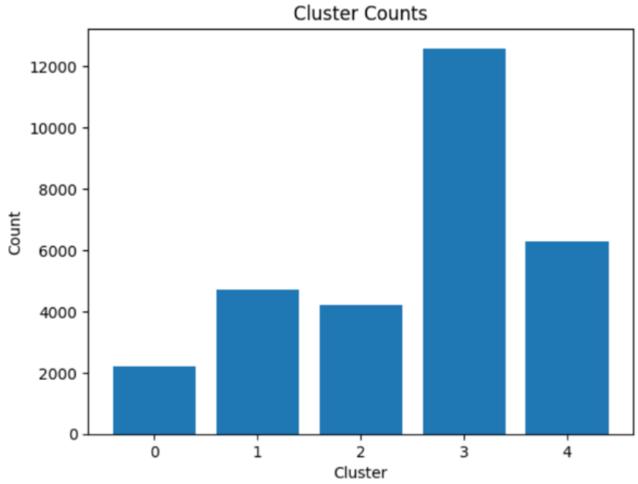


Figure 4. Fuzzy c-mean clustering representation Clustering to identify customer numbers in each cluster

Spectral clustering

Spectral Clustering is a distinctive approach within the realm of clustering algorithms that leverages the inherent connectivity among data points to establish meaningful clusters. Unlike traditional methods, which primarily rely on proximity measures, Spectral Clustering harnesses the eigenvalues and eigenvectors of the data matrix to transform the data into a lower-dimensional space, facilitating effective clustering.

The foundation of Spectral Clustering is rooted in the concept of representing data points as nodes in a graph, where the interactions or similarities between these data points are depicted as edges connecting the nodes. This graph-based perspective encapsulates the essence of the relationships between data points more comprehensively than mere distance-based approaches. By translating the data into this graph representation, Spectral Clustering can capture intricate structures and patterns that conventional methods might overlook.

Central to the Spectral Clustering process is the utilization of eigenvalues and eigenvectors, which enable the algorithm to discern the underlying structure of the data. These mathematical entities provide insights into how the data points are interconnected within the graph, offering a powerful tool to transform the data into a space that accentuates its inherent clustering tendencies. The eigenvalues serve as indicators of the importance of each eigenvector in representing the data, and by sorting and selecting a subset of these eigenvectors, Spectral Clustering effectively reduces the dimensionality while preserving crucial clustering information.

Our suggested model's main objective, as shown in the architecture, is to apply multiple machine learning techniques to create a more accurate behavioural profile of bank customers. This model starts with data that was gathered from the UCI machine learning repository. The next step is to prepare the data. Then, machine learning techniques are used to generate the customer profile. Machine learning's profiling phase recognises the components of a collection and places them in target categories. In this study, the unsupervised Gini coefficient is used to evaluate the accuracy rate of the techniques. After that, the results are used as input for supervised procedures.

In unsupervised algorithms, we have used six algorithms k-nn, Gaussian Naive Bayes, Random Forest, Decision Trees Classifier, XGBoost, and Logistic Regression. We have all output of the unsupervised algorithm which is mentioned in the architecture, that output becomes the input of the supervised technique.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12805

Algorithm 1: Proposed Spectral clustering Algorithm Pseudocode

Input: Data points X, Number of clusters k, Gaussian kernel parameter sigma
Output: Cluster assignments for each data point

1. **Compute Affinity Matrix W:**
   for each data point $x\ i\ in\ x$
   for each data point $xj\ in\ x$
   $$similarty = \exp\left(-norm(x\ i - x\ j) * * \frac{2}{2 * sigmaa ** 2}\right)$$
   $W[i, j] = similarity$

2. **Compute Degree Matrix D:**
   for each data point $x\ i\ in\ X$:
   $$D[i, j = sum\ (w[i]$$

3. **Compute Unnormalized Laplacian Matrix L:**
   for each data point $x\ i\ in\ X$:
   for each data point $x\ j\ in\ X$
   $$L[i, j] = D[i, j] - W[i, j]$$

4. **Compute Eigenvectors and Eigenvalues:**
   eigenvalues, eigenvectors = eig(L)

5. **Sort Eigenvectors:**
   sorted_indices = argsort(eigenvalues)
   sorted_eigenvectors = eigenvectors[:, sorted_indices]

6. **Select Eigenvectors:**
   selected_eigenvectors = sorted_eigenvectors $[1: k + 1]$

7. **Form Matrix U':**
   U_prime = vstack(selected_eigenvectors.T)

8. **Normalize Eigenvectors:**
   U_normalized = U_prime / norm(U_prime, axis=1, keepdims=True)

9. **K-Means Clustering:**
   kmeans = KMeans(n_clusters=k)
   cluster_assignments = kmeans.fit_predict(U_normalized)

Table I. Proposed model pseudo-code steps

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 8, August 2023
DOI: 10.17148/IJARCCE.2023.12805

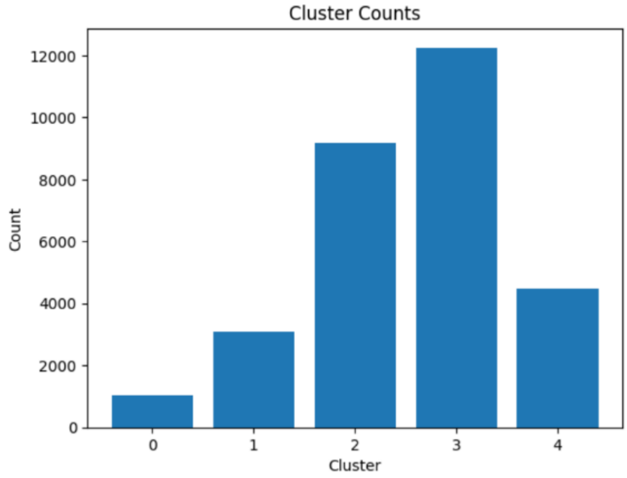


Figure5. Spectral Clustering representation Clustering to identify customer numbers in each cluster

Evaluation metrics

- Gini Index :

$$Gini = 1 - \sum_{i=1}^{n} pi\text{^}2 \qquad (3)$$

- The effectiveness of categorization was evaluated using the metrics shown in the equation below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

$$Precision = \frac{TP}{TP+FP} \qquad (5)$$

$$Recall = sensitivity\frac{TP}{TP+FN} \qquad (6)$$

$$F - measure = 2 * \frac{(Precision*recall)}{(precision+recall)} \qquad (7)$$

Where:

- True Positive(TP) : Observation is positive, and it is predicted to be positive.
- False Negative (FN): Observation is positive, but it is predicated negative.
- True Negative (TN): Observation is negative, and it is predicted to be negative.
- False Positive (FP): Observation negative but it is predicated positive.

IV. EXPERIMENT AND RESULTS

A. Data sets

The UCI (University of California, Irvine) Machine Learning Repository's repository contains the data set known as "default of credit card clients" [19]. It was just released in the 2015-obtained dataset. The dataset's attribute details appear in Table 1. The data collection includes 30000 observations. There's no missing data on any of the 23 variables or vacations it. All explanatory factors underwent normalization. Standardizing data is a stage in data preparation used on variables to scale these factors to a comparable range.

This study sought to in Taiwan and the situation of customers' defaulted payments analyze the rate of consumer profiling accuracy between four methods for machine learning. Consequently, among the four An artificial neural network is a tool for machine learning approaches the unique that he artificial neural network is the only machine learning method out of the four that can accurately profile the data set.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 8, August 2023
DOI: 10.17148/IJARCCE.2023.12805

Attribute No	Attribute Name	Description
X1	Limit_BAL	Amount of the given credit
X2	Sex	Gender (1=male,2=female)
X3	Education	Education (1=gradute school,2=university 3=high school 4= other
X4	Marital status	Marital status (1= married ; 2= single ;3= other
X5	Age	Age(year)
X6-X11	Pay_0 to Pay_6	April to September
X12-X17	Bill_AMT1 to Bill_AMT6	Amount of bill statement
X18-X23	Pay_AMT1 to Pay_AMT6	Amount of previous payment
X24	Y	Defult payment (yes=1 ,NO = 0

Table I. Data set Representation

B. Data preprocessing

The first crucial step in the data preprocessing is to remove irrelevant and unnecessary information in the dataset or if the data is noisy and unreliable. As a result, the quality and representation of data come first and are crucial before training the model. The normalization is applied on the dataset to scale the data value, initially remove the units of measurement for data and enable easy . Among the most typical methods for normalizing data is: adjusting data so that values range from 0 and 1.

Calculate the mean (μ) of the non-missing values in the dataset:

C. Machine learning techniques

Data preprocessing results in the creation of the final training set, which is then used to apply the four machine learning techniques to the final training set. The first algorithmic technique employed was K-means. The researcher's prior information is used to calculate the overall number of clusters. The researcher in this work consequently discovered that there were five clusters.

The second classifier determines that there are five clusters using an improved version of k-mean.

1. Calculating the average of all intra-cluster distances, which is simply the distance between a point and its cluster center, where N denotes the number of pixels in the image, K the number of clusters, and zi the cluster center of clusterci. Obviously, we want to take as few steps as possible.
2. Minimizing this action is the next stage. Measuring the distance between clusters, or the inter-cluster distance, must be as great as possible then compute taking the smallest value of this distance between cluster centres.
where zi and zj serve as cluster centers. The number of clusters is k.
3. The other larger values will naturally be larger than this value when only the minimum of this value is taken into account, along with the least possible distance to be maximized.
4. Finally, determine the inter-to-intra ratio that is considere* valid: $Validity = Internal/External$ (8)
5. A result of the clustering which provides a minimum value for the validity measure informs us of the optimal value (number of clusters).

The third fuzzy c-mean classifier, which was applied to the data set using five clusters is the third classifier.
The fourth classifier is the Spectral clustering Algorithm this is our main proposed method in this paper the pseudo-code which is maintained above in the spectral clustering part

DOI: 10.17148/IJARCCE.2023.12805

V. EXPERIMENT EVALUATION AND ANALYSIS

In this Experiment Evaluation and analysis we have compared the Gini index and Accuracy, precession, Recall and F1-score so our propsed system Spectral clustering have a good gini index and in terms of Accuracy , recession ,recall ,f1-score our proposed system is good as compared to previous research .

So below table give the rank of the algorithms based on gini value the Spectral has high rank and after that imporved k -meann ,Fuzzy c mean , and last rank in K-mean.

TABLE II. Result of Gini index of all unsupervised technique

Machine learning algorithms	Gini obtain	Rank
K-mean	26.948	4
Improved K-mean	34.411	2
Fuzzy C -mean	27.591	3
Spectral clustering	37.911	1

The Gini coefficient for every one of the 4 unsupervised algorithms with the best accuracy for profiling the dataset is then calculated.

Finally, submit a (ANN). To measure the correctness of a neural network, we use the outcomes of unreviewed procedures as a target. Our fuzzy C-means, K-means, improved K-means, and spectral clustering findings are used as targets to add a new label to the dataset. Then put them to the test and compare seven accuracy measures to see if they are accurate. The best classifiers for boosting bank client profiles are those with the highest accuracy.

So as you can see in table no .II our proposed Algorithm which is Spectral Clustering is giving more Gini index as compared to K-mean, improved k-mean,and fuzzy C mean.

This means the Spectral clustering is good for the other three algorithms that were used in previous research.

Below is the comparison of the Accuracy of all Machine learning techniques.

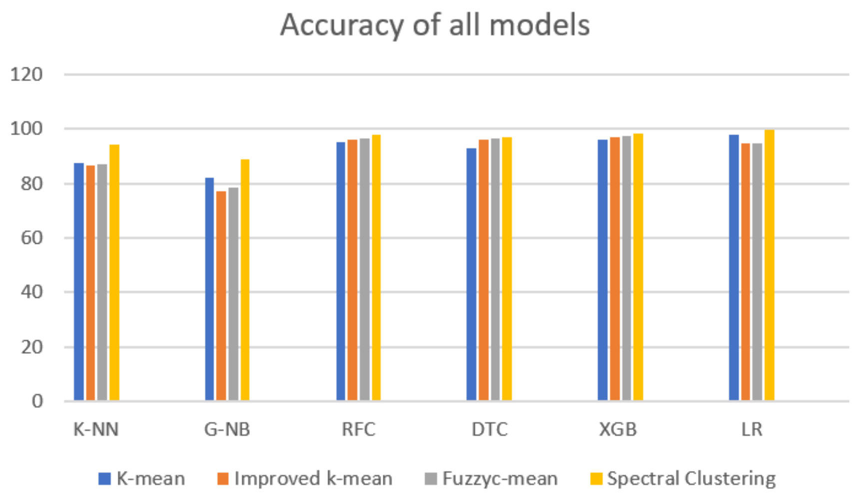
Our Proposed system (“Spectral Clustering”) is high as compared to the all-previous research model. Likewise we have compared the all model Precession, Recall and F1 -Score.

TABLE III. Accuracy for all the models

	K-mean	**Improved k-mean**	**Fuzzyc-mean**	**Spectral Clustering**
K-NN	87.61	86.48	87.01	**94.3**
G-NB	82.05	77.34	78.45	**88.79**

DOI: 10.17148/IJARCCE.2023.12805

RFC	95.13	96.19	96.58	**97.91**
DTC	93.12	95.95	96.58	**97.11**
XGB	96.12	97.14	97.33	**98.46**
LR	98.01	94.65	95.01	**99.54**



Based on these results, for classification tasks, XGBoost and Logistic Regression are top choices, while for clustering, Spectral Clustering is recommended. Sp We have done the extensive analysis of our propose model with previous research here we have compared the Accuracy of all model with our proposed model so our proposed model have performed good in this as you can see below.

TABLE III. Precession for all the models

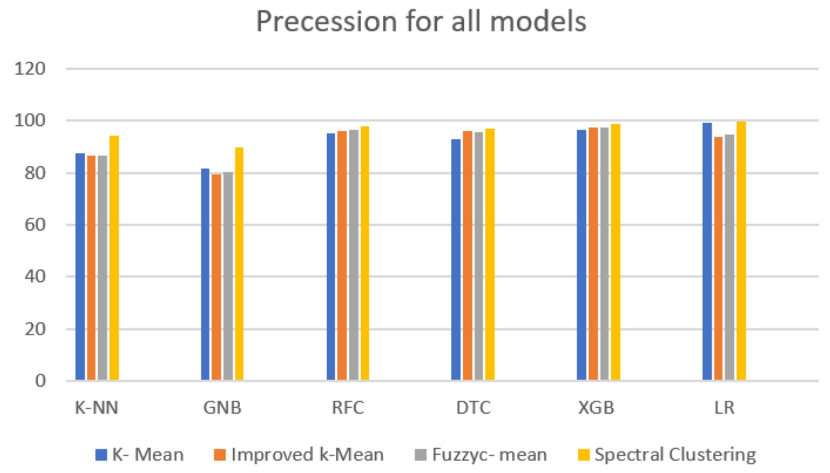
	K- Mean	**Improved k-Mean**	**Fuzzyc- mean**	**Spectral Clustering**
K-NN	87.62	86.5	86.5	**94.31**
GNB	81.61	79.63	80.12	**89.9**
RFC	95.41	96.17	96.7	**97.94**
DTC	93.01	95.96	95.68	**97.13**
XGB	96.46	97.34	97.34	**98.92**
LR	99.38	94.08	94.68	**99.54**

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁓ Impact Factor 8.102 ⁓ Peer-reviewed / Refereed journal ⁓ Vol. 12, Issue 8, August 2023

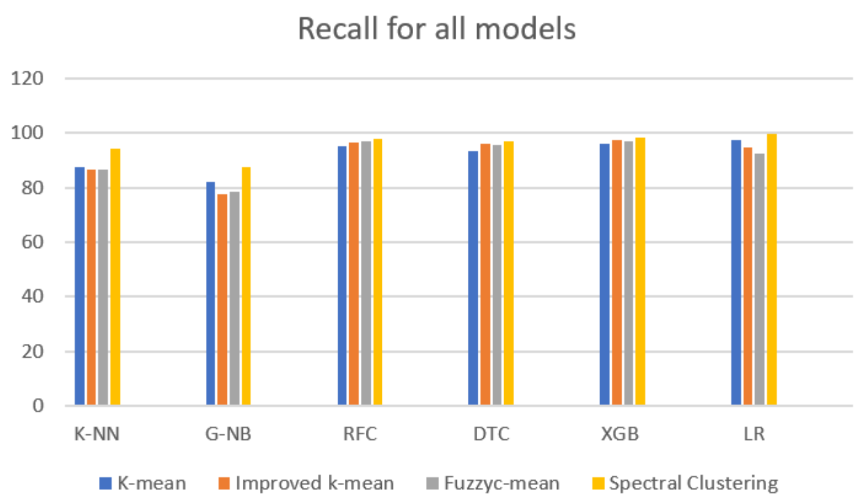
DOI: 10.17148/IJARCCE.2023.12805



In conclusion, the research findings highlight the efficacy of XGBoost and Logistic Regression for classification tasks, owing to their robustness and accuracy across different datasets. On the clustering front, Spectral Clustering's ability to reveal intricate structures within data sets it apart as a powerful tool for uncovering hidden patterns. These insights provide valuable guidance for practitioners seeking optimal algorithms in classification and clustering endeavors.

TABLE IV Recall for all the models

	K -mean	Improved K- mean	Fuzzy C mean	Spectral Clustering
K-NN	87.61	86.48	86.48	**94.31**
G-NB	82.01	77.45	78.65	**87.78**
RFC	95.39	96.78	97.12	**97.91**
DTC	93.23	96.16	95.66	**97.11**
XGB	96.13	97.33	97.13	**98.46**
LR	97.31	94.65	92.46	**99.54**

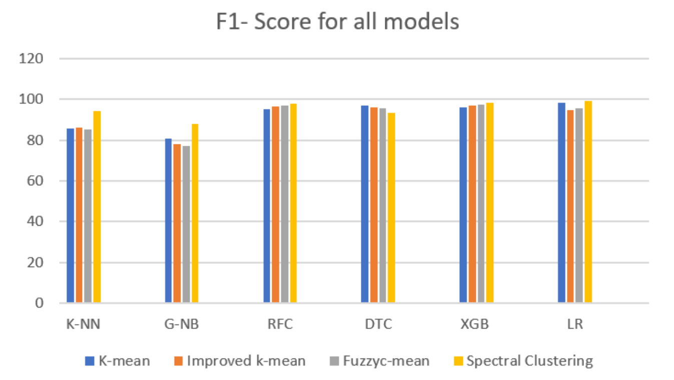


So above we have compared the recall parameters of all the Algorithms Spectral clustering stand out in this as well but as you see in this XG boost and the linear regression both are good in the classification part . So same we have done this for Recall comapred the all Algorithms the highlight vlaue is high as you can see the propose system has high value in this also.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁒ Impact Factor 8.102 ⁒ Peer-reviewed / Refereed journal ⁒ Vol. 12, Issue 8, August 2023
DOI: 10.17148/IJARCCE.2023.12805

TABLE V. F-1 score for all the models

	K-mean	Improved K-mean	Fuzzy C mean	Spectral Clustering
K-NN	85.63	86.41	.13	**94.27**
GNB	80.7	78.04	77.04	**88.08**
RFC	95.13	96.76	97.12	**97.91**
DTC	97.12	95.95	95.67	**93.22**
XGB	96.12	97.13	97.33	**98.46**
LR	98.45	94.66	95.65	**99.45**



So same as this F1 score we have analysis as we can see the spectral clustering giving great results in all the format.

After all the evaluation of all the unsupervised algorithms with supervised classification we get spectral clustering is best and the second best algorithm in Improved k-mean .

TABLE VI Cluster result from proposed Spectral Clustering.

Cluster	Name of Cluster	No of Customer
0	Platinum cluster	1026
1	Gold Cluster	3083
2	Bronze Cluster	9174
3	Silver Cluster	12249
4	Classic Cluster	4468

TABLE VII the result of the previous research is given below.

Paper	Year	Methods Used	Machine Learning Models	Accuracy rate %
[16]	2018	Kmean DBSCAN	artifical neural network k -eman	88.83%
[17]	2017	Kmean , Fuzzay	Neural network	90.99%
[18]	2017	Keman , Herachical Cluster , PCA	Naïve Bayes	81.7%
[19]	2017	K mean , C mean	Neural network	81.1%
[20]	2019	K-medoid	Deep Learning	98%
Proposed System of this Research	**2023**	**Spectral Clustering**	**ANN**	**99.54%**

Table VII demonstrates that when we compared our findings to those in papers [7], [22] and [24], we discovered that our suggested model performed the best in terms of accuracy measurements. We have discovered previous studies utilizing the identical dataset and ANN methodology Our Proposed method of spectral Clustering gives the Accuracy of 99.54.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 8, August 2023

DOI: 10.17148/IJARCCE.2023.12805

VI. CONCLUSION

Thanks to profile, the banks have been able to establish a committed connection based on humanistic experiences and trust. Clustering methods are used to separate large datasets into groups. suggested layout Spectral clustering was used to solve the accuracy level and calculation time problems, which also solved the other two main problems with K-Means clustering.

Careful investigation of the profiling environment should be done in order to effectively and efficiently segment the bank's customer base and help construct its service and product offerings to achieve customer loyalty and satisfaction. By adding a new label target to the dataset, the machine learning supervised method produced more accurate findings for profiled than the unsupervised method. So Our Proposed model gives the highest accuracy of 99.54% by Four different measures. In order for any bank to utilize this model and technique in the future, increase client profile, achieve high profitability, and lower risk. In subsequent research, we want to improve the performance and efficacy of our proposed technique using some deep learning algorithms. in relation to health informatics.

REFERENCES

[1]. A.Banduni, A. Ilavendhan, "Customer segmentation using machine learning", International journal of innovative research in technology, Volume 7, Issue 2, ISSN: 2349-6002, July(2020) research .

[2]. M.Alkhayrat, M. Aljnidi&K.Aljoumaa, K., "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA". J Big Data 7, https://doi.org/10.1186/s40537- 020-0286-0. February(2020).

[3]. C.Qiuru, L Ye, X. Haixu, L. Yijun, Z, "Telecom Customer Segmentation Based on Cluster Analysis", International Conference on Computer Science and Information Processing (CSIP), published in 2012, DOI:10.1109/CSIP.2012.6309 (2012).

[4]. Fh M. Sharahi and M. Aligholi, ''Classify the data of bank customers Usinusing dating and clustering techniques (case study: Sepah bank branches Tehran),'' J. Appl. Environ. Biol. Sci., vol. 5, no. 5, pp. 458–464, 2015.I. S. Jacobs and C. P. Bean, "Fine particles, thin films, and exchange anisotropy," in Magnetism, vol. III, G. T. Rado, and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[5]. M. Ayoubi, ''Customer segmentation based on CLV model and neural network,'' Int. J. Comput. Sci., vol. 13, no. 2, p. 31, 2016.

[6]. S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, ''Customer Profiling using classification approach for bank telemarketing,'' Int. J. Inform. Vis., vol. 1, nos. 2–4, pp. 214–217, 2017.

[7]. A. Bansal, M. Sharma, and S. Goel, ''Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining,'' Int. J. Comput. Appl., vol. 157, no. 6, pp. 1–7, 2017.-

[8]. P. S. Patil and N. V. Dharwadkar, ''Analysis of banking data using machine learning,'' in Proc. Int. Conf. IoT Social, Mobile, Analytics, Cloud (I-SMAC), Palladam, India, Feb. 2017, pp. 876–881.

[9]. S. Yang and H. Zhang, ''Comparison of several data mining methods in credit card default prediction,'' Intell. Inf. Manage., vol. 10, no. 5, p. 115, 2018.

[10]. N. H. Niloy and M. A. I. Navid, ''Naïve Bayesian classifier and classification trees for the predictive accuracy of probability of default credit card clients,'' Amer. J. Data Mining Knowl. Discovery, vol. 3, no. 1, p. 1, 2018.

[11]. A. Arshad, S. Riaz, and L. Jiao, ''Semi-supervised deep fuzzy c-mean clustering for imbalanced multi-class classification,'' IEEE Access, vol. 7, pp. 28100–28112, 2019.

[12]. S. S.-Schwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[13]. D. D. Adhikary and D. Gupta, ''Applying over 100 classifiers for churn prediction in telecom companies,'' Multimedia Tools Appl., vol. 248, pp. 1–22, Aug. 2020.

[14]. A. Idris, M. Rizwan, and A. Khan, ''Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies,'' Comput. Electr. Eng., vol. 38, no. 6,pp. 1808–1819, Nov. 2012.

[15]. S. Yang and H. Zhang, ''Comparison of several data mining methods in credit card default prediction,'' *Intell. Inf. Manage.*, vol. 10, no. 5, p. 115, 2018.

[16]. S. Imtiaz and A. J. Brimicombe, ''A better comparison summary of credit scoring classification,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, pp. 1–4, 2017.

[17]. M. Pasha, M. Fatima, A. M. Dogar, and F. Shahzad, ''Performance com parison of data mining algorithms for the predictive accuracy of credit card defaulters,'' *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 3, pp. 178–183, 2017.

[18]. V. Pyzhov and S. Pyzhov, ''Comparison of methods of data mining tech niques for the predictive accuracy,'' Tech. Rep., 2017.

[19]. A. Amin, F. Al-Obeidat, B. Shah, M. A. Tae, C. Khan, H. U. R. Durrani, and S. Anwar, ''Just-in-time customer churn prediction in the telecommunication sector,'' J. Supercomput., vol. 76, no. 6, pp. 3924–3948, Jun. 2020