UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF INFORMATION SYSTEMS

FINAL PROJECT REPORT

DATA ANALYTICS WITH R/PYTHON

TOPIC: INTEGRATION OF THE K-MEANS CLUSTERING ALGORITHM AND DIMENSIONALITY REDUCTION FOR DEEP CUSTOMER SEGMENTATION ON THE CREDIT CARD DATA

Lecturer: Mas. Nguyen Van Ho

Group: Bo Kho

Ho Chi Minh City, December 21, 2022

MEMBERS OF GROUP

NO.	**NAME**	**STUDENT ID**
1	Man Dac Sang	K204061446
2	Tran Nhat Nguyen	K204061440
3	Thai Thien Truc	K204060310
4	Nguyen Thai Ngoc Suong	K204061411

ACKNOWLEDGMENTS

We received a lot of dedicated cooperation from teachers during the research time. In addition, we cited a number of writers. We will be able to accomplish our final project to the best of our abilities because of your distribution.

Despite our best efforts, errors are unavoidable. As a result, we aim to collect your judges' opinions in order to enhance our research. These will be our great inspiration to construct our upcoming project.

COMMITMENT

We guarantee that our final DATA ANALYTICS WITH R/PYTHON project will be one-of-a-kind owing to the efforts of the entire team. There are still other materials we referred to, which we have included and quoted specifically in the report.

If all of the above is wrong, we will take all responsibilities from professors.

Ho Chi Minh City, December 21, 2021.

Committed person

GROUP Bo Kho

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACRONYMS	FULL NAME
CRM	Customer Relationship Management
SOM	Self-Organizing Map
RFM	Recency, Frequency, Monetary
DNNs	Deep Neural Networks
EDA	Exploratory Data Analysis
PCA	Principal Component Analysis
SSE	Sum of Squared Error
CHI	Calinski Harabasz Index
ANN	Artificial Neural Networks

ABSTRACT

It is well known that clustering in machine learning is a helpful method for customer segmentation. K-Means is a popular and efficient clustering algorithm, but this method has difficulty with data with too many dimensions. This study proposes a new approach that can solve the above problem, which is to use Autoencoder to reduce the data dimensionality and combine with K-Means to identify clusters. The proposed method has experimentally been conducted with a sample dataset containing information of historical credit card transactions. It was found that there are many differences between the results before and after using Autoencoder. There are also other characteristics that are illustrated in the report.

Keywords: *Customer segmentation, K-Means, Clustering, Dimension reduction, PCA, Autoencoder, Artificial neural network*

CHAPTER 1. INTRODUCTION

Due to the size of the modern business industry, it is crucial to spend money on providing good customer service. Customer is critical to any company that produces goods or services. We must do a task known as customer segmentation in order to invest in the correct customer. But not every company succeeds in winning over customers because, for one of many reasons, they struggle to develop compelling customer segmentation methods. Due to the competition to create a strategy that would draw in a big number of consumers, customer segmentation has a significant impact on the company's revenue. Customer segmentation is the process of segmenting customers based on common characteristics such as behavior, shopping habits... Effective segmentation is a factor to help businesses create appropriate marketing strategies with different customer groups, in order to attract the largest number of customers possible. Faced with this problem, we find that clustering is a reasonable choice to segment customers because the number of customers is large, but they will only belong to a few segments. Specifically, the clustering technology of unsupervised machine learning: K-Means. K-Means will do a good job of separating customers with similar characteristics into a group however, to ensure more accurate clustering, we have combined it with Autoencoder - a kind of Artificial Neural Networks Analysis (ANN) to reduce the data dimension, remove unnecessary data from which the clustering results are more accurate. The goal of the target after obtaining the standard clustering result is that it can be applied to the marketing field such as identifying customer buying patterns, market basket analysis, predicting customer purchases, target marketing, direct marketing, market analysis, CRM, and customer segmentation.

CHAPTER 2. MOTIVATION AND SOLUTION

2.1 Motivation

Due to the application of machine learning, businesses can better understand customer behavior and launch effective campaigns. A specific example for this phenomenon is customer segmentation. In order to increase customer loyalty, companies cluster their customers into different groups based on the collected data. Then, they can perform marketing strategies focused on targeted customer segments. However, with the dramatically increasing customer data, and number of attributes in the dataset it is challenging for companies to segment because of the curse of dimensionality [1]. The literature surveys [2], [3] show that one of the major applications of K-Means is customer segmentation. Yet, they have bad performance on high-dimensional datasets and produce inefficient clustering results because of the inefficiency of similarity measures used in them. In addition, using these techniques with large-scale datasets usually suffer from high computational complexity [3]. Therefore, it is necessary to have proper technique for processing high-dimensional datasets and combine with clustering techniques to get useful insights from customer data.

2.2 Solution

With high dimensional dataset, K-Means will suffer from the curse of dimension and dimensional reduction is widely used to relieve the problem [4] .There are many methods for dimensionality reduction. Deep learning is a subset of machine learning, they have been applied in various fields. Especially, Autoencoder is an effective technique for dimensionality reduction, which will reduce the size of the inputs for the subsequent clustering model. One more advantage is feature transformation that a set of features with high signal-to-noise ratios will be produced [5].Our team will use Autoencoder for data reduction before clustering by K-Means to enhance accuracy and processing time, mitigates constraints evaluated on large scale dataset [3]. Moreover, this research can be deployed in practice by businesses that have large-scale datasets with a lot of attributes that can cause the curse of dimensionality. Therefore, businesses can accurately segment customers into different clusters and give better insights.

CHAPTER 3. RELATED WORK

Customer segmentation is a powerful tool that can guide business toward more efficient ways to group customers into groups with different needs and characteristics .[6] reveals that segmentation is necessary since firms with limited resources must identify and focus on potential customers, not all customers. Clustering by unsupervised machine learning algorithms has been demonstrated to implement customer segmentation [7]. K-Means clustering algorithm has been used in different customer segmentation articles [8], [9]. [8] used k-Means to cluster customers with an accuracy of 95%. [10] has used RFM model and K-means algorithm to perform customer segmentation and analysis by online sales dataset. K-Means and SOM algorithms to cluster customers in the insurance field and detect customer features and demands in [11] article. [12] has implemented K-means, the Hierarchical clustering, and the Principal Component Analysis to define customer segments and propose marketing strategies for a credit card company. All researches above agree that K-Means is the most simple and suitable way to cluster customers into segments.

With more and more customer data being collected, a lot of new techniques have been developed to improve performance of K-Means algorithms on high dimensional data. To achieve better clustering performances, deep clustering models use deep neural networks (DNNs) with stronger non- linear representational capabilities. [5] have conducted deep customer segmentation based on a combination of a deep neural network and a self-supervised probabilistic clustering technique on Vietnamese supermarkets' data. [13] proposed a new deep clustering method named Deep Embedded K-Means that learns the deep embedding space and identifies clusters. [14] proposed a deep embedding network considering two constraints on learned representations and then followed by K-means to find clusters. Principal Component Analysis and Autoencoder Neural Network have been used to perform dimensional reduction before clustered by K-Means in [3]. This article reveals that after comparing the results of Principal Component Analysis and Autoencoder Neural Network methods, autoencoder neural networks obtain the best results and play an vital role in dimensional reduction. Deep autoencoders have been demonstrated to be extremely useful in dimensional reduction in [15].

CHAPTER 4. PRELIMINARIES

4.1 Customer Segmentation

Segmentation in marketing is a strategy for categorizing consumers or other entities based on characteristics such as demographics, behavior, lifestyle, and so on, allowing you to understand each segment in its entirety. This provides information that can be used to market to each segmented cluster of customers more effectively.

Credit card companies are performing customer segmentation to better understand their clients and offer user-centric solutions for targeted marketing, thanks to the increased usage of machine learning algorithms. Cluster analysis is an effective tool in this effort because one of its goals is to help analysts and marketers understand—cluster analysis is used to find influential groups in objects that share characteristics that help analysts and marketers in their daily analysis, description, and use of information hidden in groups [16].

4.2 K-Means Clustering

K-Means clustering is a machine learning algorithm that arranges unlabeled data points around a specific number of clusters. K-Means clustering is one of the techniques that is practical for consumer segmentation.

Unsupervised machine learning is used in the K-Means clustering algorithm. Unsupervised algorithms lack labeled data or a ground truth value to compare their performance against. Arrange the data into clusters that are more similar is the fundamental concept behind K-Means clustering. Each cluster is represented by its center (i.e., centroid) in K-Means clustering, which corresponds to the mean of the observation values given to the cluster [17].

To implement K-Means clustering includes these steps:

S1: In K-Means clustering firstly we have to specify the number of clusters k.

S2: Initialize centroids by first shuffling the dataset and then randomly selecting k data points for the centroids without replacement.

S3: Keeps iterating until there is no change to the centroids, i.e assignment of data points to clusters is not changing.

S4: Computes the sum of the squared distance between data points and all centroids.

S5: Assign each data point to the closest cluster (centroid).

S6: Compute the centroids for the clusters by taking the average of all data points that belong to each cluster.

The loss function:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2 \tag{1}$$

Where:

E: SSE of all of the data points

k: number of clustering

x: data point

x_i: mean of C_i

Euclidean distance formula:

$$d(\mathbf{x}_i, \mathbf{y}_i) = \left[\sum_{i=1}^{n} (x_i - \mathbf{y}_i)^2\right]^{\frac{1}{2}} \tag{2}$$

4.3 Standardization

Standardization [18] is another scaling technique that centers the results on the mean with a single standard deviation. This signifies that the attribute's mean becomes zero and the resulting distribution has a standard deviation of one. Z-score is defined by the below formula:

$$Z = \frac{x - \mu}{\sigma} \tag{3}$$

Where:

x: the previous value

μ: sample mean

σ: sample standard deviation

4.4 Elbow methodology

The Elbow method [19] is one of the most popular methods used to select the optimal number of clusters. The Elbow method is illustrated as a curve graph with the horizontal axis as the number of k clusters, and the vertical axis as SSE (Sum of Squared Error). When data points or observations are close together, they will have similar characteristics which are classified in a cluster.

The optimal k is the point at which the SSE starts to decrease steadily, the "elbow" position will be the best k number for the algorithm.

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_j} d\left(x_{ij}, m_i\right)^2 \tag{4}$$

Where:

x: data points in cluster i

m: mean value cluster i

k: number of clusters

4.5 Silhouette Score

Silhouette score [20] is also recognized as one of the most popular methods for determining the ideal k for clustering issues by analyzing the distances between clusters. The Silhouette score will indicate which data points or observations are inside the cluster (good) or close to the edge of the cluster (not good) to evaluate the clustering efficiency. Silhouette score has the value in range [-1; +1]. The silhouette score formula is below:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{5}$$

Where:

$s(i)$: silhouette score

$a(i)$: the average distance between data point i and the others in the same cluster.

$b(i)$: the average distance between data point i with respect to all other clusters.

4.6 Calinski Harabasz Index

The Calinski Harabasz Index [21] also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and the within-cluster dispersion for all clusters, the higher the score, the better the performances. CHI is computed for a collection of k clusters as follows:

$$CHI = \frac{T_r\left(B_k\right)}{T_r\left(W_k\right)} \times \frac{N - k}{k - 1} \tag{6}$$

Where:

N: is the number of points in our data; k is the number of the cluster

T_r: represents dispersion matrix

B_k: is the between-group dispersion matrix

W_k: is the within-cluster dispersion matrix

B_k and W_k are defined by the following equations:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{7}$$

$$B_k = \sum_{q}^{k} n_q (c_q - c)(c_q - c)^T \tag{8}$$

Where:

C_q: is the set of points in the cluster q

c_q: is the center of the cluster q

c: is the center of the whole data set which has been clustered into k clusters

n_q: is the number of points in the cluster q

4.7 Principal component analysis

Principal Component Analysis [22] is a well-known approach for dimension reduction, feature extraction, and data visualization. PCA is used to find patterns in data and present the data in a way that highlights similarities and contrasts. While reducing the dimension, it simultaneously improves interpretability. It makes data easier to plot in 2D and 3D and aids in identifying the dataset's most important properties. Finding a series of linear combinations of variables is made easier by PCA. This method also solves the problem of correlation among the variables.

4.8 Autoencoder

Artificial Neural Networks (ANN) are brain-inspired algorithms that are used to foresee problems and model complex patterns. The idea of biological neural networks in the human brain gave rise to the ANN, a deep learning technique. An effort to simulate how the human brain functions led to the creation of ANN. Although they are not exactly the same, the operations of ANN and biological neural networks are very similar. Only structured and numeric data are accepted by the ANN algorithm.

Autoencoder is a type of Artificial Neural Networks that is used to perform data encoding or representation learning. The concept originated in the 1980s and was subsequently popularized in a significant study published in 2006 by Hinton and Salakhutdinov, see [15].

It consist two parts:

- Encoder: is a set of convolutional blocks that compress the high dimension input to the model into a lower-dimensional space.

- Decoder: is a set of upsampling and convolutional blocks that reconstructs from the compressed version provided by the encoder

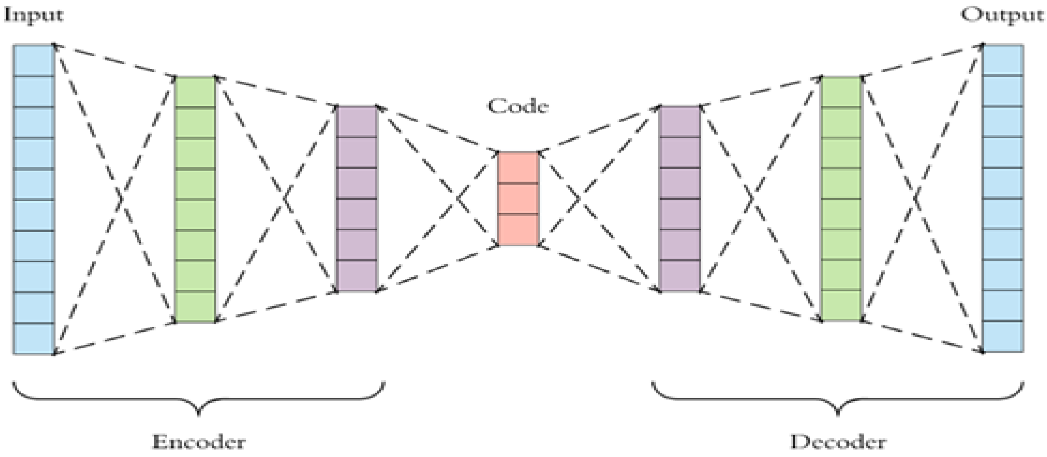


Figure 4.1: Schema of Autoencoder architecture

Compared to principal component analysis, PCA essentially only conducts linear dimensionality reductions, complete autoencoders can perform large-scale non-linear dimensionality reductions [5] [23].

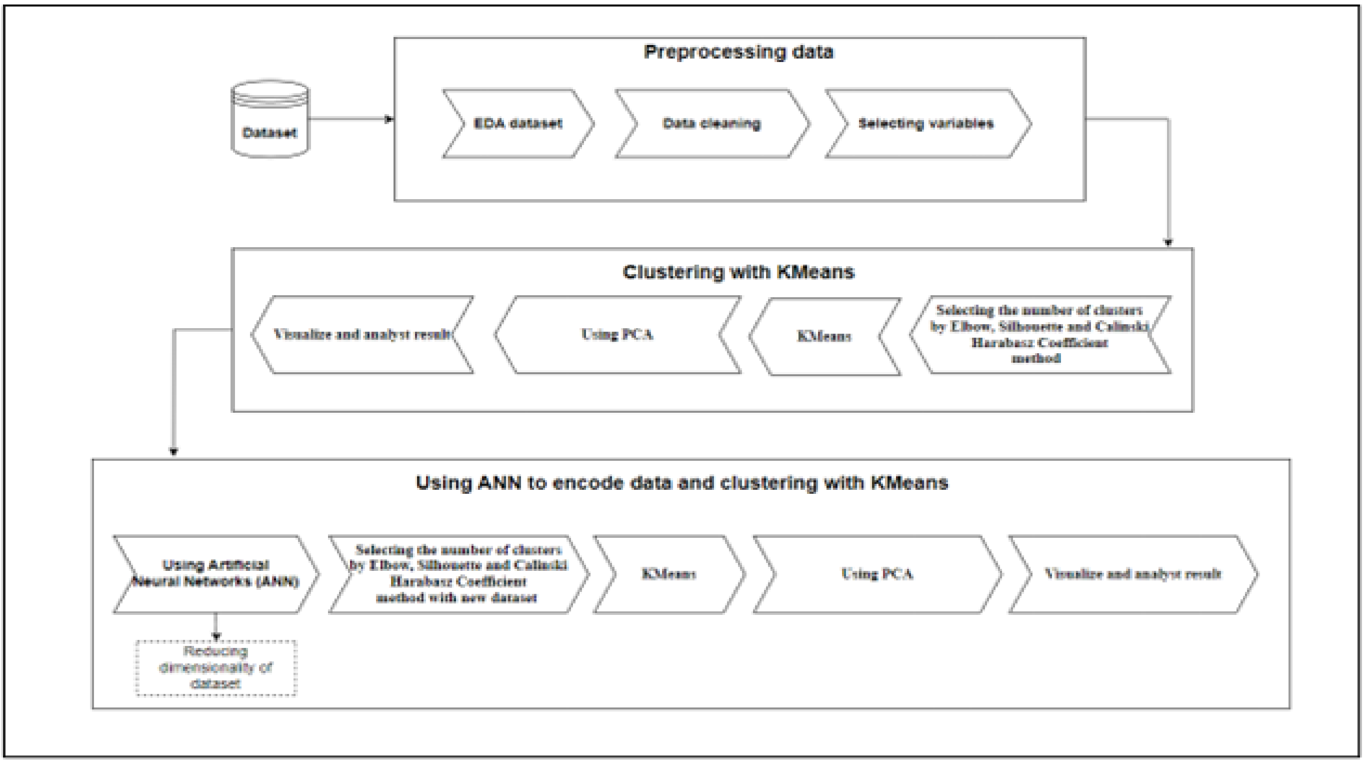

Preprocessing data
Dataset
EDA dataset
Data cleaning
Selecting variables
Clustering with KMeans
Visualize and analyst result
Using PCA
KMeans
Selecting the number of clusters by Elbow, Silhouette and Calinski Harabasz Coefficient method
Using ANN to encode data and clustering with KMeans
Using Artificial Neural Networks (ANN)
Reducing dimensionality of dataset
Selecting the number of clusters by Elbow, Silhouette and Calinski Harabasz Coefficient method with new dataset
KMeans
Using PCA
Visualize and analyst result

Figure 5.2: Methodology

As shown in Figure X, the research process involves three main stages:

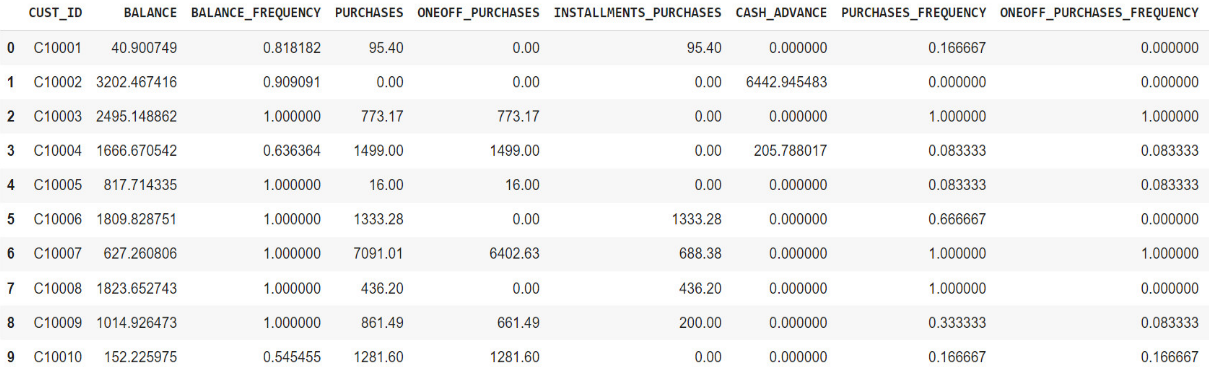
1. *Preprocessing data:* After collecting data from the Kaggle website, we started to explore the data and process the unnecessary data before applying machine learning and deep learning methods.

2. *Clustering with K-Means:* At this step, to be able to perform clustering by K-Means, we used 3 methods, Elbow, Silhouette and Calinski Harabasz Coefficient to choose the optimal number of clusters - k. Next, the K-Means machine learning method is applied to cluster the processed data files into K subgroups. We used PCA machine learning to reduce the data dimension and visualize the clusters to make the results more intuitive.

3. *Using ANN to encode data and clustering with K-Means:* After having the results from clustering with K-Means, we use the processed dataset and apply Deep learning ANN to get a new dataset with a reduced number of attributes. Next, we repeat the work in step 2 and make an assessment and compare the results from the two cases.

5.1 Preprocessing data

The dataset we use is “Customer Credit Card Dataset” it describes information related to the user’s credit and transactions within a year. It contains 8950 rows of 8950 customers.



	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY
0	C10001	40.900749	0.818182	95.40	0.00	95.40	0.000000	0.166667	0.000000
1	C10002	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	0.000000	0.000000
2	C10003	2495.148862	1.000000	773.17	773.17	0.00	0.000000	1.000000	1.000000
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.00	205.788017	0.083333	0.083333
4	C10005	817.714335	1.000000	16.00	16.00	0.00	0.000000	0.083333	0.083333
5	C10006	1809.828751	1.000000	1333.28	0.00	1333.28	0.000000	0.666667	0.000000
6	C10007	627.260806	1.000000	7091.01	6402.63	688.38	0.000000	1.000000	1.000000
7	C10008	1823.652743	1.000000	436.20	0.00	436.20	0.000000	1.000000	0.000000
8	C10009	1014.926473	1.000000	861.49	661.49	200.00	0.000000	0.333333	0.083333
9	C10010	152.225975	0.545455	1281.60	1281.60	0.00	0.000000	0.166667	0.166667

Figure 5.3: A part of the dataset

Following is the Data Dictionary for the Credit Card dataset:

Table 5.1: Description of data's variables

No.	Variables	Description
1	CUST_ID	Credit card holder ID
2	BALANCE	Monthly average balance (based on daily balance averages)
3	BALANCE_FREQUENCY	Ratio of the last 12 months with balance
4	PURCHASES	Total purchase amount spent during last 12 months
5	ONEOFF_PURCHASES	Total amount of one-off purchases
6	INSTALLMENTS_PURCHASES	Total amount of installment purchases
7	CASH_ADVANCE	Total cash-advance amount
8	PURCHASES_FREQUENCY	Frequency of purchases (Percent of months with at least one purchase)
9	ONEOFF_PURCHASES_FREQUENCY	Frequency of one-off-purchases
10	PURCHASES_INSTALLMENTS_FREQUENCY	Frequency of installment purchases
11	CASH_ADVANCE_FREQUENCY	Cash-Advance frequency
12	CASH_ADVANCE_TRX	Average amount per cash-advance transaction
13	PURCHASES_TRX	Amount per purchase transaction
14	CREDIT_LIMIT	Credit limit
15	PAYMENTS	Total payments in the period
16	MINIMUM_PAYMENTS	Total minimum payments due in the period
17	PRC_FULL_PAYMENT	Percentage of months with full payment of the due statement balance
18	TENURE	Number of months as a customer

First, checking for duplicates and missing values shows that the data has no duplicates, but there are many null data. We replaced 313 null values in the MINIMUM_PAYMENTS column and 1 null value in the CREDIT_LIMIT column with the average value of that column. With outlier data, we had a problem that there were quite a few outlines in the dataset. An effective method to deal with this problem and also applied in the project is KNN.

Realizing that all columns have their own meaning, are useful in customer clustering, and all have a numeric data type, we select all columns except CUST_ID column. For more accurate results, the data will be scaled to the range from 0 to 1 before performing the next steps.

5.2 Clustering with K-Means

After the dataset has been processed in step 1, we proceed to select the number of clusters based on 3 indexes: Elbow, Silhouette and The Calinski–Harabasz. To make the results more objective for the study, we will choose the common results of 2 or more methods. The results after each implementation are as follows:

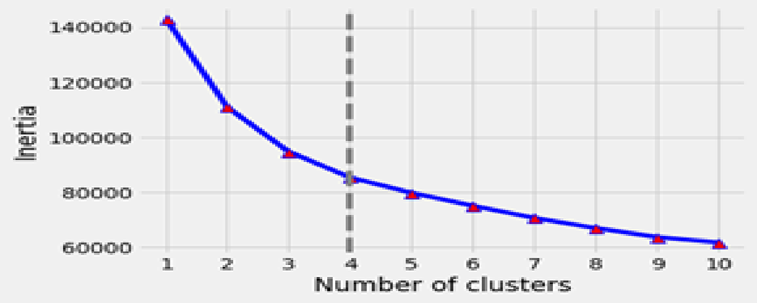


Figure 5.4: Result of Elbow method

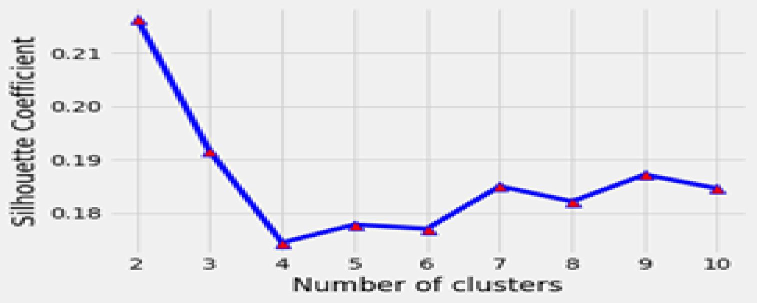


Figure 5.5: Result of Silhouette method

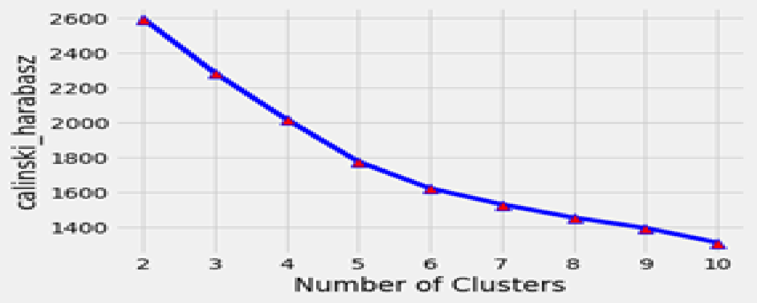


Figure 5.6: Result of Calinski–Harabasz method

The results show that the two methods Sihouette and Calinski–Harabasz say that the number of clusters k=3 is the most optimal. Therefore, we have chosen the number of clusters to be 3 to perform clustering using K-Means. After the session of K-Means, the result is that the customer has been divided into 3 clusters with the number of each cluster as follows:

Table 5.2: Number of each cluster

Cluster	Number of customers
0	2559
1	2569
2	3819

We use PCA to reduce the result dimension for the purpose of visualizing the results. The following is a part of the result:

Table 5.3: A part of data after using PCA

	pca1	pca2
0	-1.673564	-1.100200
1	-1.167053	2.530958
2	0.962476	-0.373595
3	-0.878599	0.035371
4	-1.595336	-0.709113

This is the visualization result after using PCA to reduce the data dimension. Figure 5.7 shows the distribution of data points in the 3 clusters.

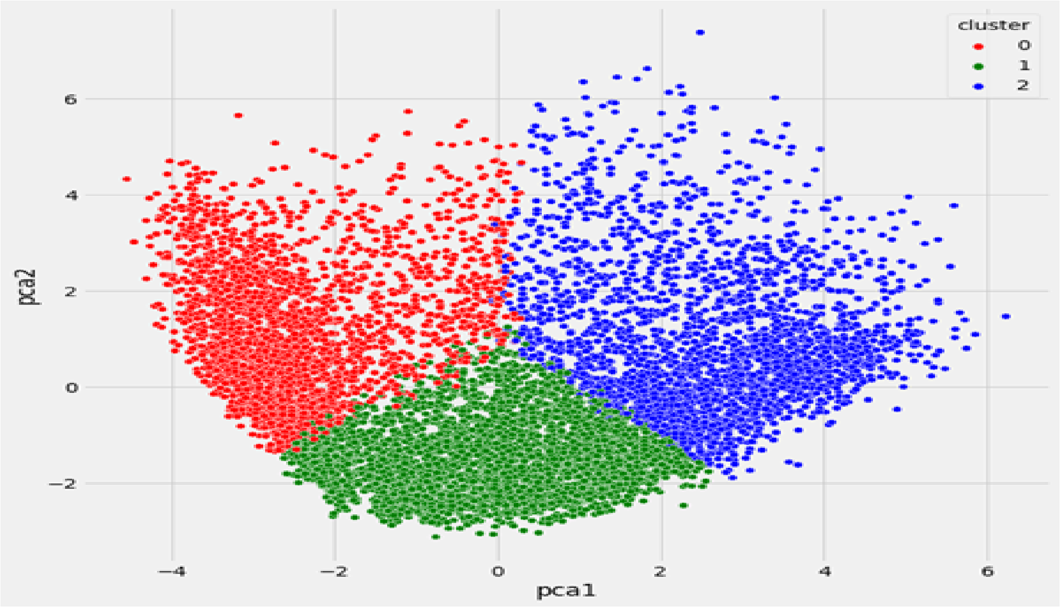
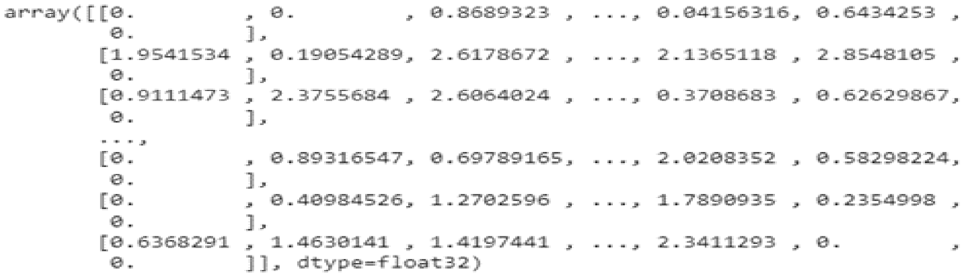


Figure 5.7: Distribution of 3 clusters on 2D space

5.3 Using ANN to encode data and clustering with K-Means

In this step, the ANN method is used as a tool to reduce the dimensionality of the processed data set. As a result, the number of columns of the dataset is reduced by 7 columns, to 10 columns compared to the previous 17 columns.



```
array([[0.        , 0.        , 0.8689323 , ..., 0.04156316, 0.6434253 ,
        0.        ],
       [1.9541534 , 0.19054289, 2.6178672 , ..., 2.1365118 , 2.8548105 ,
        0.        ],
       [0.9111473 , 2.3755684 , 2.6064024 , ..., 0.3708683 , 0.62629867,
        0.        ],
       ...,
       [0.        , 0.89316547, 0.69789165, ..., 2.0208352 , 0.58298224,
        0.        ],
       [0.        , 0.40984526, 1.2702596 , ..., 1.7890935 , 0.2354998 ,
        0.        ],
       [0.6368291 , 1.4630141 , 1.4197441 , ..., 2.3411293 , 0.        ,
        0.        ]], dtype=float32)
```

Figure 5.8: Part of the data after applying ANN

After acquiring new data, we repeat step 2 with this dataset. Cluster selection methods are applied and have the following results:

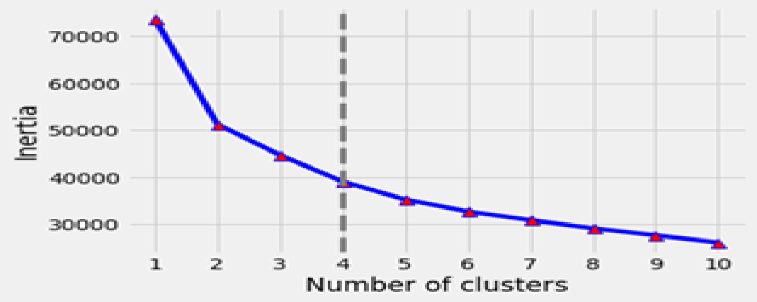


Figure 5.9: Result of Elbow method after applying ANN

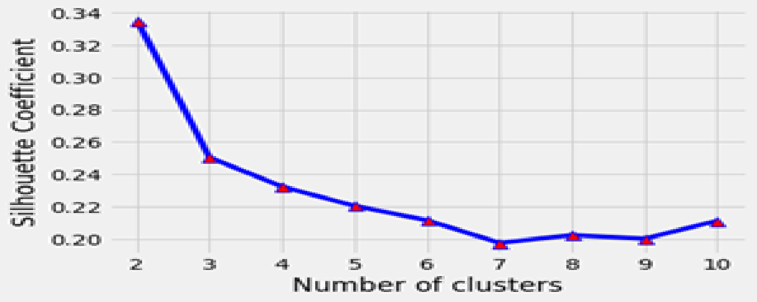


Figure 5.10: Result of Silhouette method after applying ANN

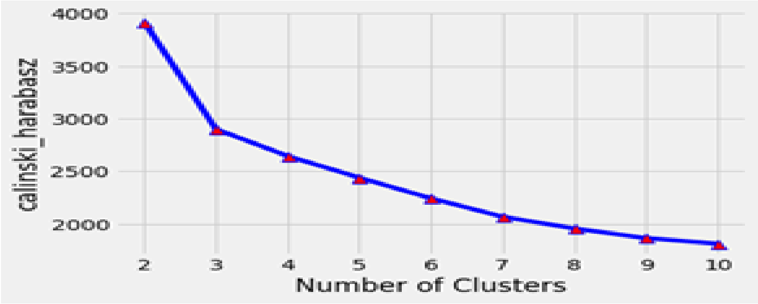


Figure 5.11: Result of Calinski–Harabasz method after applying ANN

There are two methods, Sihouette and Calinski–Harabasz that give k=3, so we choose the number of clusters k=3. With k=3 the K-Means method results in the number of clusters as follows:

Table 5.4: Number of each cluster after applying ANN

Cluster	Number of customers
0	2245
1	1937
2	4765

The figure below is the distribution of data points after using PCA machine learning to reduce the dimensionality:

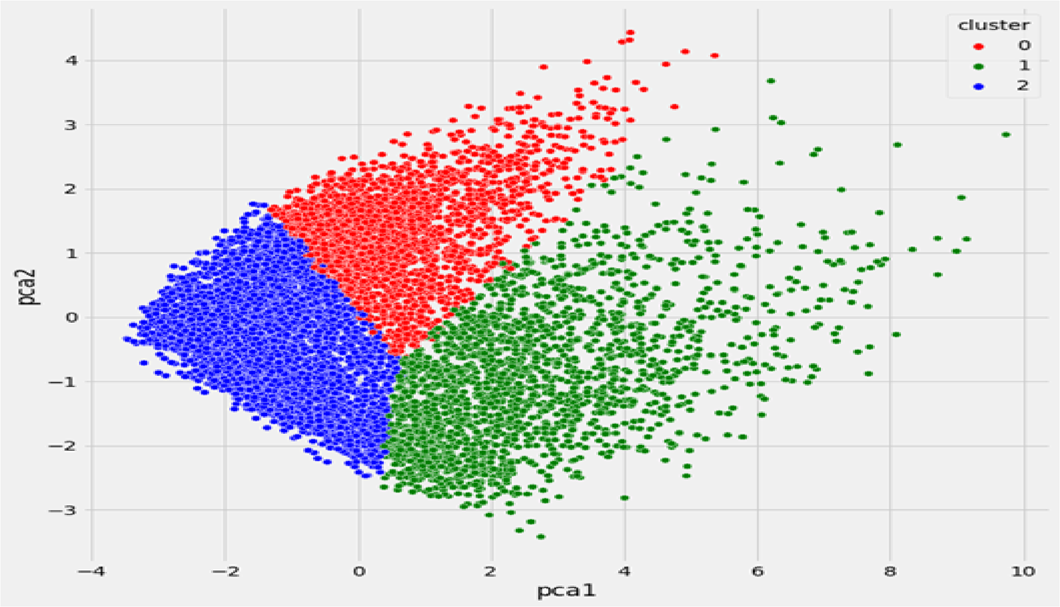


Figure 5.12: Distribution of 3 clusters on 2D space after using ANN

The clustering result shows that there are three clusters for both using ANN and without using ANN.

The chart below illustrates 3 segments clustered by K-Means without using ANN for dimension reduction. Each customer will have different motivations regarding purchasing decisions, and customer behavior refers to how and why they make a decision. So, out of 17 variables, we selected 5 that we found to be the most important to customers using a credit card to analyze their behavior: BALANCE, PURCHASES, CREDIT_LIMIT, PAYMENTS, ONEOFF_PURCHASES. These are important data that can help us predict why customers use credit cards to make decisions. For example, the more customers buy, the more likely they are to make larger purchases, or customers with higher credit balances are more likely to have higher credit limits and also more cash advances. or customers who buy more also make more payments.

The result with the combination between K-Means and ANN:

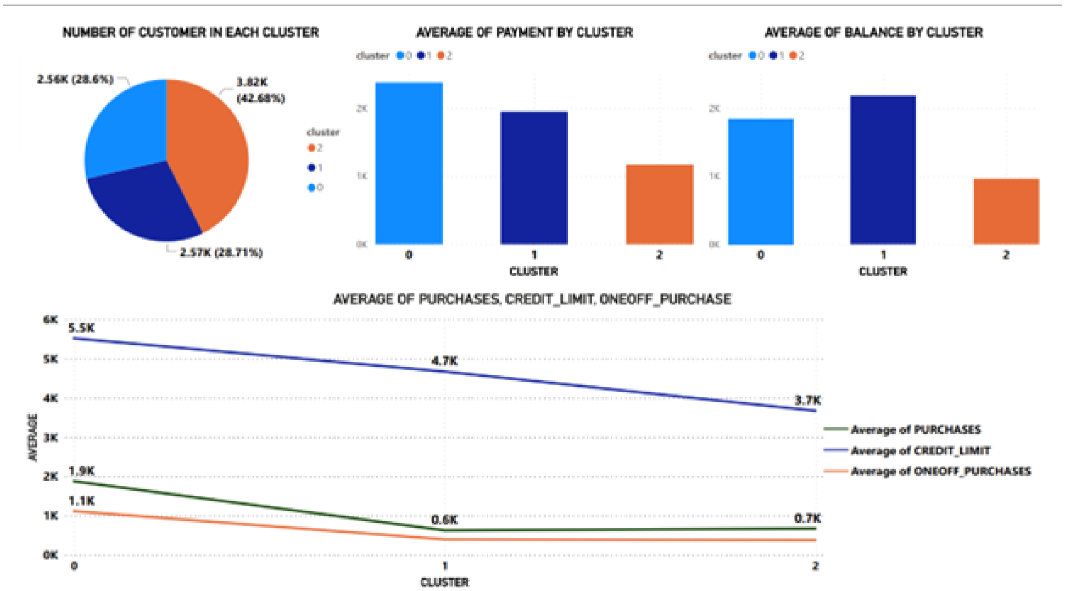


Figure 6.13: Clustering result without using ANN

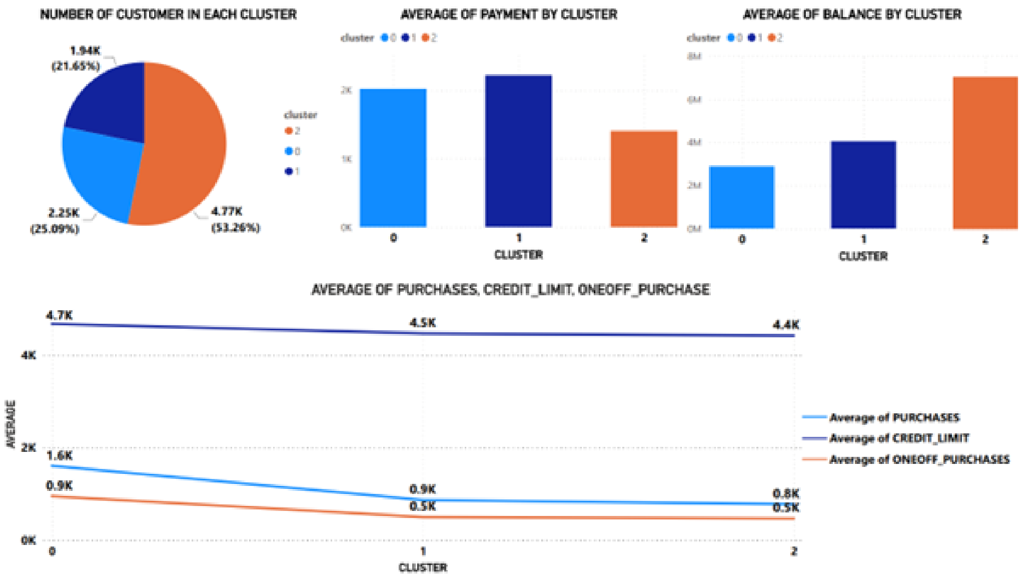


Figure 6.14: Clustering result using ANN and K-Means

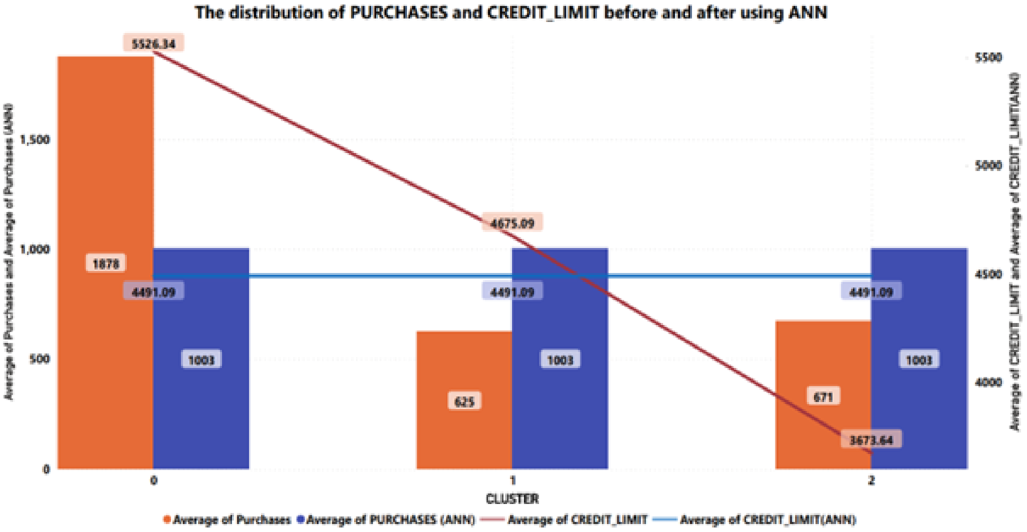


Figure 6.15: The distribution of PURCHASES and CREDIT_LIMIT after using ANN

These two graphs describe the correlation between PURCHASES and the average of CREDIT_LIMIT before and after using ANN. It can be clearly seen that the fluctuation of the two lines is proportional to each other after using ANN, which is more reasonable. For example, in the first graph, cluster 1 with CREDIT_LIMIT is 4675,09 which is higher than that of cluster 2 (3673,64) but the PURCHASES in cluster 1 (624,79) are lower than that of cluster 2 (670,81). Because customers with high purchase demand will increase their credit limit, it is unreasonable with the result in the first graph (without ANN).

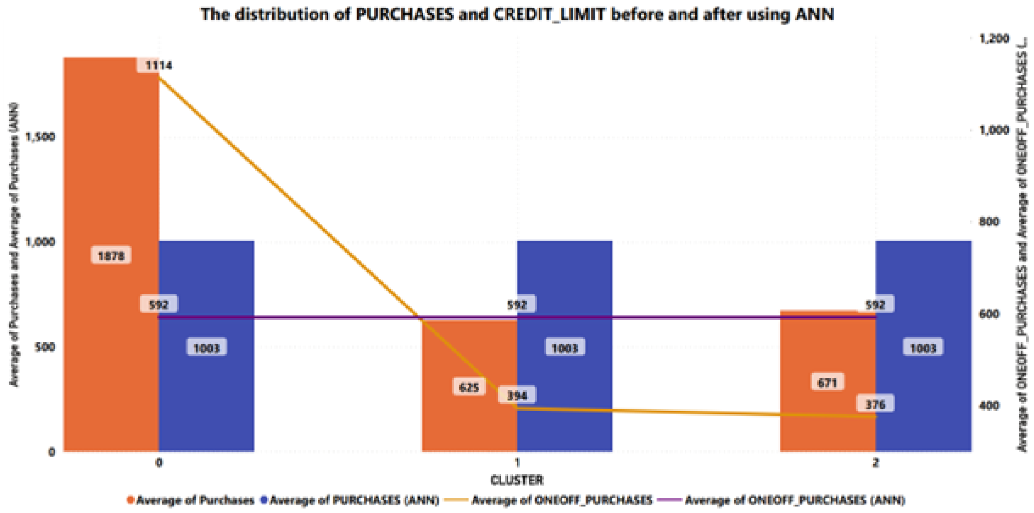


Figure 6.16: The distribution of ONEOFF_PURCHASES and PURCHASES after using ANN

The same as PURCHASES and CREDIT_LIMIT, in the relationship between ONEOFF_PURCHASES and PURCHASES, there are dramatical differences between clustered without ANN and with ANN. The ONEOFF_PURCHASES and PURCHASES are proportional to each other, therefore, in graph 1 (without using ANN), customers of cluster 1 have lower ONEOFF_PURCHASES (624.79) than cluster 2 (670.81) but have higher PURCHASES than cluster 1 (394.47 > 376.06). This result shows extreme illogicality.

With the two illustrations above, it is obvious that K-Means with ANN will provide better results compared to those without using ANN. Our team just performs clustering on a dataset of 8000 rows. If we continue to increase the scale of data, the problems mentioned will be much bigger. Therefore, clusters with ANN are chosen for segmentation labeling.

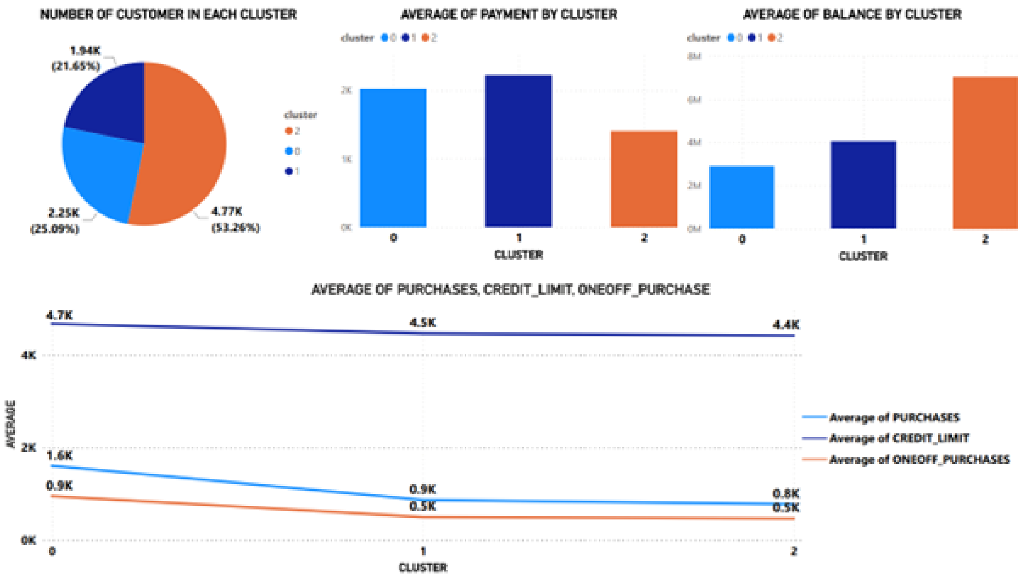


Figure 6.17: Clustering result using ANN and K-Means

Cluster 0 - ***Willing-to-pay customer:*** This cluster involves 2245 customers with the highest credit limit, high purchases, and high payment. These are the most valuable customers whom credit card providers can entice to increase their spending amount by enhancing their credit limits.

Cluster 1 - ***Thrifty customer:*** This cluster includes 1937 customers with a high credit limit, the highest balance, and the highest payment but their amount of purchases is a bit low. Businesses can encourage them to simulate their activities by providing cash back, promotions, etc. These customers are potential ones who can bring back much revenue for the business.

Cluster 2 - ***New customer:*** This cluster accounts for 4765 customers, which is the biggest cluster. In this cluster, they have the lowest purchases and payments but with a medium balance. Most of the customers in this cluster are new customers who have just opened credit accounts and made small transactions, not spending much money on making purchases. The credit card providers should have preferences policies for new customers and introduce to them more special offers.

CHAPTER 7. CONCLUSION

Realizing the importance of customer segmentation for businesses and the difficulty of businesses when the data is too many dimensions, a customer segmentation model combining ANN and K-Means method has been proposed. The model is tested with a dataset of customer information and transactions on credit cards. ANN method is significant in reducing the data dimension and K-Means has a key role in clustering customers into small clusters. ANN method is significant in reducing the data dimension and K-Means has a key role in clustering customers into small clusters. The use of ANN before clustering positively changes the results, the difference can be easily seen when using a combination of PCA method and visualization plot types. For businesses, this proposed model can be applied easily and quickly. Businesses no longer have to have a hard time choosing which variables to use for customer segmentation, the collected data will make the most of it, so there will be many improvements in business results.

Overall, the study has solved the problem posed. However, the data used for the model experiment is not satisfactory because of the small amount of data and only within 1 year, so the obtained results are not really superior to the traditional model. In addition, there are some properties in the 3 obtained clusters that are not too different from each other such as CREDIT_LIMT variables. In the future, in order to make the model more accurate and reliable, we propose to experiment on datasets with longer collection times and large enough data sizes.

REFERENCES

[1] E. Page, "Adaptive control processes: A guided tour," 1962.

[2] S. Yoo, J. Song, and O. Jeong, "Social media contents based sentiment analysis and prediction system," *Expert Systems with Applications*, vol. 105, pp. 102–111, 2018.

[3] M. Alkhayrat, M. Aljnidi, and K. Aljoumaa, "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca," *Journal of Big Data*, vol. 7, no. 1, pp. 1–23, 2020.

[4] C. Ding, *Dimension Reduction Techniques for Clustering.* Boston, MA: Springer US, 2009, pp. 846–846. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_612

[5] S. P. Nguyen, "Deep customer segmentation with applications to a vietnamese supermarkets' data," *Soft Computing*, vol. 25, no. 12, pp. 7785–7793, 2021.

[6] A. Palmer, "The evolution of an idea: an environmental explanation of relationship marketing," *Journal of Relationship Marketing*, vol. 1, no. 1, pp. 79–94, 2002.

[7] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using k-means clustering," in *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS).* IEEE, 2018, pp. 135–139.

[8] C. P. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of k-means algorithm for efficient customer segmentation: a strategy for targeted customer services," 2015.

[9] A. Aziz, "Customer segmentation basedon behavioural data in e-marketplace," 2017.

[10] J. Wu, L. Shi, W.-P. Lin, S.-B. Tsai, Y. Li, L. Yang, and G. Xu, "An empirical study on customer segmentation by purchase behaviors using a rfm model and k-means algorithm," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[11] W. Qadadeh and S. Abdallah, "Customers segmentation in the insurance company (tic) dataset," *Procedia computer science*, vol. 144, pp. 277–290, 2018.

[12] A. Abdulhafedh, "Incorporating k-means, hierarchical clustering and pca in customer segmentation," *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021.

[13] W. Guo, K. Lin, and W. Ye, "Deep embedded k-means clustering," in *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2021, pp. 686–694.

[14] P. Huang, Y. Huang, W. Wang, and L. Wang, "Deep embedding network for clustering," in *2014 22nd International conference on pattern recognition*. IEEE, 2014, pp. 1532–1537.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[16] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[17] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[18] I. B. Mohamad and D. Usman, "Standardization and its effects on k-means clustering algorithm," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, no. 17, pp. 3299–3303, 2013.

[19] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336, no. 1. IOP Publishing, 2018, p. 012017.

[20] G. Ogbuabor and F. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," *International Journal of Computer Science & Information Technology*, vol. 102, no. 2018, pp. 27–37, 2018.

[21] J. Li, D. Hassan, S. Brewer, and R. Sitzenfrei, "Is clustering time-series water depth useful? an exploratory study for flooding detection in urban drainage systems," *Water*, vol. 12, no. 9, p. 2433, 2020.

[22] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.

[23] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.