SEGMENTATION FOR FACTORING CUSTOMERS USING UNSUPERVISED MACHINE LEARNING ALGORITHMS

NUR SEHER AYYILDIZ

MEF UNIVERSITY
JUNE 2023

MEF UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MASTERS'S IN INFORMATION TECHNOLOGIES

M. Sc. THESIS

SEGMENTTAION FOR FACTORING CUSTOMERS USING UNSUPERVISED MACHINE LEARNING ALGORITHMS

Nur Seher AYYILDIZ

ORCID No: 0009-0003-8644-8550

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

JUNE 2023

ACADEMIC HONESTY PLEDGE

I declare that all the information in this study is collected and presented in accordance with academic rules and ethical principles, and that all information and documents that are not original in the study are referenced in accordance with the citation standards, within the framework required by the rules and principles.

Name Surname: Nur Seher AYYILDIZ

Signature:

ABSTRACT

SEGMENTATION FOR FACTORING CUSTOMERS
USING UNSUPERVISED MACHINE LEARNING ALGORITHMS

Nur Seher AYYILDIZ

M.Sc. in Information Technologies

Thesis Advisor: Asst. Prof. Dr. Tuna ÇAKAR

June 2023, 80 Pages

Nowadays the fact that technology facilitates data collection is an important opportunity, as well as making the management of all this data difficult and makes no sense unless it is well processed. This stored data is extremely important, and companies use data provided by their customers. Catching the needs of the customer profiles of the changing world is now a necessity and takes the first place for companies. With the increase in the amount of stored data over time, it has become difficult to establish a relationship between the data and to separate them from each other. At this point, machine learning methods have become more involved in our lives.

In this study, what segmentation is and its change over the years are mentioned. It has been mentioned which machine learning techniques will be useful in data selection. Then, possible machine learning methods are shown using the local factoring company's customer check data.

Since this study aims to group unlabeled data, unsupervised learning techniques are emphasized. Among these methods, Hierarchical Clustering, DBSCAN, Gaussian Mixed Modeling methods, Fuzzy c - Means were used besides the most popular K-Means. The success criteria for each algorithm were examined and the appropriate cluster numbers were found, and the results were measured.

When the clustering outcomes were examined, the optimal number of clusters was calculated very high with GMM, DBSCAN could not assign clusters, and Hierarchical clustering has been found to be very costly in terms of time. It was observed that the best results were obtained with the K - Means and FCM.

Keywords: Customer Segmentation, Clustering Algorithms, Factoring Customers, Machine Learning, Segmentation Model

Numeric Code of the Field: 92404

ÖZET

GÖZETİMSİZ MAKİNE ÖĞRENMESİ ALGORİTMALARI
KULLANILARAK FAKTÖRİNG MÜŞTERİLERİ İÇİN SEGMENTASYON
YAPILMASI

Nur Seher AYYILDIZ

Bilişim Teknolojileri Tezli Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Tuna ÇAKAR

Haziran 2023, 80 Sayfa

Günümüzde teknolojinin veri toplamayı kolaylaştırmasının önemli bir fırsat olmasının yanı sıra tüm bu verilerin yönetimini zorlaştırmakta ve veriler iyi işlenmedikçe bir anlam ifade etmemektedir. Depolanan bu veriler son derece önemlidir ve şirketler, müşterileri tarafından sağlanan verileri kullanır. Değişen dünyanın müşteri profillerinin ihtiyaçlarını yakalamak artık bir zorunluluk haline gelmekte ve firmalar için ilk sırayı almaktadır. Zamanla depolanan verinin artması ile artık veriler arasında ilişki kurmak ve bunları birbirinden ayırmak zor bir hal almıştır. Bu noktada hayatımıza makine öğrenmesi yöntemleri daha fazla dahil olmaya başlamıştır.

Bu çalışmada, segmentasyonun ne olduğu ve yıllar içindeki değişiminden bahsedilmiştir. Hangi makine öğrenmesi tekniklerinin veri seçiminde faydalı olacağına değinilmiştir. Ardından olası makine öğrenmesi yöntemleri yerel bir faktoring şirketinin müşteri çek verileri kullanılarak gösterilmiştir.

Bu çalışma etiketsiz verilerin gruplanmasını hedeflediğinden gözetimsiz öğrenme teknikleri üzerinde durulmuştur. Bu yöntemler arasında en popular olan K – means algoritmasının yanı sıra Hiyerarşik Kümeleme, DBSCAN, Gauss Karışık Modelleme ve Fuzzy c - Means yöntemleri kullanılmıştır. Her bir algoritma için başarı ölçütleri incelenerek uygun küme sayıları bulunmuş ve bulunan sonuçlar karşılaştırılmıştır.

Kümeleme sonuçları incelendiğinde GMM ile optimal küme sayısı oldukça yüksek hesaplanmış, DBSCAN küme atayamamış, Hierarchical clustering ise zaman açısından maliyetli bulunmuştur. En iyi sonuçların K - means ve Fuzzy c - Means algoritmalarıyla elde edildiği gözlemlenmiştir.

Anahtar Kelimeler: Müşteri Segmentasyonu, Kümeleme Algoritmaları, Faktoring Müşterileri, Makine Öğrenmesi, Segmentasyon Modeli

Bilim Dalı Sayısal Kodu: 92404

Tüm süreç boyunca bana destek olan

eşime, aileme ve arkadaşlarıma ithaf ediyorum.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

GMM	: Gaussian Mixture Model
ML	: Machine Learning
DBSCAN	: Density-Based Spatial Clustering of Applications with Noise
EM	: Expectation-Maximization
PCA	: Principal Component Analysis
T – SNE	: T-Distributed Stochastic Neighbor Embedding
BIC	: Bayesian Information Criterion
WCSS	: Within Clusters Sum of Square
CRM	: Customer Relationship Management
NBD	: Negative Binomial Distribution
FCM	: Fuzzy c – Means
ANOVA	: Analysis of Variance
KNN	: K Nearest Neighbors
LTV	: Lifetime Value
VSM	: Vector Space Model
CLV	: Customer Lifetime Value
RFM	: Recency, Frequency, Monetary
BA	: Business Analytics
BI	: Business Intelligence

INTRODUCTION

Estimation of recent requirements of customers are becoming harder with unstable factors in daily life and developing technology. Although today's technology gives opportunities to collect data, collecting is never worth it if it is not used since storing, managing, and protecting data is a big deal.

Understanding customer needs and predicting risks which could be caused by customers are two of the main topics of companies. If a company has known its customers, this is a perfect opportunity to find the ways to reach them and protect its own reputation.

Customer can be an individual person or a company who uses a product or a service provided by companies. Conceptually, even though a customer exists for a very long time it was not the focus point for producers or companies till 1990's. In the past companies were running against each other to produce new products, then they aimed to mass produce with low cost with the development of industry and technology. When each target reaches certain maturity new targets have emerged that need to be reached. After mass production activities, quality of the productions started to be the main topic and customer relations occurred in 1990's. Customer relationship is an important value for various sectors like finance, retail, health, telecommunication. In a recent study customer relationship management for the factoring sector will be discussed and segmentation models will be developed. Companies incur remarkable costs to achieve new customers and increase the pleasure of present customers for maximizing customer returns [1]. To achieve this goal understanding customers and separating them into meaningful groups and service them according to their needs is extremely important. For this reason, there are some customer metrics to analyze customers.

Although providing services according to customers' expectations allows meeting specific customer needs it is not possible while considering the time and costs. It is a preferred method to group the customers by taking into account their demographic information, psychological status and shopping characteristics and to provide services according to the determined groups. A 5 % increase in customer retention enhances benefits by 25 % to 95 % [2]. While being familiar with customer's behaviors and getting to know their habits is beneficial for the companies to reach and retain their customers, there is another highly important point that companies have the opportunity to predict their risks which may be caused by their customers.

Financial organizations especially have to overcome different types of risks and the customer risk is the one of them. Recent years with easier access to information, organizations develop their point of view while approaching the customers. Current study, factoring firm’s customer profiles will be examined therefore business to business (B2B) segmentation will be concerned.

Purpose of Thesis

This paper aims to find the most appropriate customer segmentation according to their credit limit and risk status and the checks they apply. To achieve this goal various unsupervised learning methods and data cleaning methods have been used. In addition to these, it has been tried to make the selection of variables in an optimal way by interviewing people who have sectoral knowledge, especially those who work in the credit departments of the banking sector. The data set subject to the study was obtained from a local factoring company and consists of 1103996 rows and 231 columns. In this study, it is desired to categorize the customers whose applications are processed by the company and the applications made to the company between 2021 and 2022 were examined with K - Means, Fuzzy c – Means (FCM), Gaussian Mixture Model (GMM), DBSCAN, Hierarchical Clustering and a clustering study was carried out. Then, these methods were compared, and it was determined which application produced more efficient results in such a data set.

Literature Review

Segmentation in terms of data science may be done using statistical methods and ML algorithms. In this study ML algorithms will be applied for clustering customer data and statistical methods will be used for data preprocessing, feature selection and defining success rate. Many impressive clustering algorithms have been found over the past few decades, such as K - means, FCM, DBSCAN, Hierarchical Clustering and more [3].

In 2012, business intelligence analytics and development were reviewed by Hsinchun Chen, Roger H. L. Chian, and Veda C. Store. The field of business intelligence and business analytics (BI&BA) and big data analytics has gained increasing importance and value in both the academic and business circles over the past two decades. This development has also been highlighted by industry research. According to the IBM Tech Trends Report published in 2011, more than 4000 information technology experts from 93 countries and 25 industries were surveyed. According to this survey, it was emphasized that business analytics will be one of the four main technologies in the 2010s. According to the BA survey carried out by Bloomberg Business in 2011, it was determined that 97 % of companies with incomes exceeding $100 million use some form of BA. [4]

According to a report prepared by the McKinsey Global Institute, it is predicted that only in 2018, there will be a lack of 140,000 to 190,000 people with vast analytical knowledge, and there will be a shortage of managers with approximately 1.5 million data insights and big data knowledge to make effective decisions. [5]

In a study published in the Marketing Science Journal in 1995 declared that there was research for customer selection to do direct mail campaigns and historically as it is said RFM method was the most frequently used method. In the declared study Recency measured the count of unanswered consecutive mails and the passed time since the last order. Frequency meant that the count of orders in a specific term and monetary was the paid cash in that period. In the RFM method each variable has several values that are decided by a researcher and probabilities assigned to each of them. This method has some disadvantages like it has 3 measure elements but for customers response behaviors there are more than recency, frequency and monetary to be effective [6].

In the European Journal of Operational Research market segmentation study with K - means and Frequency-Sensitive Competitive Learning Algorithm (FSCL) has been declared in 1996. Nonhierarchical K - means method has been applied in SAS application. For segmentation study 207 data which belongs to 18 coffee brands was used to find customers brand changing rates. K value was set as 6 and by the end of the 11th iteration the algorithm had finished its work. FSCL and K - means have produced different results and they have looked for the best method. According to the study [7] synthetic data had been produced. Then, in SAS they were compared by the Analysis of Variance (ANOVA) method. Finally, results found with the FSCL method were used as initial seeds for the K - means and it was proven that combining algorithms give better results and selecting mean value closed initial points gives better results in K - means algorithm [8].

In 1995, P. Bradley, U. Fayyad stated that they could identify better starting points for the application of K - means based on their work. They called this method constrained K - means. Thus, they stated that with the improved starting points, the K - means algorithm will also work much faster with big data. They stated that the method they found is applicable to both discrete and continuous data sets. The optimization runtime is significantly shorter than the time needed to cluster the entire database. Their method is scalable and possible to combine with a scalable clustering algorithm to solve large-scale clustering issues in data mining [9].

In 1998, Zhexue Huang mentioned K - means efficiency with large data sets, unless they have categorical values. Since, in real life data sets contain numerical and categorical variables, some extensions were applied to K - means who handle categorical data sets and mixed type of datasets. They used K - modes and K - prototypes algorithm for this problem’s solution and used popular soybean disease and credit approval data sets. The K - modes algorithm mentioned that it handles categorical variables the same way with K - means. K - prototype algorithm was declared as a combination of K - means and K - modes [10].

In the study by Pelleg and Moore in 2000, a new method that generates solutions for K - means was developed and this is called X - means. K - means algorithm suffers with insufficient computation, user defined the cluster count, and the search sloped to the local minima problems. X - means proposes answers for the first two problems and fractional solutions for the last one. X - means uses statistical methods to make local decisions, and it maximizes posterior probabilities. Experimental results show that X - means performing faster and better than K - means on synthetic and real-life data [11].

An empirical study with widespread document clustering techniques was conducted by Steinbach, M., Karypis, G., and Kumar, V. in 2000. In this study, especially Agglomerative Hierarchical Clustering and K - means were compared. Hierarchical clustering is frequently referred to as a higher quality clustering method however is constrained by quadratic time complexity. K - means and alternatives have a time complexity that is linear in document count but is thought to generate smaller clusters. In some cases, better results can be obtained by using a combination of these two algorithms. In this study, a proposal is made for the results based on the analysis of the properties of the clustering algorithms and the nature of the document data. For the K - means, a standard K - means and a variant of K - means bisecting the K - means were used. Study outputs showed that the dividing K - means technique was better than the standard K - means approach and was nearly better than the hierarchical approaches exercised. Also, the running time of bisecting K - means is much more convenient to implement compared to agglomerative hierarchical clustering techniques. [12]

According to the study published in the journal Pattern Recognition in 2002, A. Likas, N. Vilasis, and J.J. Verbeek made suggestions that will enable better selection of the starting point for the K - means algorithm and minimize time and cost loss with less iteration. They named the method they found as Global K - means. The suggested clustering methods have been exercised on popular data sets. The method they found is independent of any initial condition. They compared the success of their work with random reboots with K - means and observed that the method was successful [13].

In 2003, a study by Hyunseok Hwang, Taesoo Jung and Euiho Suh studied customer segmentation according to customer value. The concept of relationship management in the field of marketing began to gain importance in the early 1980s. Among the main concerns of the companies is performing the right targeted campaigns to acquire the most profitable customers and then to retain these customers. It is important to learn about customer value in order to manage customer relationship management effectively. For this reason, a lot of research has been done to determine the customer lifetime value (LTV). In the study, a lifetime value model is proposed that takes into account a customer's past expedience contribution, potential advantage and probability of leaving. By analyzing the customer value, the segmentation study is handled over this customer value. Customer value is analyzed in three categories, these are current value, potential value, and customer loyalty. A case study was conducted to divide the customer value into segments by calculating these three customer values by randomly selecting 2000 pieces of 6-months 16,384 customer service data belonging to a wireless communication company in Korea. These 3 values are scaled in the range of 0 - 1 and their distribution is observed by drawing a 3-dimensional scatterplot. According to this distribution, customers are divided into 8 segments. As a result, it has been suggested that customers with high current value can be given coupons as a reward, customers with high potential value can use some services free of charge for a few months, and in cases where customer loyalty is low, continuity can be ensured by giving loyalty cards to these customers. [14]

There is another customer segmentation study published in 2012 which involves online store customers and, for this study psychographic data were used. Dataset for this study collected 196 online store customer surveys in Korea. The purpose of this study is to present new and is to provide a personalized marketing strategy for existing customers. This study has three phases. In the first phase, the factors affecting the customers' intention to buy from online stores were determined by Structural Equation Method (SEM). With this method statistically significant features were selected. In the second phase, the Self Organized Model(SOM) and K - means algorithm were applied for the factors selected in phase 1. If a customer was assigned to the same group for both models, then this group was determined for the customer. Else they were assigned a new cluster which is a combination of the results of two models. Three clusters (A, B,C) have been predefined for both models , after implementation five clusters(A, AB, B ,BC, C) have been achieved. Then, ANOVA test was performed to confirm the homogeneity of the groups and it was seen that the groups were significantly different. At the last phase, K Nearest Neighbor

Method implied for the customers which were not included in the clustering process. It has been stated that the K - means is not practical enough for very large data and why the KNN algorithm is needed. Evaluation for the KNN algorithm success test was performed by determining the k value between 1-10 and the most accurate result was obtained when k=1. With k = 1, the accuracy of the KNN approximation was calculated as 89.74%. This means that it is possible to identify segments with an error of about 10% without knowing about customers' purchase intention [15].

In 2013, a study was conducted by Soumi Ghosh and Sanjay Kumar Dubey comparing FCM and K - means algorithms. The behavior patterns of both algorithms were analyzed according to the count of data points and the count of clusters. In this study, the UCI Machine Learning Repository, which is frequently used by database researchers, was used. Algorithms were run in Mathlab with the iris dataset. In this study, a total of 5 features, 4 numerical and 1 non-numeric, were used for the Iris data set. The non-numeric attribute has 3 different values. With K - means, the data set could be divided into 3 clusters in 13 iterations and 0.443755 seconds. It could be divided into 3 clusters in 0.781679 seconds with 30 iterations with Fuzzy c - Means. Both algorithms were compared according to time complexity, it was seen that time complexity increased in both algorithms as the number of iterations increased, but this value was calculated higher for FCM. The FCM algorithm produced approximate results with K - means, but fuzzy logic calculation methods took more time. Considering these situations, the K - means algorithm was found to be more successful. [16]

Clustering analysis was implemented for the insurance sector in 2016. The data set used in clustering purposes belongs to one of Turkey's most important insurance companies and the data consists of individual policy information. During the data preparation phase private information was removed like identity number and policy number. For missing variables in the data set the Data Audit tool was used. After all these studies the dataset prepared consists of demographic information and has 3662 with 5 columns. Five columns refer to gender, resident city, job, age, and coverage. In this study K - means was chosen and was applied for 3 clusters. For the categorical values Chi Square test and the continuous variables t-test were implemented and important variables were specified. For each cluster important variables were different. Also, important variables were showing the rule set for each cluster [17].

In the study conducted in 2019, video-based image clustering was requested. In this study, several algorithms are tried and compared. These algorithms are Fuzzy c - Means, K - means and Agglomerative clustering. A methodology is presented for automatic grading of video frames based on a combination of SIFT features and clustering method. An assessment model based on Silhouette analysis and Adjusted Rand Index was used to define resemblance between generated tags (with their suggested method) and manually assign tags. A comparative analysis was made between the available methods and the presented methodology. In this study, Vector Space Model (VSM) was created and then clustering algorithms were applied. The success value of the clusters for each algorithm was calculated with the silhouette score. According to the calculated scores, the algorithms that provide effective clustering for the relevant data set were K - means, K - medoids, FCM and Agglomerative clustering, respectively [18].

Another study is related with customer segmentation and published in 2020, it supports e-retail customer segmentation research with RFM method. First of all, for each of the three variables, a score was made in the range of 1-5, with 5 for the data entering the first 20% slice and 1 for the data entering the last 20% slice. Then, these scores were combined and segments such as 515 were obtained. 515 refers to the customer group who shopped recently, has a high monetary value in the specified period, but has a low shopping frequency. According to this segmentation, the most important customer segment is 555. With this partitioning technique, 81 groups were obtained. Another approach, the method of dividing the data set into equal intervals, was also tried and 20 groups were obtained, so this method was not sufficiently interpretable. Both approaches have been tested with the Cluster Evaluation methods Silhouette Coefficient, Average Cohesion, Average Separation. Only with Average Separation, the equal range approach was successful. In this study, the best customers, the most valuable customers, churn customers and more were analyzed [19].

In 2020 there was another study related to the popular K - means algorithm, algorithm was applied to data which have mixed types of features. K - means [20], X - means [11], Constrained K - Means [9], K - prototype [10], and Kernel K - means [13] algorithm was applied to Wisconsin Diagnostic Breast cancer, KDD Cup 1999 (10%) and Epileptic Seizure datasets. Study compares the different K - means performances. As a result of this experimental analysis, it is declared that there is no generic resolution for the issues of K - means algorithms, and algorithms are specific for the applications and the data set [21].

Based on the 2018 order data of a pizza brand in 2021 customer clusters and behavioral patterns were determined. This data consists of 24 million rows and contains 52 variables as a result of data preprocessing. The three commonly used clustering algorithms in the litterateur were used. These are K - means, Gaussian Mixture and DBSCAN algorithms. Based on these clustering algorithms, Silhouette, Davies Bouldin and Calinski Harabasz indexes were tested to detect meaningful clusters. The data was divided into 4, 6 and 8 clusters with K - means, Gaussian Mixture algorithms and each index was examined, and it was seen that the most successful result was obtained with 4 clusters. Since the DBSCAN algorithm is unsuitable for the data set used in the experiment, it was not compared in the clustering analysis. The elbow method used to define appropriate cluster count for K - means was applied and again the number of clusters was found to be 4. It has been determined that 52 features in the data set do not make a difference between other clusters. For this reason, these data were removed from the data set and clustered again, and it was observed that the data set was divided into 3 basic clusters. Then, clusters with high data were divided into sub-clusters, common independent variables were determined, and multiple regression analysis was performed for each cluster. Then, behavioral rule sets were determined for each cluster [22].

In 2021, there is a study was promoted by the National Natural Science Foundation of China. GMM clustering has been extensively analyzed for its efficiency and effectiveness. Although this algorithm gives excellent results, it has been observed that it does not produce effective results for missing data. In this study, unlike other studies, firstly the absences in the data were eliminated and then the GMM algorithm was used.

Missing data were filled according to the result of the GMM algorithm, and then the algorithm was applied again. This two-step algorithm has been tested on 8 data sets and the accuracy of this study has been proven [23].

Although there are too many studies for segmentation there are not enough studies for the finance sector especially factoring companies. Therefore, current study will be the introduction for segmenting factoring customers.

Overview

The paper is organized as follows. First Chapter begins with background information about factoring sector information, the risks of factoring and the customer segmentation. The process flows of checks and the history of customer segmentation are explained. Chapter 2 presents details about machine learning algorithms, cluster counts and the ways to find success rate of the algorithms. Chapter 3 describes the factoring data set. Implementation details and test results of this work are presented in detail. In chapter 4, developed work is discussed and in final part conclusion is made and some future work is stated.

1. BACKGROUND

1.1 What is Factoring?

Factoring companies play an important role in providing working capital, especially for businesses that have difficulty in obtaining funds through banks [24]. Factoring is a financial term which refers to relation with debtor, factoring company, and customer. Customers refers to the seller who is expecting payment and debtor refers to the buyer who is expected to pay and the factoring company mediates both. Literally, sellers apply a factoring company to provide payment from the buyer and avoid the risks. Sellers negotiate the agreements, bills, or checks which are obtained from the buyer to the factoring company and get payment from the company. Payment amount is affected by maturity dates, debtor count, debtors' financial risks, seller's endorsement, bill amount and bill count and is approximately 80%. Missing amount is a benefit of the factoring company for its service.

There are 3 or 4 parties in factoring transactions, depending on whether the transactions are domestic or international. Domestic factoring transactions become 3-sided, with the buyer and the seller being in the country and the factoring company involved in the process.

In case the buyer is abroad, and the seller is in the country, both the factoring company and the correspondent factoring company are involved in order for the factoring company to collect the payment.

Table 1.1: Brief information about factoring parties

Factoring Parties	Roles
Seller	The party that sells the product and receives a check in return
Buyer	The party that makes a purchase and issues a check in return
Factoring Company	Organization providing factoring services
Correspondent Factoring Company	Correspondent factoring firm informs the factoring firm about the buyer's limit.

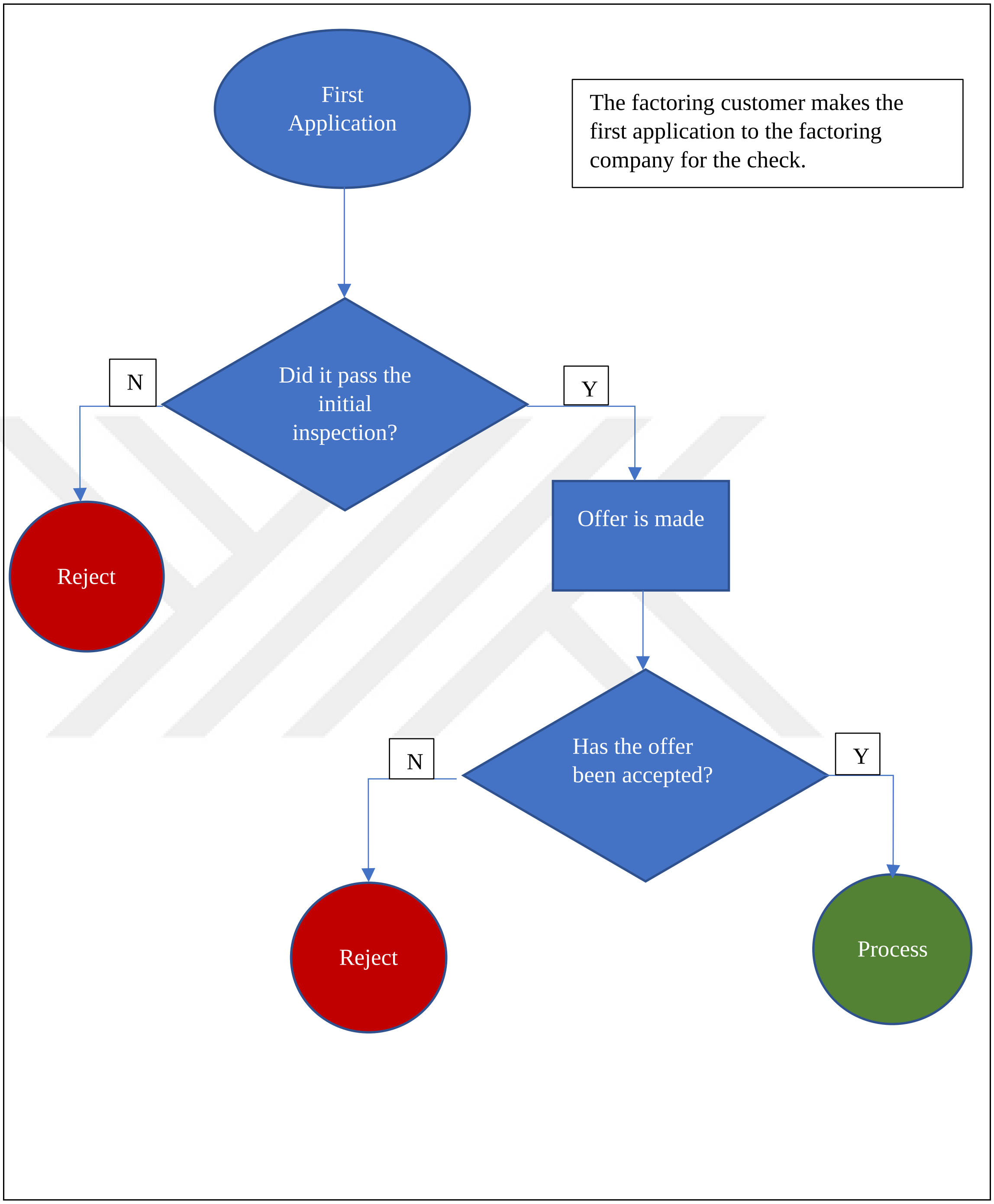

First Application
The factoring customer makes the first application to the factoring company for the check.
Did it pass the initial inspection?
N
Y
Reject
Offer is made
Has the offer been accepted?
N
Y
Reject
Process

Figure 1.1: Brief information of factoring processes.

While this flow was being drawn up, the data owner factoring company was interviewed, and they expressed their operations with this flow.

1.2 Risks of Factoring

Financial markets are always faced with risks, which are grouped into financial risks and non-financial risks. While financial risks consist of credit risks, liquidity risks and market risks and non-financial risks consist of operational risks, political and legal risks, and economic risks. [25]

1.3 Customer Segmentation

Customer segmentation is explained by Wendell R. Smith as the process of obtaining homogeneous customer groups from a heterogeneous customer group. [26] While digitalization was not that much popular, firms were advertising by global channels like TV, radio, and newspaper. This caused unnecessary spending since expenditures did not provide sufficient return. Engaging the attention of the population is not too easy while the advertisements are not individual.

In the digital age, collecting data about the customers and reaching them with different channels became capable and manageable. As a result of technological improvements firms have a chance to collect overmuch data thus traditional management methods are not enough to evaluate this data. In the age of big data Machine Learning (ML) is used for evaluation and clustering data. Machine Learning algorithms can qualify the data in different ways which cannot be possible with the eye. Usage of the ML algorithms helps companies to cluster their customers according to customers' similarities. While with business knowledge companies may have their own segmentation methods, ML algorithms can help to improve their models and make them realize different opportunities while evaluating customers.

In the current study there have been interviews with finance sector Analytic CRM department employees and common usage of K - means for customer segmentation has been confirmed. In finance companies there are many types of segmentation for the specific usage areas. While customer value segmentation is meaningful to use in the marketing department, customer risk segment is more helpful to use in the credit or check department.

There must be a problem, or a reason to divide customers into the clusters. After that, according to the problem, dividing customers into meaningful groups makes sense. For instance, if a problem is understanding customer’s behavior for the campaign management department, customer communication data can be observed. Further thought, the campaign management department needs to find the best way to communicate with the customers and get responses from them for each campaign. They have hesitation that they could not reach customers effectively. In this case the most important step must be evaluation of the previous behaviors of the customers. Campaign distribution channel may be an important attribute for this research.

The credit department in a company may have different observation areas. They are mostly interested in customer’s possible risks. In this case how they answered a campaign is not a key variable. Depending on the situation may have effects but as a first step it will make no sense to find the customer's risk segment. In such a case financial companies need customer’s endorsement information, shareholders and loan information which was taken out already and more. Taken out loans can be checked via integrations from controller companies.

As mentioned above for various needs there are different segmentation types and true questions, and approaches convey organizations to the required ones. In some companies preparing segmentation rules are based on business knowhow however clustering customers according to their similarities shows more than the rules known before.

Pareto Principle supports the idea that 80 % of the results happen because of 20 percent of the reasons. For marketing it can be said 20 % customers cause 80% income [27]. Based on this idea, clustering data for a specific purpose is helpful to organizations to understand and give a service or response to a specific group of their customers. While segmenting mostly the main purpose is clustering the customers and predicting income or risks which would be caused because of them. These predictions are affordable by machine learning methods, supervised or unsupervised. In the current study some of these methods are explained deeply and a study for segmentation implemented for factoring organization. As it mentioned before there are different algorithms for different segmentation problems. Topic 3 contains some of them, which is helpful for the current study.

1.3.1 Customer Metrics

1.3.1.1 Customer Lifetime Value

Customer orientation is to offer the customer the level of service they deserve. The measure of what a customer deserves in CRM is lifetime value. (Customer Lifetime Value (CLV)) [28]. Customer lifetime value is the money a customer will bring to a brand over time. CLV tells how valuable a customer is to a brand and gives an idea of its overall value. This helps to understand how much investment is required to retain the customer. Not only that, the CLV also gives an idea of whether any customer will become a recurrence. The higher the customer lifetime value, the higher the brand loyalty and purchasing habit is expected. For a customer, if this value is low, he or she is probably a passive one-time shopper and is more difficult to retain. CLV concentrate on customers who are constantly in touch with your brand or show potential to be. The best way to maximize customer lifetime value is to invest in retention.

The CLV value is simply obtained by multiplying the average order value, average purchase frequency, and customer age (time since becoming a customer). Many models with different assumptions and different foundations have been developed to determine CLV.

Most of these models can be broadly classified as scoring models, probability models, and econometric models. In scoring models, simple scores are generated based on user' purchasing behaviors (for example, novelty, frequency, and monetary value (RFM) model). In probability models, user habit is viewed as an expression of an underlying stochastic process determined by individual characteristics (eg, the negative binomial distribution (NBD) model). In econometric models, user habit is explained as a function of a set of covariates.

However, the two key steps in evaluating CLV are estimating the net cash flows the firm waits to get from the customer over time and calculating the present value of that cash flow [29]. The factors that influence the CLV model are generally considered in three categories, these are revenue, costs, and retention rate [30].

1.3.1.2 Recency, Frequency, Monetary

RFM is one of the classification methods if frequency is an important variable, then this method can be used for segmentation problems. According to the interviews done for this study, in banks, the most popular finance area, RFM method is not used. Nevertheless, to decide that if frequency is an important variable the data can be classified by using K - means algorithm and then, specific class can be called as a segment. Regression model may implement the new data variables and effects of frequency to the segment variable can be gauged. RFM does not apply to research for new customers, as transaction information is not available to potential customers [31].

Recency means the time passed since the last transaction. If recency is getting longer, the last transaction is very far, this is a signal that customer’s behaviors have changed [32]. Such a negative course of customer behavior is a negative situation for the market, and this should be prevented. Since this situation also means that the customer is lost, situations that distract the customer from the brand should be determined. This change in behavior in the customer can be specific to that customer, but it can also represent potential customer losses.

Frequency represents the number of transactions in a specific term, to illustrate, once a year, once a quarter, or once a month. The higher the frequency, the greater the F [33]. This specific period can be changed according to the sector and the problem looking for a solution. If a customer's purchasing pattern can be presumed, then future sales turnover can be presumed by reminding the customer of their next merchandising.

A high frequency value for a customer may be an indicator of customer loyalty, but the root cause of this loyalty may be very different. For example, in a sector where the customer physically transacts, the presence of a single institution around may have caused the customer to frequently shop with this institution. In case of competition, the loss of the customer may occur. While examining these values, supporting them with customer satisfaction surveys may be effective in producing healthier results.

Monetary means to the amount of money consumed in a given term. The more money, the bigger the Monetary [33]. Monetary value relates to the total amount of sales generated by the customer. It is ineluctable that the customers who spend the maximum money are in the marketing focal point of the business. Still, if marketing spends all its attention on incitation these customers to continue their purchases, it can cause them to miss out on potential customers and customers they haven't seen their actual buying potential still.

2. MACHINE LEARNING ALGORITHMS BASED ON SEGMENTATION

2.1 Supervised Learning

Supervised and unsupervised learning algorithms have shown great potential in extracting information from large datasets. Supervised learning reflects the algorithm's ability to generalize information from existing data with target or tagged cases so that it can be used to predict new (unlabeled) situations [34].

Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Supervised learning is divided into two problems in data mining, classification, and regression.

Classification problems use an algorithm to separate test data from each other and correctly assign data to specific categories. For example, it separates dogs and cats into separate classes. To give an example from daily life, these algorithms can be used to categorize junk e-mail or text messages in an aside folder from our inbox. Support vector machines, linear classifiers, decision trees, and random forest are all well-known classification algorithms.

Regression finds the relationship between dependent and independent variables. Regression models help estimate numeric values based on different data points, such as annual income projections for businesses. Linear regression, logistic regression, and polynomial regression are some of the well-known regression algorithms [35].

2.2 Unsupervised Learning

Unsupervised learning refers to the process of grouping data into clusters using automated methods or algorithms on unclassified or uncategorized data. In this case, algorithms must "learn" underlying relationships or features from existing data and group cases with similar characteristics [34].

Unsupervised learning machine learning algorithms are used to analyze and cluster unlabeled datasets. These algorithms are called "unsupervised" because they found hidden models in data without the necessity of human interference. Unsupervised learning models are used for three main purposes: clustering, association, and dimensionality reduction.

The clustering method is a data mining technique used to group unlabeled data by determining whether they are similar or different. For example, K - means clustering algorithms, one of the most used unsupervised algorithms, assign similar data points to groups. The K value refers the number of clusters for the algorithm.

In association problems, the relationships between features in a given dataset are found using different rules. Association algorithms use rules to detect connections between variables and their occurrence patterns.

Dimensionality reduction when a dataset has a high number of features, can be employed as a learning technique to manage the inputs. With this method, the number of dimensions is reduced in a manageable way while maintaining data integrity and quality. Size reduction, which is generally used in the preprocessing stage, can increase the efficiency of machine learning algorithms [35]. The aim of clustering algorithms is to discover clusters among data defined by various variables. There are many clustering methods that exist and generally fall into two categories.

The first one looks for distance's similarity or dissimilarity. Hierarchical clustering which builds trees and K-means which categorize data according to the cluster count can be an example for this group. The other one is a model-based method which tries to do optimization of data and the model. In this model every cluster presented by a parametric distribution like Gauss [36].

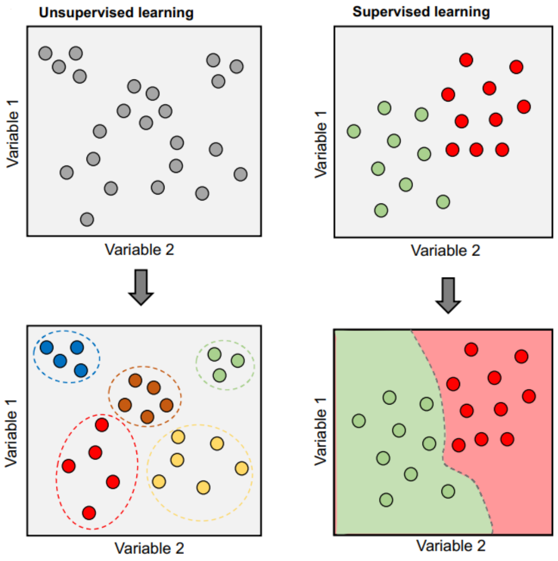


Figure 2.1: Supervised and Unsupervised Learning.

The figure on the left shows the clustering problem and the figure on the right shows the classification problem [37].

2.2.1 Clustering with K-means

The K - means algorithm, discovered by Mac Queen in 1967, is one of the easiest unsupervised learning algorithms that resolves the clustering problem [20]. K -means is one of the most used methods for data analysis since its computing velocity and success [38]. It is one of the earliest algorithms [39] and one of the partitioning clustering techniques, is the most widely used clustering algorithms in scientific and industrial applications [40] but it is very affordable for extremely large datasets [20]. This algorithm is one of the most important clustering approximations based on the sum of squares criterion. When we go back to the origin of this algorithm, it is seen that it has been proposed by many scientists in different ways and under different assumptions [41].

This algorithm finds the K number of clusters, K refers to the number of clusters given to the program by the user. The K - means algorithm basically works as described below.

a. Initial centroids have to be chosen. (Count of centroids shows the dataset how many clusters are divided into). We can refer to them c1, c2, c3, etc.

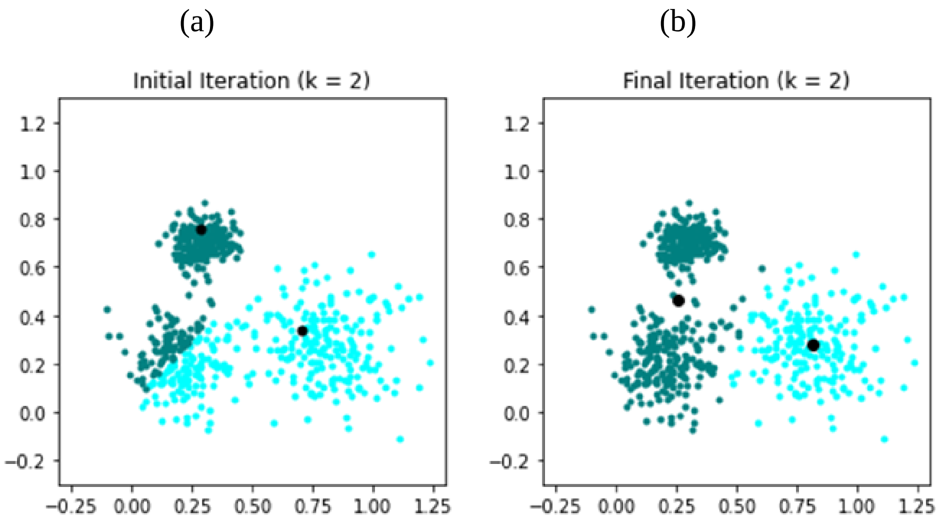
b. Each data point is assigned to the closest centroid. Data points can be referred to as I1, I2, I3, etc. while checking the nearest centroid , Euclid, Manhattan, or Minkowski methods can be used. After calculation, each data point is assigned to the closest centroid. There are few ways to find distances between two points, which are explained below.

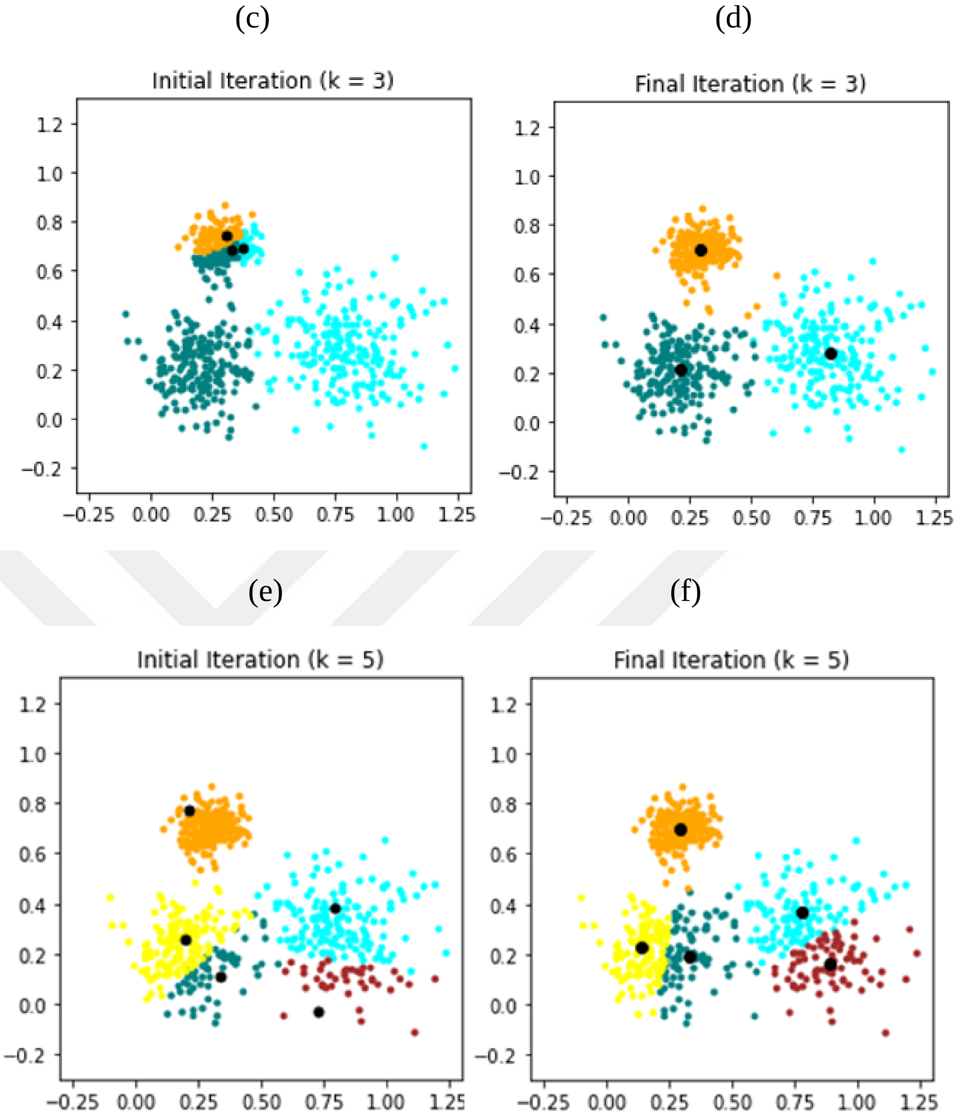
c. New centroid is calculated. When all data points are divided into K clusters, the sum of each dimension's value of all points are divided by the count of points. For example, when there are 3 points in a cluster in 3D space a new centroid will be calculated formula 3.1 as below, in this 3 points' assigned cluster.

$$c = \sum_{i = 1} n(X_i)/n$$

$$cnew = ((x1 + x2 + x3)/3, (y1 + y2 + y3)/3) \tag{3.1}$$

d. b, c steps are repeated until there will be no change between calculated last two centroids.




(c) (d)

(e) (f)

Figure 2.2: Clustering visualization for K – Means

In Figure 3(a) shows the initial clusters when k is chosen 2, (b) shows the final clusters when k is chosen 2. (c) shows the initial clusters when k is chosen 3, (d) shows the final clusters when k is chosen 3. (e) shows the initial clusters when k is chosen 5, (f) shows the final clusters when k is chosen 5.

2.2.2 Gaussian Mixture Model Clustering

Gaussian Mixture Model clustering (GMM) is an unsupervised learning algorithm like K - means, but there are some differences between them. The GMM algorithm is based on the Gaussian distribution, which makes it effective in solving problems that cannot be solved with K - means. Many datasets can be modelled by using Gauss distribution. It means that the datasets can be created from several different Gauss distributions. As a result of this, in cases where algorithms acting assuming that the dataset consists of a single distribution are insufficient, GMM can be quite successful as it enables the modeling of different distributions.

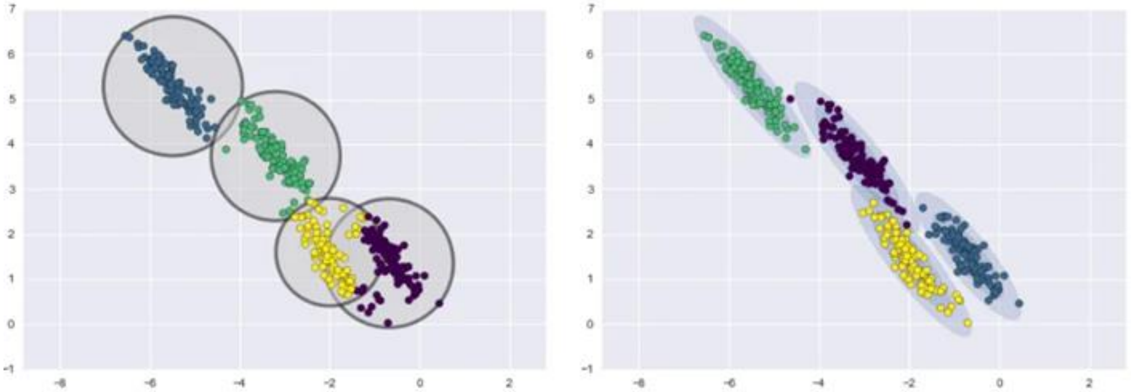


Figure 2.3: Difference between K-Means and GMM. [42]

The picture on the left shows K – means clustering and the picture on the right shows GMM clustering.

The formula 3.2 below is the probability density function of the Mixture model consisting of n-gaussian distributions. The w parameter is the weight of the sample for all Gaussian distributions.

$$P(X|\Theta) \ = \sum_{i=i}^{n} WiN(X|\mu i, \sigma i) \tag{3.2}$$

Suppose there are k-numbers in the model. We need to have the mean and variance separately for each of the k-number clusters. We need to arrive at these values by “Maximum Likelihood Inference”. However, this approach is insufficient in cases where the solution is not analytical. Therefore, the Expectation-Maximization (EM) Algorithm is used.

The EM algorithm is an iterative algorithm that finds the maximum likelihood [43] estimates for parameters when there are missing observations or some hidden variables of the sample. Since it is an iterative algorithm, it can also be used in problems that cannot be solved analytically. The goal of the Expectation Maximization algorithm is to maximize the probability of $P(X|\Theta)$ relative to Θ. ($P(X|\Theta) = = P(X1,X2|\Theta)$, X is independent samples, X2 is missing observations, Θ is matrix $[\mu,\Sigma,\omega]$). According to the formula, $P(X|\Theta)$ is maximized with the inferences made for the parameter Θ. The Expectation Maximization algorithm has two basic steps: Expectation step and Maximization Step.

Expectation step: $Q(\theta|\theta m-1)=Ep\theta m-1(X1|X2)\{\log\ p\theta(X1,X2)\}$. For k-number clusters, the best probabilities for the unknown data are estimated with arbitrary means and variances. The expectation of log $p\theta(X, Y)$ of all data log likelihood is calculated according to the conditional probability density function ($p\theta m-1$ $(X|Y)$) of the latent variables. The conditional probability function is calculated using $\theta m-1$ parameter values, which is the last estimate of the parameters and calculated in the previous iteration.

Maximization Step: $\theta m = \text{argmax}\theta Q(\theta|\theta m-1)$. New estimates of the parameters are obtained by substituting the estimated missing value and calculating the maximum likelihood over the data. The parameters are updated to maximize the distribution of the data and the Hidden Variable. The algorithm starts with a value of θm-1, this initial value is either randomly chosen or determined with the help of other clustering algorithms. These steps are performed sequentially until a certain criterion is met or the maximum number of iterations is reached [44].

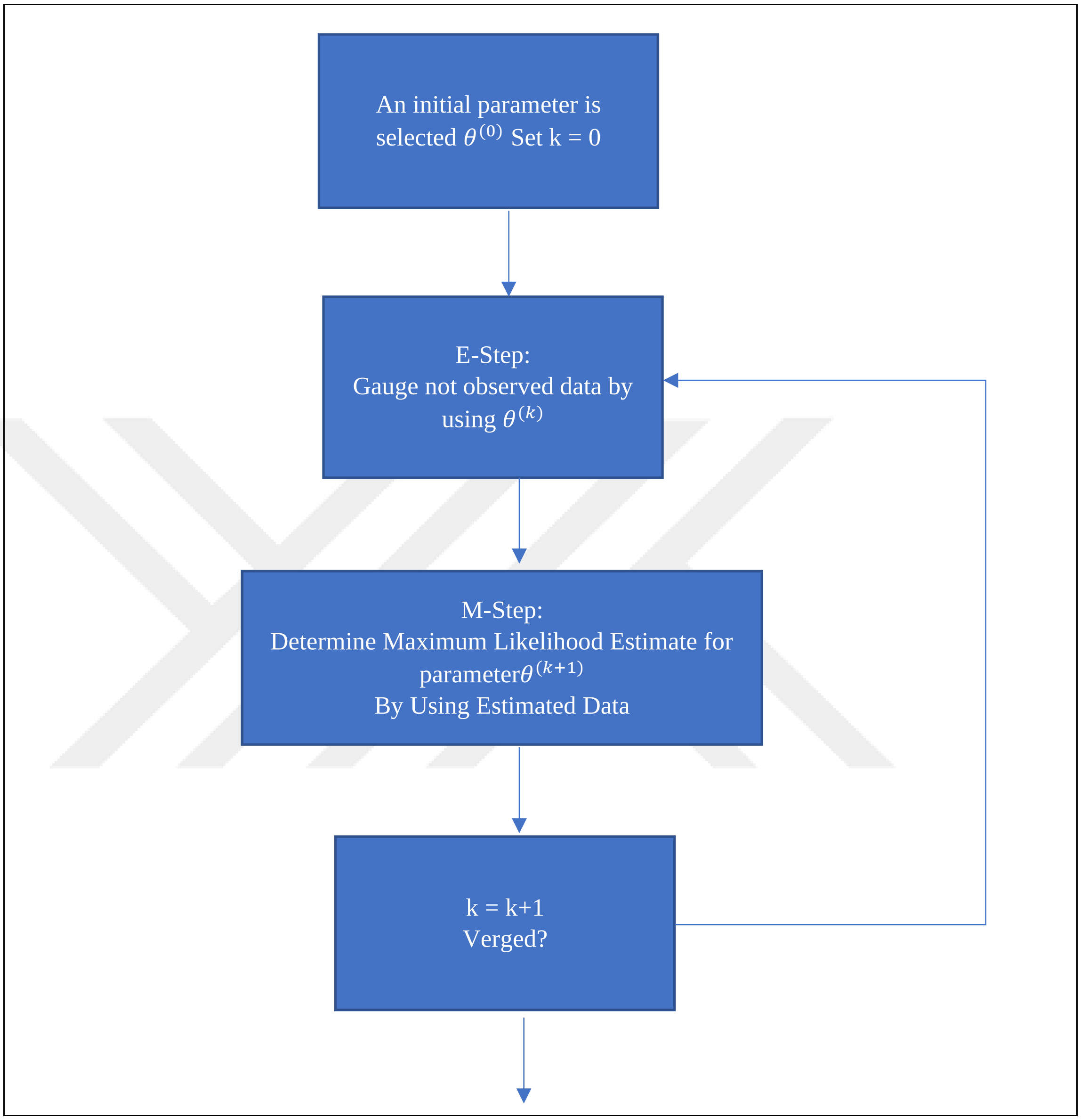

An initial parameter is selected $\theta^{(0)}$ Set k = 0
E-Step:
Gauge not observed data by using $\theta^{(k)}$
M-Step:
Determine Maximum Likelihood Estimate for parameter$\theta^{(k+1)}$
By Using Estimated Data
k = k+1
Verged?

Figure 2.4: Brief information for EM Algorithm.

E - Step estimating the unobserved data and M - Step maximum probability calculation continue until there is no change in the estimation. [44]

2.2.3 Density-Based Spatial Clustering of Applications with Noise

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was introduced in 1996 and formed the basis of density-based clustering techniques [45]. DBSCAN algorithm, unlike K - means algorithm, evaluates clusters as density. In this way, the DBSCAN algorithm gives successful results in detecting clusters that are separated in any way (nested clusters or moon-shaped datasets, etc.).

The algorithm takes the densities of objects into account when creating sets. Clusters are described by high-density data objects, while clusters of low-density objects point out outliers or noisy points.

In addition, it is frequently used to determine clusters of different sizes and shapes [46]. The DBSCAN algorithm uses min_samples and eps (ϵ) parameters to determine density. The eps parameter defines the maximum distance between samples, while the min_samples parameter is defined as the number of samples that must be found for the point to be considered a seed point. With these two parameters, it is aimed to have an idea about the density.

While the min_samples parameter principally controls how tolerant the algorithm is to noise (in noisy and large data sets it may be desirable to increase this parameter), it is very important that the eps parameter is chosen appropriately for the dataset and the distance function and cannot usually be left at the default value. Check the local neighbors of the points. When very small is selected, most data are not clustered (and is labeled -1 for “noise”). Selecting too large causes close clusters to be merged into a single cluster, eventually returning the entire dataset as a single cluster.

The advantages of the DBSCAN algorithm can be listed as follows; DBSCAN does not require pre-specifying the cluster count, performs well with randomly shaped clusters, and is resistant to outliers. The disadvantages of the DBSCAN algorithm are as follows; cannot cluster well datasets with large differences in densities. It can be difficult to choose a significant eps value if the data is not well understood and finally DBSCAN is not completely deterministic. This is because the algorithm starts with a indiscriminate point. Therefore, boundary points that can be reached from more than one cluster can be member of any cluster.

2.2.4 Hierarchical Clustering

As it is common, the k - centered clustering method has a disadvantage. The number of clusters must be determined beforehand. Hierarchical clustering has been developed to eliminate this disadvantage. The general logic of the hierarchical clustering algorithm is based on the combination of similar features or vice versa. According to this working logic, there are two basic approaches: agglomerative and divisive.

In the unifying approach, also known as induction (bottom up), initially all objects are separate from each other. In other words, each of the available data is considered as a separate set. Then, clusters with similar attributes are combined to form a single cluster.

In the top bottom approach, unlike the inductive way, a discriminatory strategy is predominant. In this way, there is just one cluster as an initial. In every stage, objects are separated from the main cluster according to the distance/resemblance matrix, and different subsets are formed. As a result of the process, every data becomes a cluster [47].

In hierarchical cluster analysis, resemblance and discrepancy calculations between data are updated in every step. In the agglomerative hierarchical cluster algorithm, each unit is initially considered as a separate cluster, and similar units are brought together, and n units are gradually placed in n, n-1, n-2, n-r clusters, respectively. The general working structure of the algorithm is as in the following steps a, b, c, d [48].

a. Operations are started with n individuals and n clusters.
b. The two closest clusters are combined.
c. By reducing the number of clusters by one, the iterated distance matrix is found.
d. Steps b and c are repeated (n-1) times.

2.2.5 Fuzzy c – Means Clustering

A pioneering application for Fuzzy c – Means (FCM) clustering theory was made by Ruspini in 1969 [49]. The foundations of the FCM algorithm were first laid by J.C Dunn in 1974. He also proved the algorithm he presented mathematically in his study [50]. Then, the algorithm was developed by Bezdek in 1981 [49].

The K - means algorithm calculated the distance of every cluster element from the cluster nuclei and assigned the cluster elements to the cluster to which they were closest. Fuzzy cluster analysis, on the other hand, allows a data to belong to more than one cluster and gradually assigns [0,1] values to the data points. The sum of these values is equal to 1 [51].

Standard FCM uses the Euclidean distance as a cost function to be minimized and is expressed by the equation 3.3 [52].

$$J_{FCM}(U,V) = \sum_{\mathrm{i}=1}^{C} \sum_{j=1}^{N} \mu_{ij}^{m} \, ||x_j - \, c_i||^2$$

μ_ij : membership value

x_j: data point

c_i : center of the ith cluster

m: fuzzifier

(3.3)

The first step is the initial period, the second step is randomly chosen initial class centers and fuzzy c-sections. Then the fuzzification parameter m with $1 \leq m \leq \infty$ is given; with the coefficient ξ and the value ε> 0.

In the third step, the membership matrix is calculated. In the fourth step, class centers are updated. In step five, Δ is calculated, Δ > ε (if ε is a termination criterion between 0 and 1) to step two, otherwise step five. In the sixth step, the outcomes for the final centroids are found and the FCM is terminated. The figure below illustrates these steps [52].

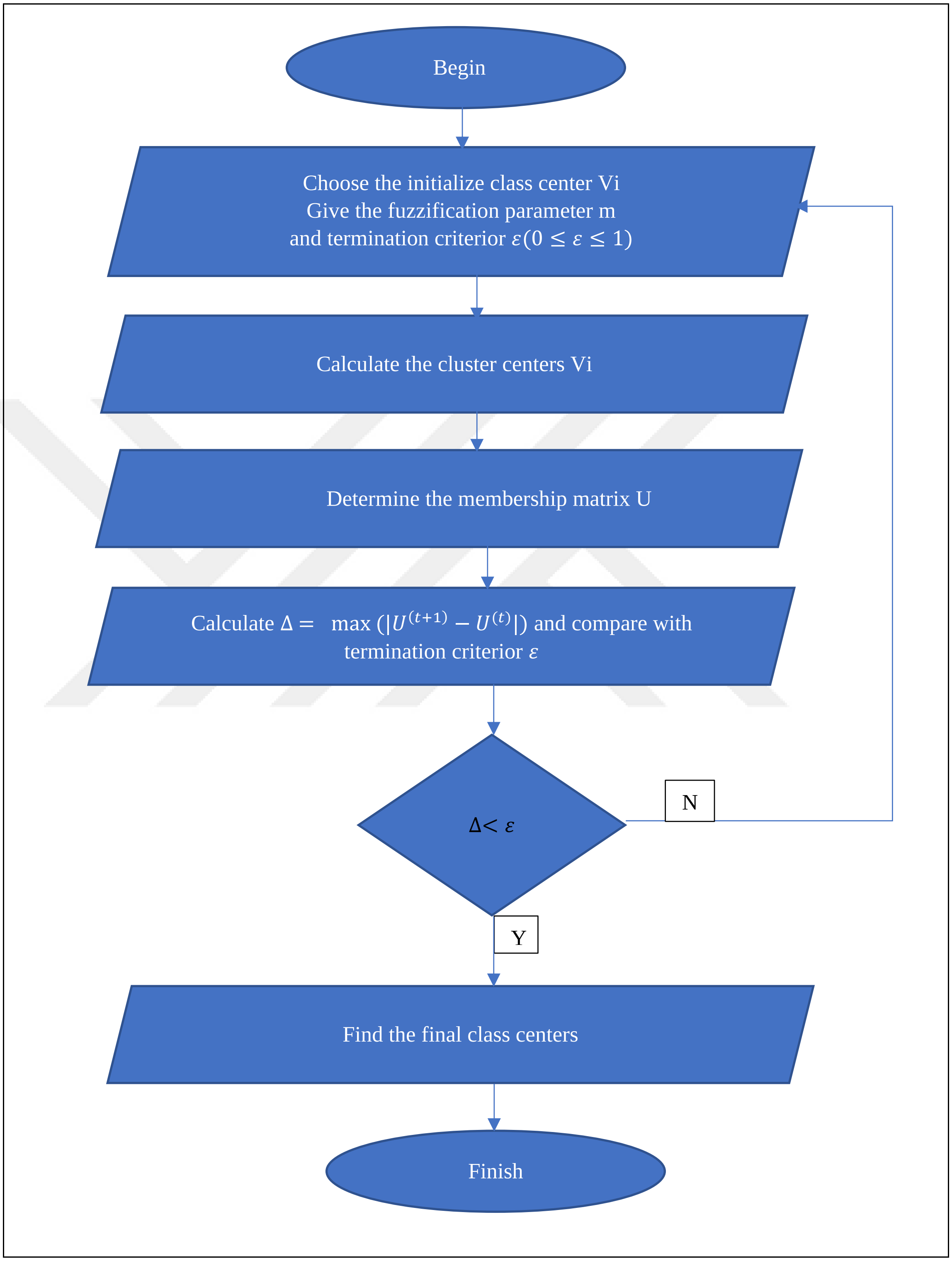

Begin
Choose the initialize class center Vi
Give the fuzzification parameter m
and termination criterior $\varepsilon(0 \leq \varepsilon \leq 1)$
Calculate the cluster centers Vi
Determine the membership matrix U
Calculate $\Delta =$ max $(|U^{(t+1)} - U^{(t)}|)$ and compare with
termination criterior ε
$\Delta < \varepsilon$
N
Y
Find the final class centers
Finish

Figure 2.5: Fuzzy c – Means Algorithm

2.3 Machine Learning Distances

Distances play an important role in machine learning. It provides the basis for many popular and effective machine learning algorithms, such as KNN from the supervised learning models and K - means from the unsupervised learning models.

The most widely used distance measurement method is Euclidean, but there are many different measurement methods. Figure 2.6 contains images of different distance measurement methods.

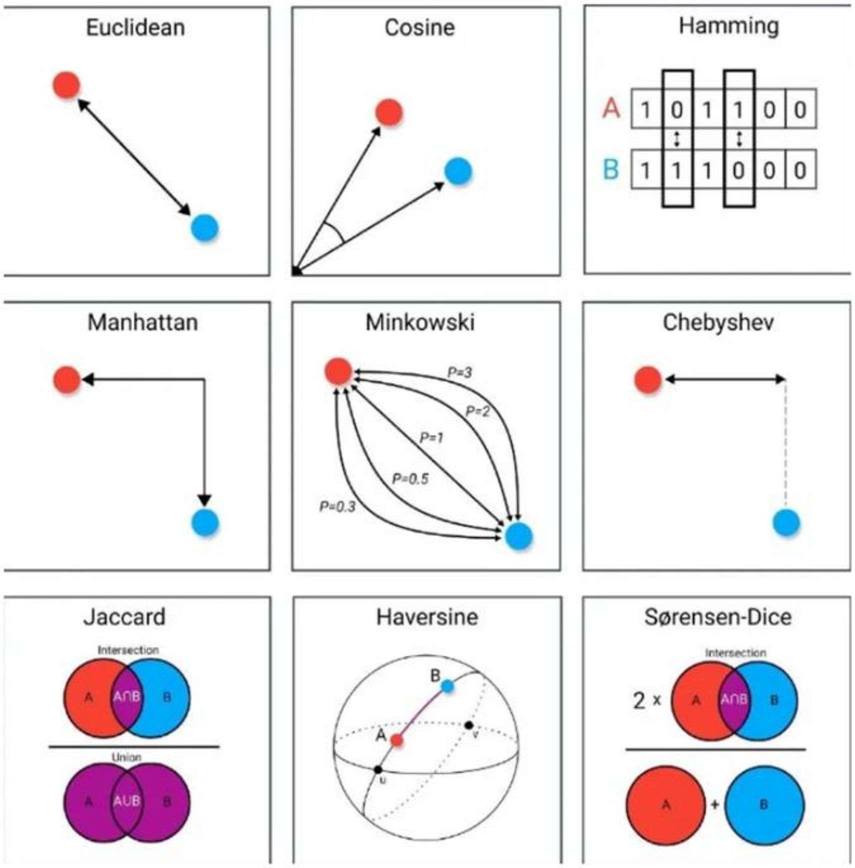


Figure 2.6: Distance measurements used in machine learning techniques [53]

2.3.1 Euclidean Distance

While P1 = (x1, y1) and P2 = (x2, y2) distance of these points will be like below, by using Euclidean distance.

$$d = \surd([(x2 - x1)\text{\textasciicircum}2 + (y2 - y1)\text{\textasciicircum}2\ ]\ ) \tag{3.4}$$

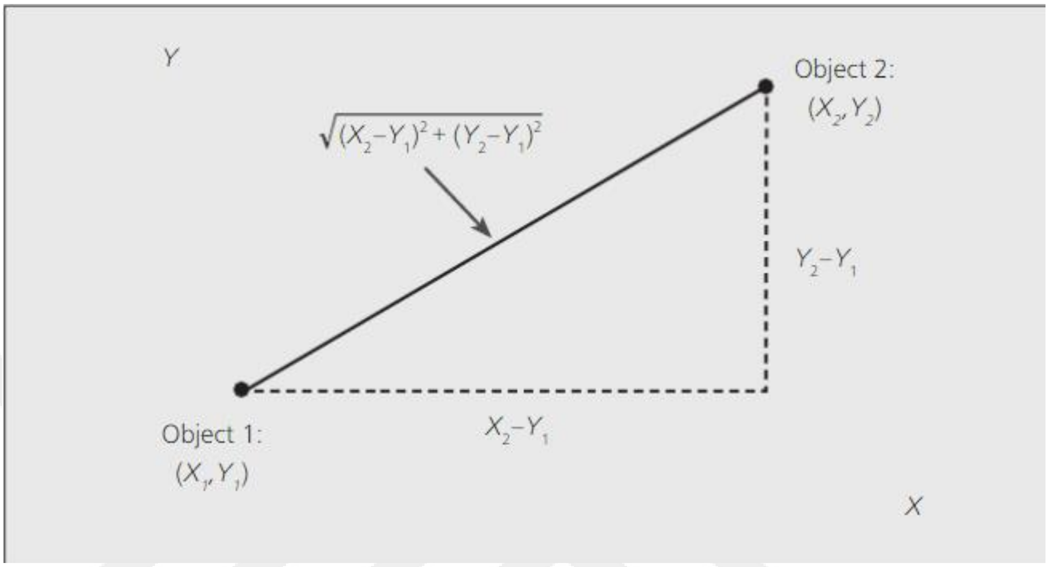


Figure 2.7: Euclidean distance between two points [54]

2.3.2 Manhattan Distance

The Manhattan distance is often used where the lines run parallel to the X or Y axis only. [55] While $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ distance of these points will be like formula 3.5 below, by using Manhattan distance.

$$d = |x2 - x1| + |y2 - y1| \tag{3.5}$$

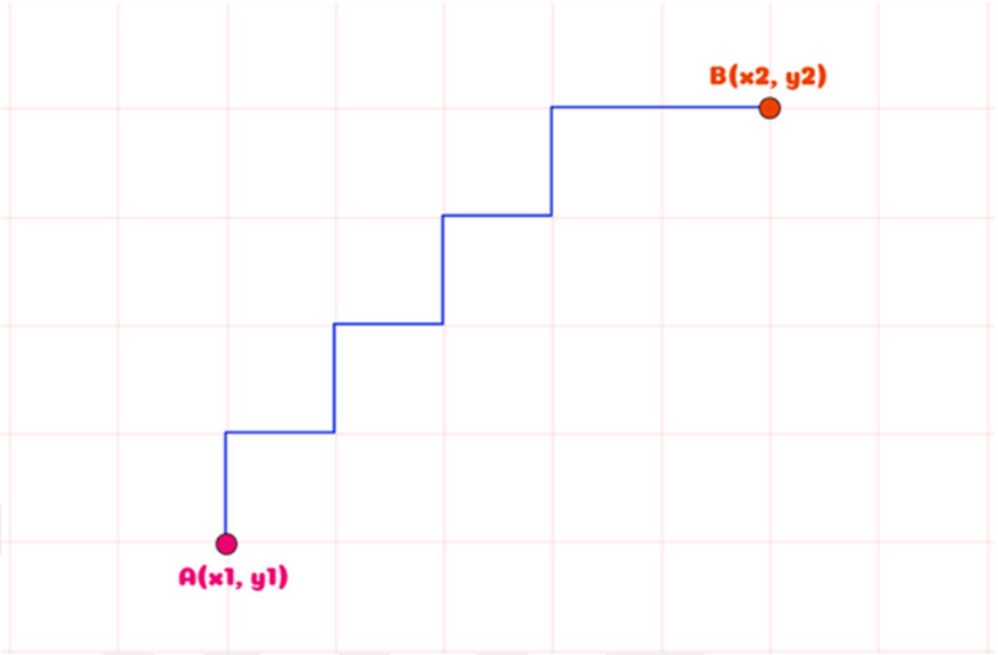


Figure 2.8: Manhattan Distance [56]

2.3.3 Minkowski Distance

It is a generalization of Manhattan and Euclidean distance. While X and Y are two points with N dimensions when p equals to 1, the formula will be the same with Manhattan distance and when p equals to 2, the formula will be the same as Euclidean distance formula.

$$\sum d = (\sum_{i=1}^{n} |xi - yi|^{\frac{1}{p}}) \tag{3.6}$$

2.4 Dimensionality Reduction

It is thought that more information is better than less information, but the number of variables in a single sample can cause problems in the big data world [57]. As the data size increases, the samples (points) become too dispersed in space. The density of the points or the distance between them is very important for many problems. As the data size grows, the density and distance information become meaningless and negatively affects the performance of these algorithms. This situation is called the curse of dimensionality

[58]. High dimensionality statistics and size reducing techniques are often used for data visualization.

2.4.1 Primary Component Analysis

Primary Component Analysis (PCA) method is also referred to as the Karhunen-Loeve method. It is a multivariate statistical method used in the fields of classification and image compression, which clarifies the variance covariance structure of a set of variables, through linear combinations of these variables, and ensures size reduction and commentary. In this method, (p) variables with the number of measurements (n) showing the interdependence structure; linear, vertical (orthogonal) and being independent from each other (k) are transformed into new variables. PCA is a very effective method to reveal the necessary information in the data.

By detecting common features in high-dimensional data, it lowers the number of size and compresses the data. It is certain that some features will be lost with size reduction; but the intention is that these missing features include very few info about the dataset [59]. In PCA, the goal is to find a new dimension set that best summarizes the data set [40]. While PCA facilitates the visualization of the data, it also contributes to the prevention of noise.

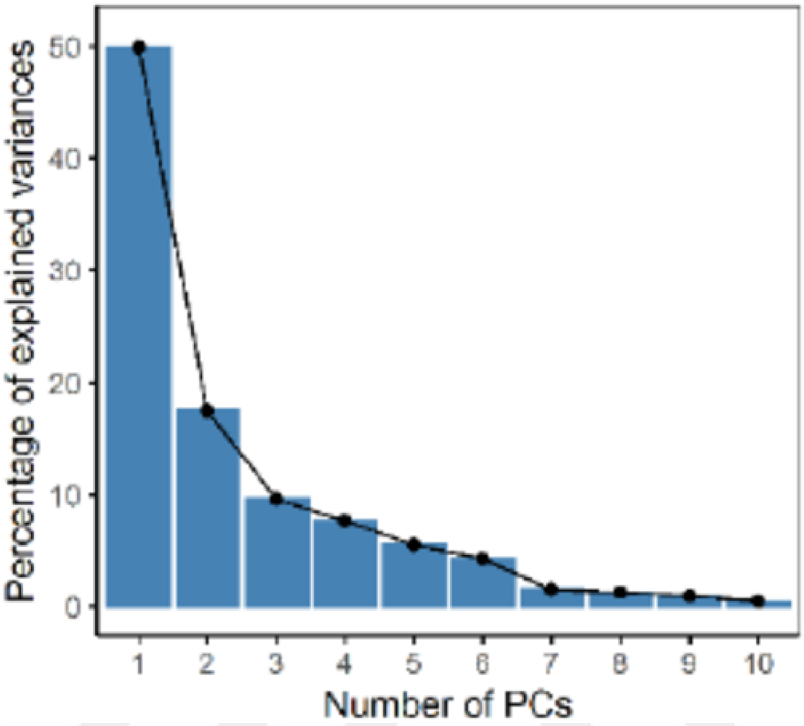


Figure 2.9: Percentage of variance while PCA = 10 [60]

2.4.2 T-Distributed Stochastic Neighbor Embedding

Manifold learning algorithms are primarily used for data visualization. T - Distributed Stochastic Neighbor Embedding (t-SNE) is one of the most practical manifold learning algorithms.

The focus of the t-SNE algorithm is to find a low-dimensional presentment in a way that retain the distances between points as far as possible [61]. t-SNE begin with a indiscriminate low-dimensional presentment for every data point and tries out to retain immediate points close jointly and distant points far apart in the original space. t-SNE places more importance on points that are immediate instead of maintaining the distance between points that are far from each other.

2.5 Selecting Optimal Cluster Count

Another major subject in clustering problems is finding the correct cluster count. The fact that the clusters are sufficiently homogeneous and discrete is an indication of successful clustering. It is possible to state correct cluster count or to find the success rate of the clustering with different methods to be applied for different algorithms. In the following headings, 5 different methods that can be applied for 5 different algorithms are mentioned.

2.5.1 Elbow Method

With this method the sum of the square of the distances of the points from the cluster center according to every K value is figured out. This method is called Within Clusters Sum of Square (WCSS). According to these values, a graph is drawn for every K value. The elbow point on the graph where the distinction between the WCSS starts to diminish is stated as suitable K value [62].

In fact, every point is a cluster and is also the cluster center. If every point is a cluster and cluster center at the same time, the distances will always be zero. So, in the model where every point is a cluster, WCSS will be zero.

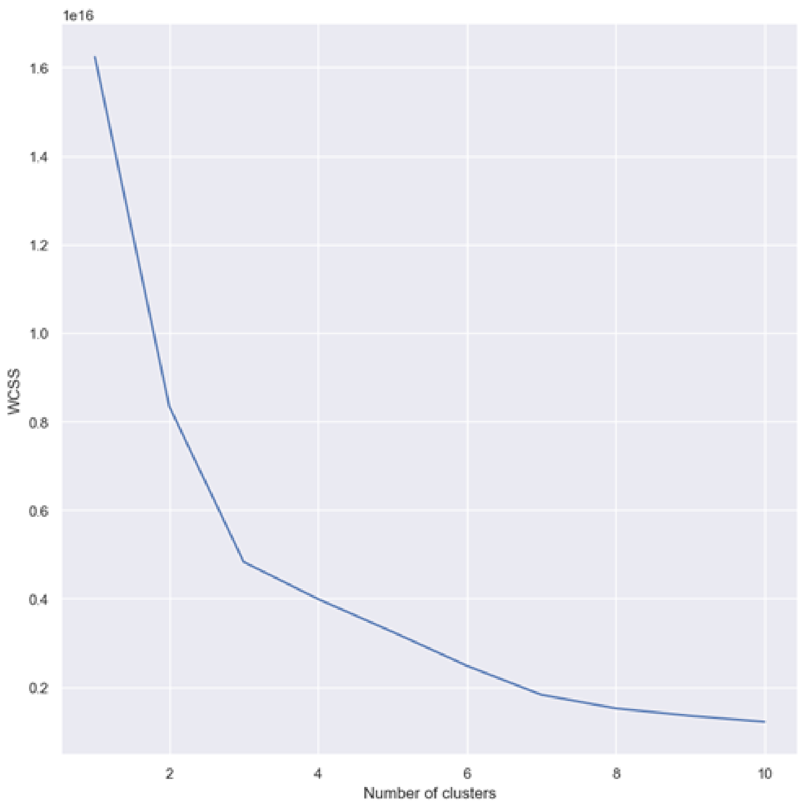


Figure 2.10: WCSS and optimal cluster count relationship

The elbow method output is shown in the graphic above. A good model will have a small WCSS value. It is seen that the WCSS decreases very fast up to 3, then the graph follows a more horizontal course. According to this graph, the ideal cluster count for the relevant data set should be determined as 3. More clusters do not reduce WCSS significantly and will reduce the interpretability of the model.

2.5.2 Silhouette Score for K-Means

The silhouette score value takes a value between 1 and -1. Higher values indicate that the algorithm is successful, while negative scores indicate that the algorithm has failed clustering.

Conceptually, the Silhouette score uses several distance parameters to measure how far a point is from its own cluster compared to the center of a different cluster. If this value is negative, this data point indicates that it is closer to another cluster than the assigned cluster.

2.5.3 Dendrogram for Hierarchical Clustering

Dendrogram is a graphical presentment of clustering. Ordinarily , it is drawn backwards starting from the last set containing all objects and resemblance 0. In the analogy where 2 clusters are combined to form the recent cluster, the recent cluster is divided into 2 main clusters, and et cetera.

When resemblance or differences is stated, the corresponding axis is the elucidative axis. The longness of the perpendicular lines gauges the partition between aggregated clusters. Therefore, cutting the dendrogram to the resemblance corresponding to the lengthiest branches is one of the common practices used to obtain significant clusters [63].

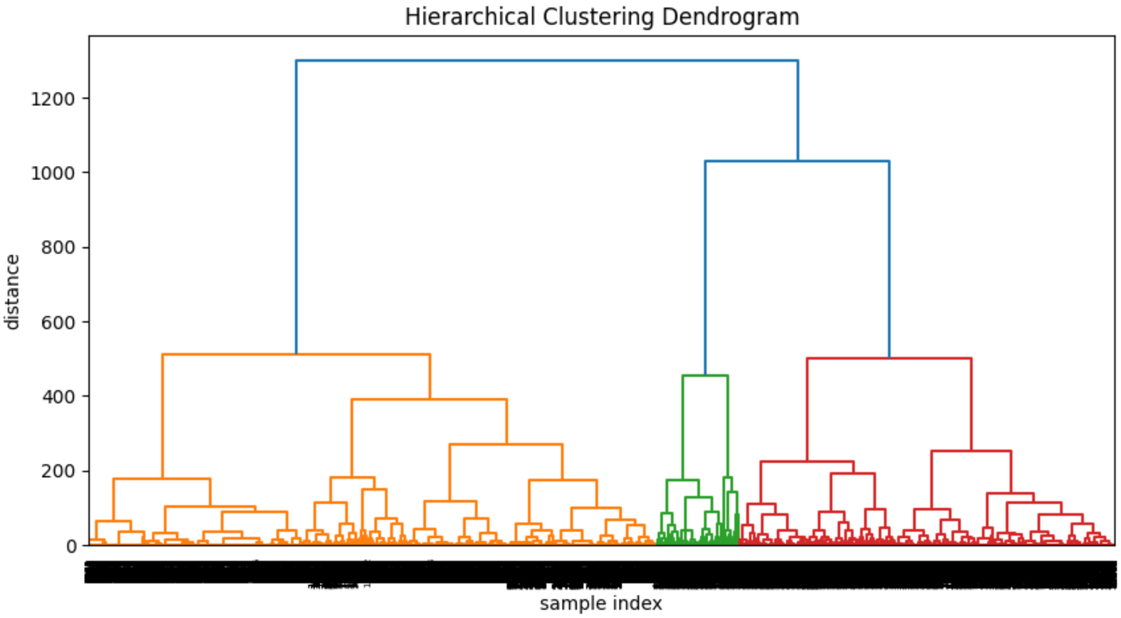


Figure 2.11: Dendrogram visualization

2.5.4 Bayesian Information Criterion for Gaussian Mixture Model

Bayesian Information Criteria is a model introduced by Gideon Schwarz in 1978 [64]. Among the most well-known and widely used tools in statistical model selection is Bayesian information criterion (BIC). Its demand is due to its computational primitiveness and effective performance in many modeling frameworks, including Bayesian applications [65].

The Bayesian Information Criterion is a consistent measurement method for selecting the number of components in mixture models [66].

$$BIC = k\,Ln(n) - 2Ln(L) \tag{3.7}$$

L = the maximized value of the likelihood function of the model

n = number of data points

k = number of free parameters to be estimated

2.5.5 Akaike Information Criterion for Gaussian Mixture Model

The Akaike information criterion (AIC) is one of the most widely used tools in statistical modelling. It is the first widely accepted model selection criterion and introduced in 1973 by Hirotugu Akaike as an extension of the maximum likelihood principle. Traditionally, maximum likelihood is applied to predict the parameters of a model after the structure and size of the model have been formulated [67]. AIC is simply calculated as in the following formula 3.8.

$$AIC = 2k - 2Ln(L) \tag{3.8}$$

L = the maximized value of the likelihood function of the model

n = number of data points

k = number of free parameters to be estimated

3. RESULTS

3.1 Dataset

The data set handled in this study comprise of the check data of a local factoring company that applied to the company between 2020 and 2021. Data contains 1.103.996 rows and 231 columns which have bool, float, int, object data types. This dataset has 2 boolean values, 115 float, 68 int, 46 object types of columns.

Table 3.1 Brief information about dataset.

Data Type	Data Count
Boolean	2
Float64	115
Int64	68
Object	46
Total	1103995

3.2 Data Preprocessing and Feature Engineering

3.2.1 Experiment 1: Transaction Based Clustering

First of all, the 231 fields in the data set were taken into account, and the one hot encoding (with the use of get_dummies function from pandas library) method was applied for all categorical variables and 400+ columns were obtained. Then the correlation matrix was run, and the highly correlated data were removed, leaving about 215 columns at the end of the run. Examining the data set with 215 columns and running the clustering algorithms did not produce interpretable results as well as creating a systemic load.

Considering this situation, the number of singular values for each of the 231 columns, columns containing data, containing 0 data, and non-data were examined separately and variables that should not be included in the clustering algorithm were determined.

Columns which have less than 30 % data count were 19, and they were removed from the data set. After removing large percentages of missing valued columns, the data set has a maximum 72,9% missing valued column and less.

There were 3 columns with order ID and process ID removed. Unique value counts were checked for each 228 columns and the columns with less than 2 unique values were removed. Columns which contained only nan and zero values were also removed from the data frame. Some columns include future information for itself. These columns are not informative for this study's clustering problem, and they were removed from the data frame. After these removals the data frame had 107 columns.

Boolean values were converted to the int type. Other boolean types of columns which seem as categorical are converted to 1-0 values and some columns which have numerical relationship but seem categorical are converted to numeric values. Some computations were done for a few columns, for example age of firms were calculated and foundation year columns were removed. Maturity date for checks also was removed for the same reason. When all necessary calculations are done, fields which contain date information were removed.

Some categorical data contained more than one information. Such columns were divided into 2 columns and each column was compared with its own variables and it was decided whether to use the get_dummies function. In some of them, the get_dummies function was applied, while in others, numerical relationships were found between the variables and included in the study in this way.

After these processes, 93 columns remained in the data frame. Null values were filled with the mean values for each numerical column. Dataset was extended with the categorical variables; each value was added as a new column which has only 1,0 values. For this process get_dummies function is used. Special characters in the column labels were removed to make the dataset much clearer. After all column's reduction and extension processes, we have 76 columns.

For the data set correlation matrix was created to check highly correlated variables and find the variables which cannot give extra information about the data set. To minimize the complexity of the algorithm's highly correlated variables was removed. While the correlation is more than 80% one of these variables were removed. After highly correlated fields were removed, we had 35 columns total. Optimal cluster numbers were examined by applying PCA to 35 columns, which were obtained as a result of all data cleaning and feature engineering studies, by taking 35 first, then 10, then 5 and finally 3 columns. Although 5 components represent approximately 80% of the data set, the ideal cluster count is quite high for both GMM and K - Means algorithms. Too few clusters cannot represent the differences in the data set, and too many is not meaningful in terms of manageability.

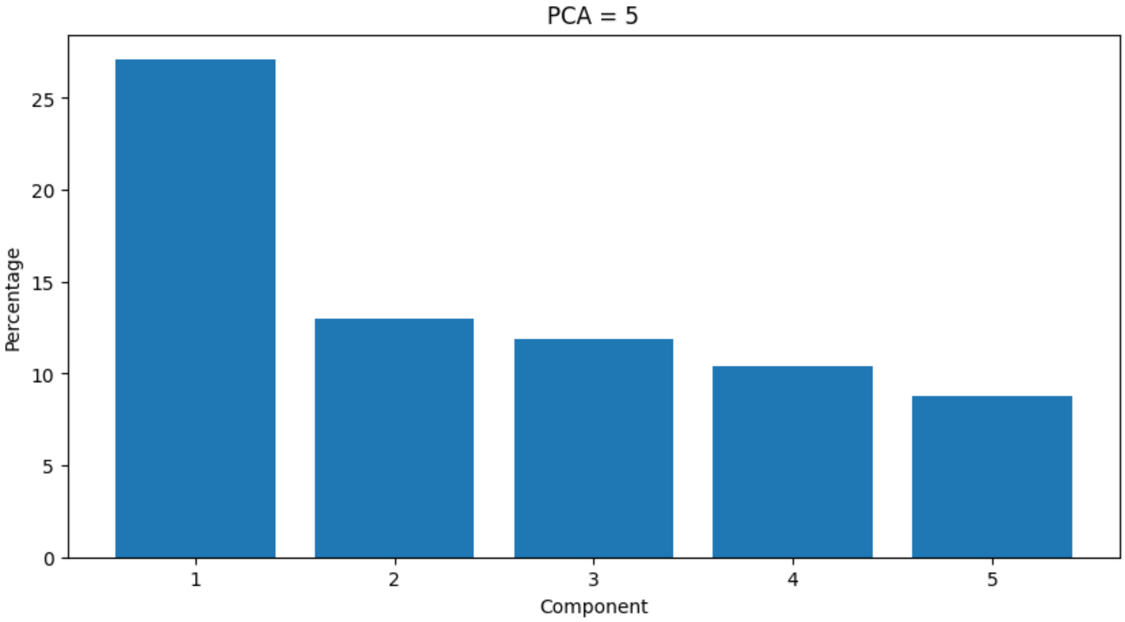


Figure 3.1 Data set representation while PCA = 5

In this experiment, 4 unsupervised learning algorithms were applied to the factoring dataset. These algorithms are DBSCAN, Hierarchical Clustering, K-Means, Gaussian Mixture Model algorithms.

The first trial aimed at transaction-based segmentation, each algorithm was applied with data of 10,000 or 100,000 rows randomly selected from the dataset of 1,103,996 rows, but these trials did not produce meaningful outputs.

Table 3.2 : Data counts according to algorithms

Algorithm	Data Count
DBSCAN	10.000
Hierarchical Clustering	10.000
K - Means	100.000
Gaussian Mixture Model	100.000

Since DBSCAN and Hierarchical Clustering are inappropriate for enormous data sets, algorithms were run with 10,000 rows of random data. To find the appropriate number of clusters for the DBSCAN, the Nearest Neighbors algorithm was used, and the epsilon value was calculated as 0.1. With this calculated epsilon value, the ideal number of clusters was determined. Since the optimal cluster count is found to be 1, it is assumed that this algorithm and the data set are not compatible. In Hierarchical Clustering, the creation of the dendrogram with 10,000 randomly selected data was very costly in terms of time.

3.2.2 Experiment 2 : Customer Based Clustering with 3 Scores

The second trial was conducted with three different scores obtained using certain variables from the data set for customers who made transactions in the last 1 year. Algorithms were applied on 3 scores of approximately 170.000 customers. These scores consist of the customer's activity score, trust score and potential score.

In this experiment, 3 unsupervised learning algorithms were applied to the factoring dataset. These algorithms are K-Means, GMM and FCM. The optimal number of clusters could not be calculated with GMM, FCM and could be produced only with K – means as 8. The outputs obtained as a result of this experiment were evaluated with the data owner institution and it was decided to conduct the third experiment.

3.2.3 Experiment 3: Customer Based Clustering with 2 Scores

The third trial was conducted with customers whose five or more applications were processed in the last year. The number of unique customers obtained using this filter was found to be 15829. Since there are applications returning to the process, only the activity and trust scores were calculated for these customers, and a clustering study was carried out with 5 different algorithms using these two variables. The clustering and segmentation studies that will be mentioned hereafter are the studies of the third experiment.

3.3 Applying Algorithms

The table 4.3 indicating the outcome of the algorithms used in the third trial is as follows.

Table 3.3 Applied algorithms and the results

Algorithm	Result
K - Means	the ideal cluster count was calculated as 4
Fuzzy c - Means	the ideal cluster count was calculated as 4
Gaussian Mixture Model	the ideal cluster count could not be calculated
Hierarchical Clustering	the ideal cluster count was calculated as 3
DBSCAN	the ideal cluster count could not be calculated

3.3.1 Implementation of K – Means Algorithm

Firstly, the K - means was applied because it is the most popular algorithm and the most commonly used in clustering problems, it is convenient to use with big data. To find the ideal cluster count, both elbow method and silhouette score were calculated.

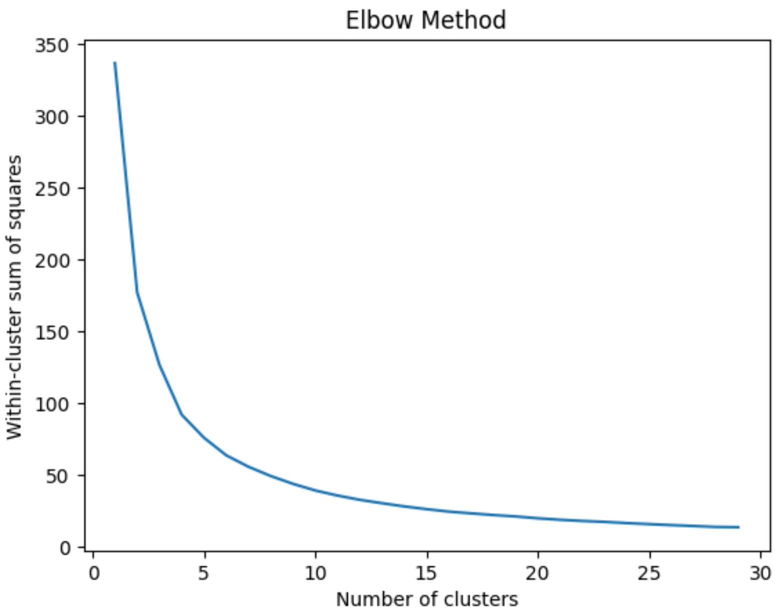


Figure 3.2: Finding optimal cluster count by Elbow Method

When the output produced by the Elbow method is examined, it is seen that the optimum cluster count can be considered as 4. Silhouette score values were also calculated, and different possibilities were also wanted to be observed.

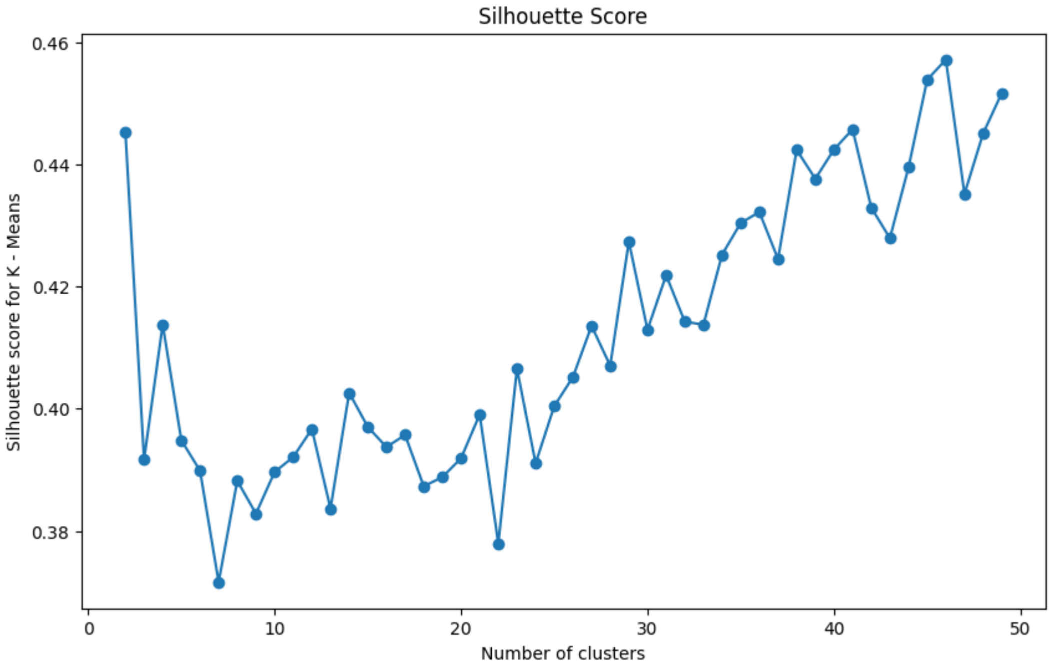


Figure 3.3: Silhouette Score for K – means algorithm

Since the silhouette score scales from -1 to 1 and values close to 1 represent the number of clusters that are best differentiated, the maximum values were examined. The highest point of silhouette score calculated up to 50 clusters is k = 46 point. Since it is not a manageable situation to assign 46 different segmentations to customers, it was decided to determine the highest manageable score level, k = 4, as the optimum cluster count.

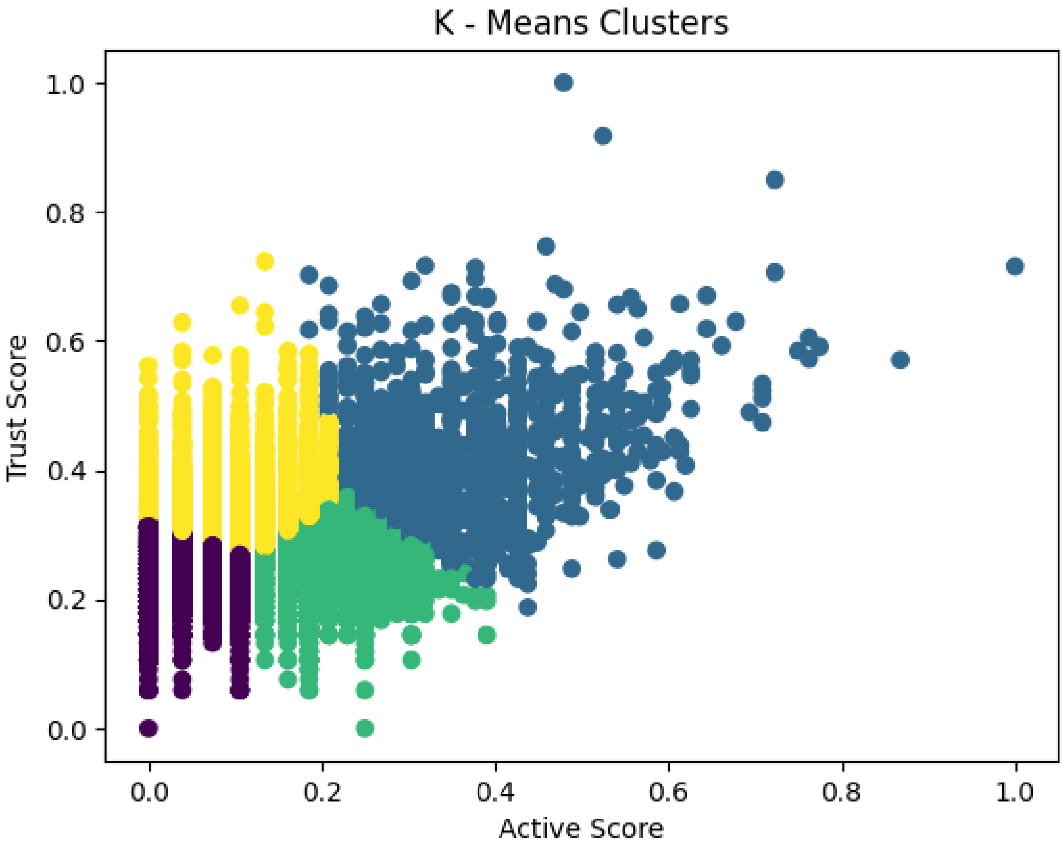


Figure 3.4: K - Means clusters while k = 4

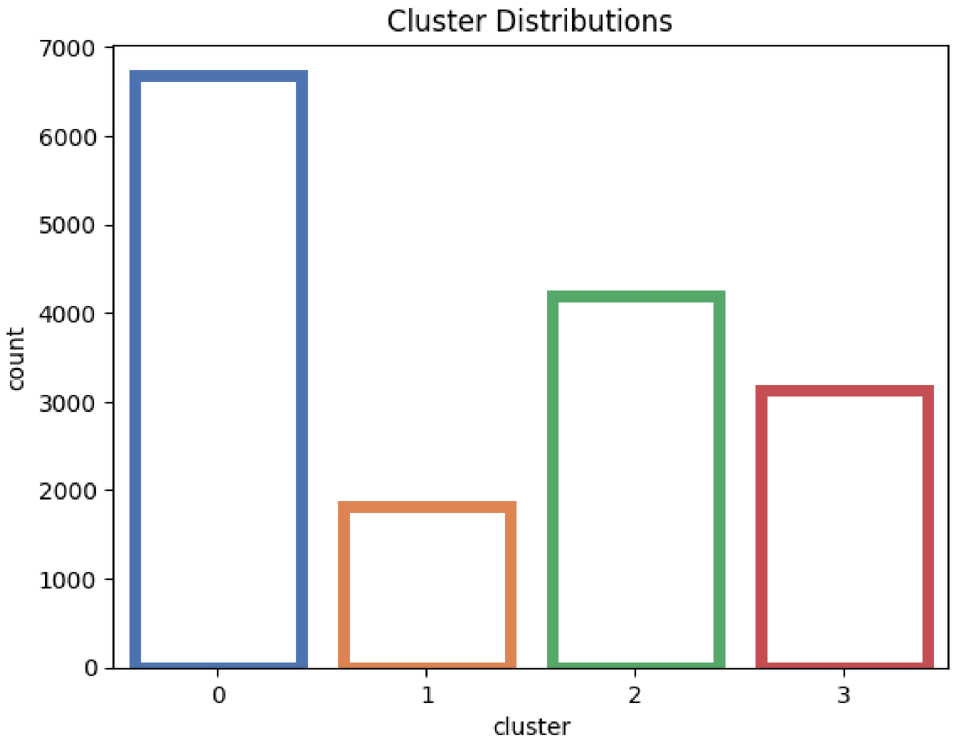


Figure 3.5: Cluster distribution according to K- Means

3.3.2 Implementation of Gaussian Mixture Model Algorithm

The GMM algorithm was another algorithm included in the study because it is an algorithm suitable for working with big data. AIC and BIC scores were used to specify the optimal cluster counts for this algorithm. The point where the AIC and BIC scores are minimum indicates the number of appropriate clusters, but when the calculated scores are examined, it is seen that this algorithm is not compatible for this clustering problem. For this reason, it was not possible to perform segmentation according to the GMM algorithm.

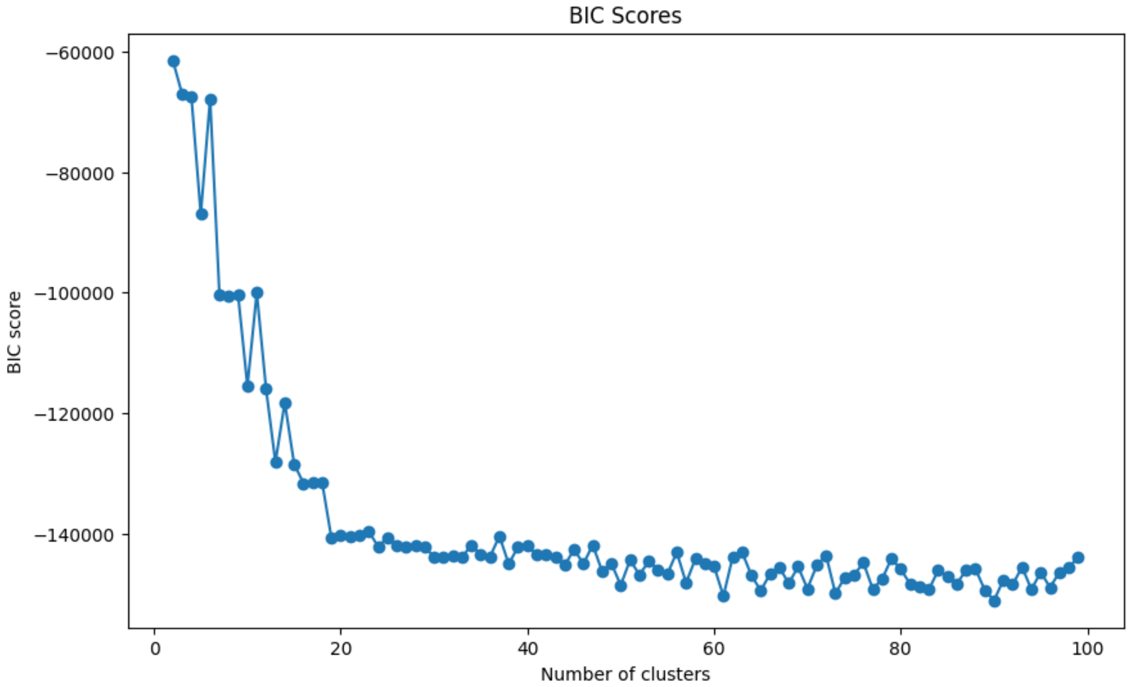


Figure 3.6: BIC Scores

BIC score value was calculated up to k = 100. It was observed that the score was in a downward trend as the number of clusters increased. For this reason, the BIC score was not suitable for specifying the optimum cluster count and the AIC score was estimated.

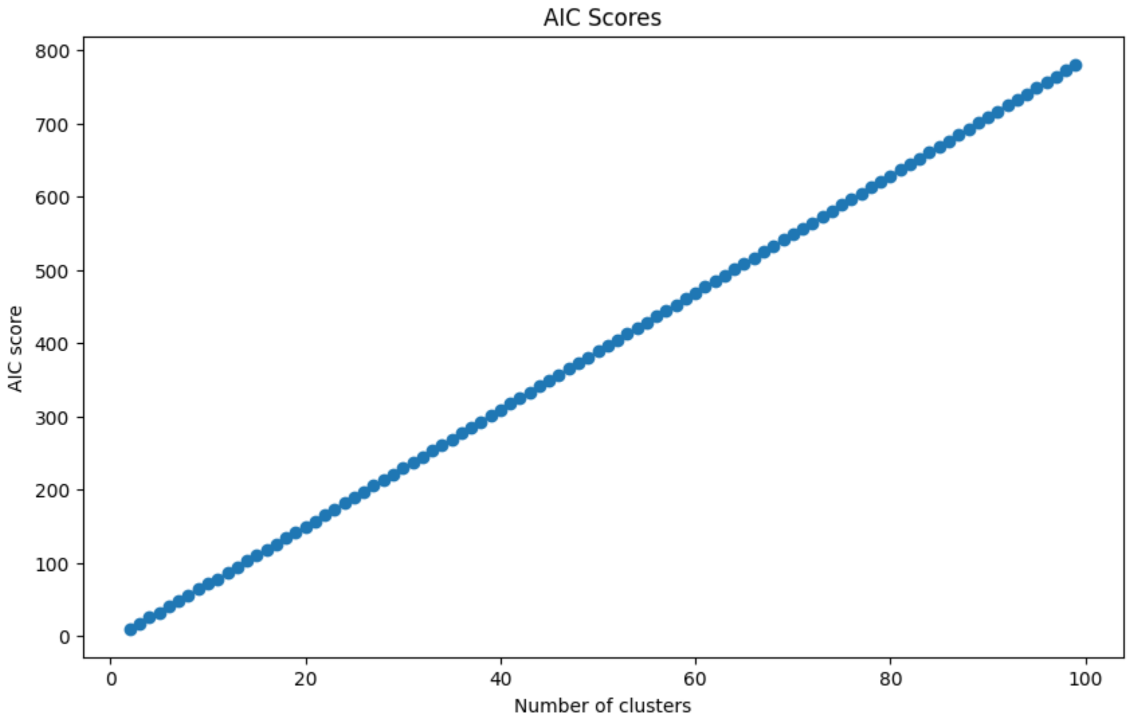


Figure 3.7: AIC Scores

When the AIC score was examined, it was sighted that the AIC score value increased linearly as the number of clusters increased. This scoring pattern also did not help to specify the optimum cluster count. By considering both scores, it was concluded that the GMM algorithm is not suitable for the problem in this study.

3.3.3 Implementation of Fuzzy c - Means Algorithm

Before applying the FCM algorithm, Elbow method was applied to determine the optimal cluster numbers and Silhouette score values were calculated. The cluster count could not be specified clearly from the Elbow method output, but the Silhouette score graph produced a more specific result.

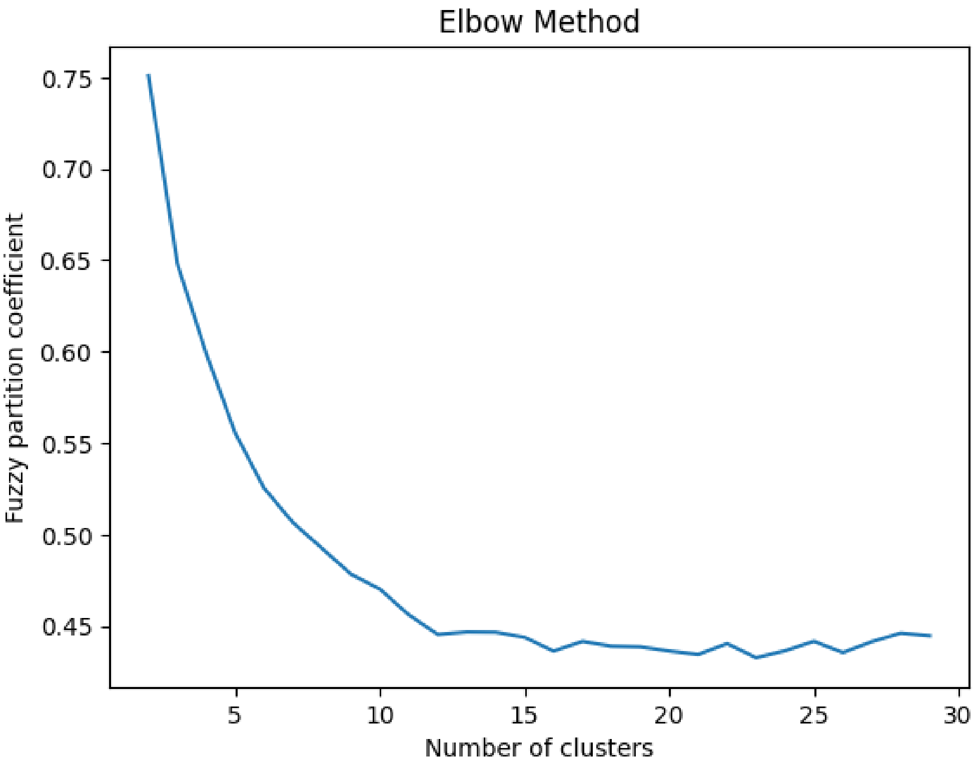


Figure 3.8: Finding optimal cluster count by Elbow Method

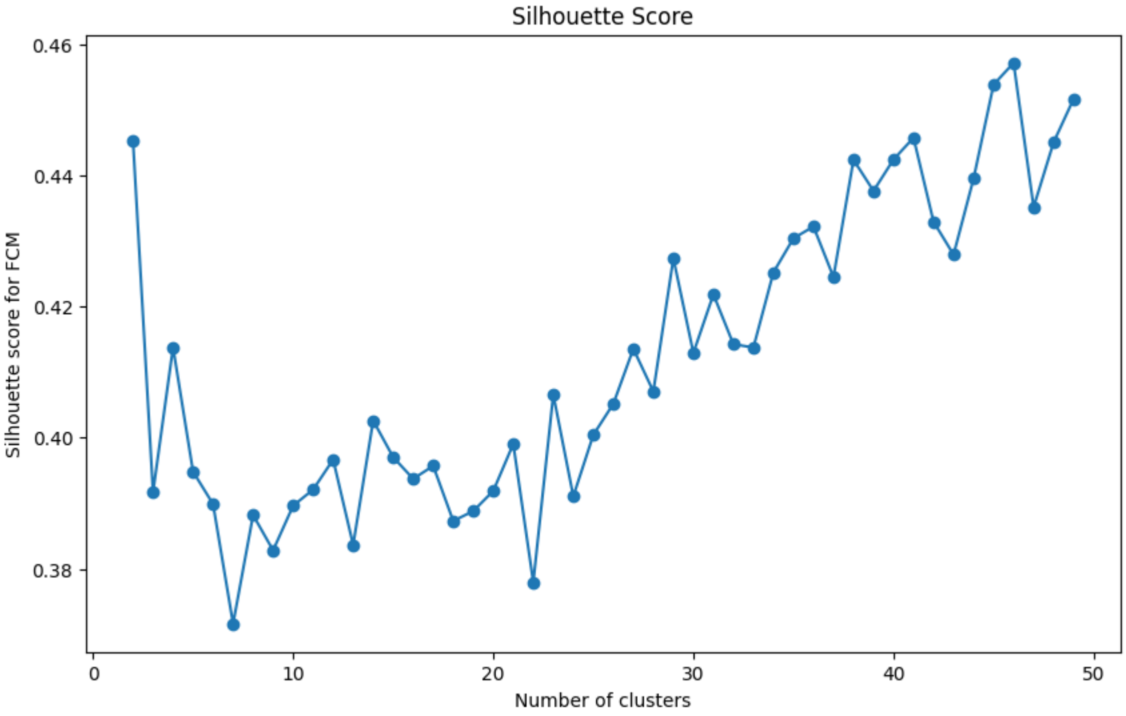


Figure 3.9 : Finding optimal cluster count by Silhouette Score

The closest value of the silhouette score to 1 is in cases where the number of clusters is between 40 and 50. The fact that count of clusters is very high will prevent the manageability of this segmentation problem. When the graph is examined, it has been determined that the most manageable number of clusters is 4.

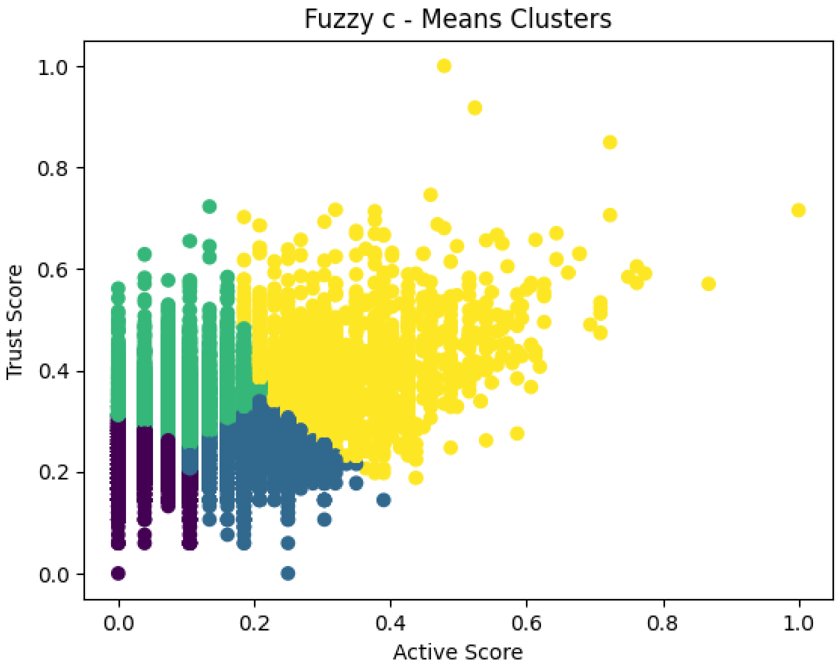


Figure 3.10: Fuzzy c – Means clusters while c = 4

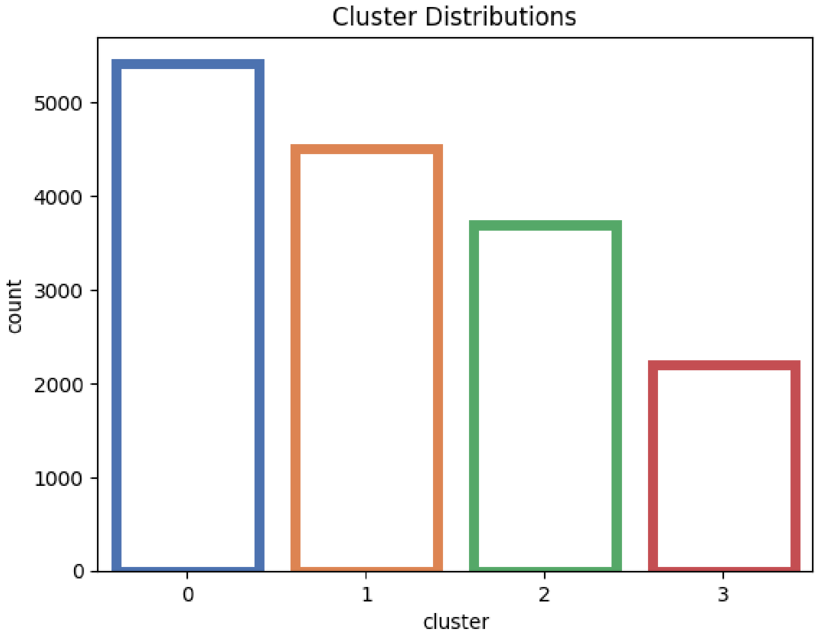


Figure 3.11: Cluster distribution according to Fuzzy c - Means

3.3.4 Implementation of Hierarchical Clustering Algorithm

Since hierarchical clustering is different from other approaches, it is desired to be discussed in this study. Dendrograms were created for 15829 customers and ideal number of clusters was specified as 3. Although the data is 2-dimensional and the number of lines is not very large, dendrogram drawing is very costly in terms of time.

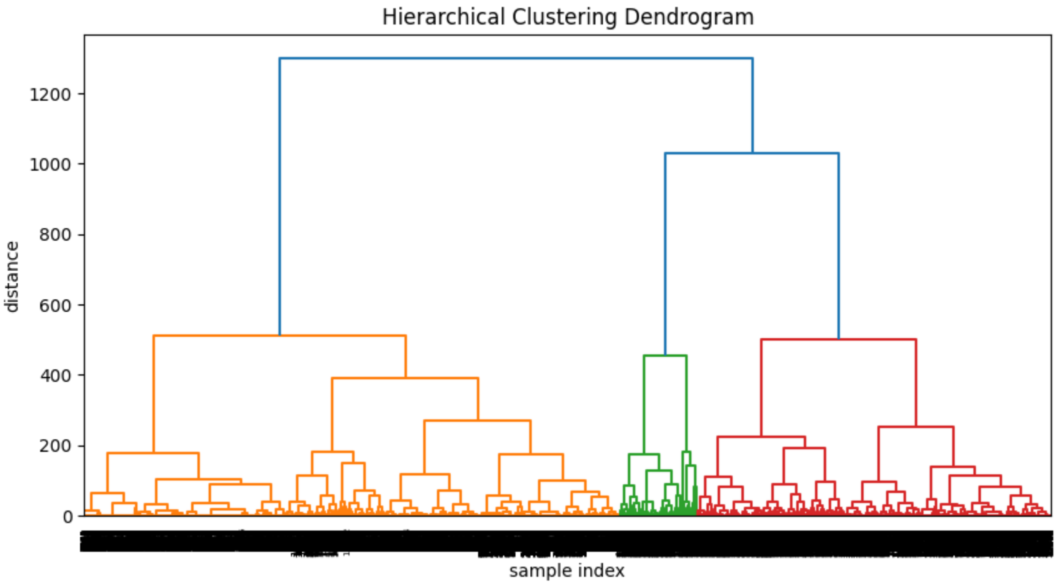


Figure 3.12: Dendrogram for Hierarchical Clustering

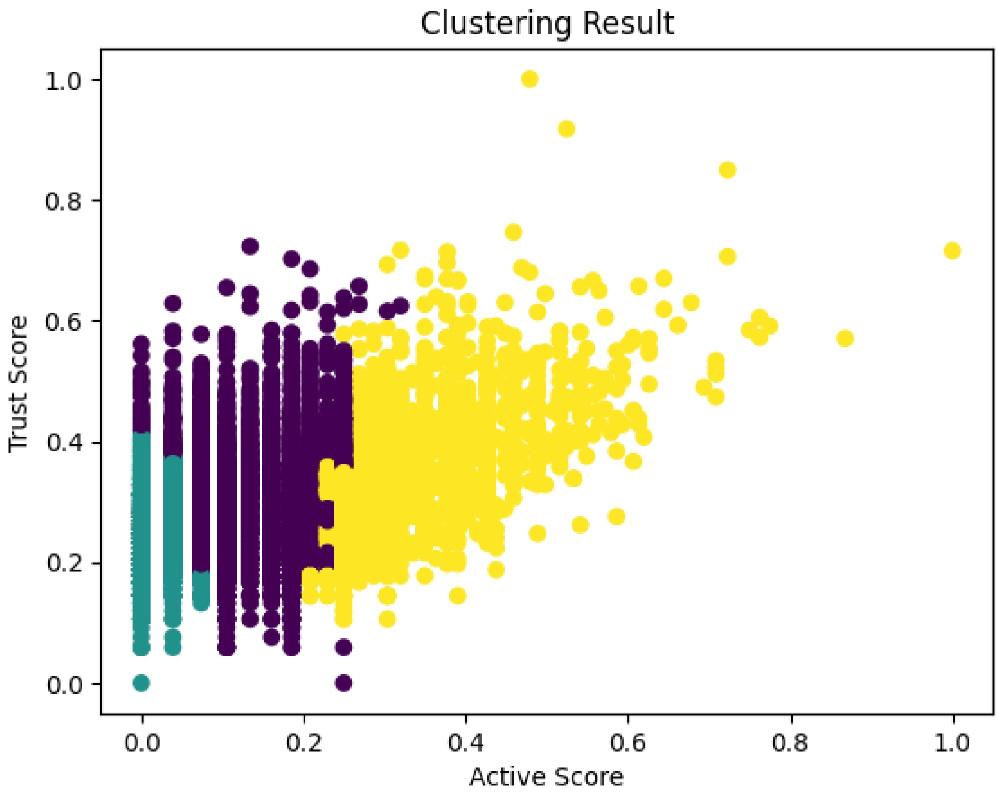


Figure 3.13: Clusters according to Hierarchical Clustering

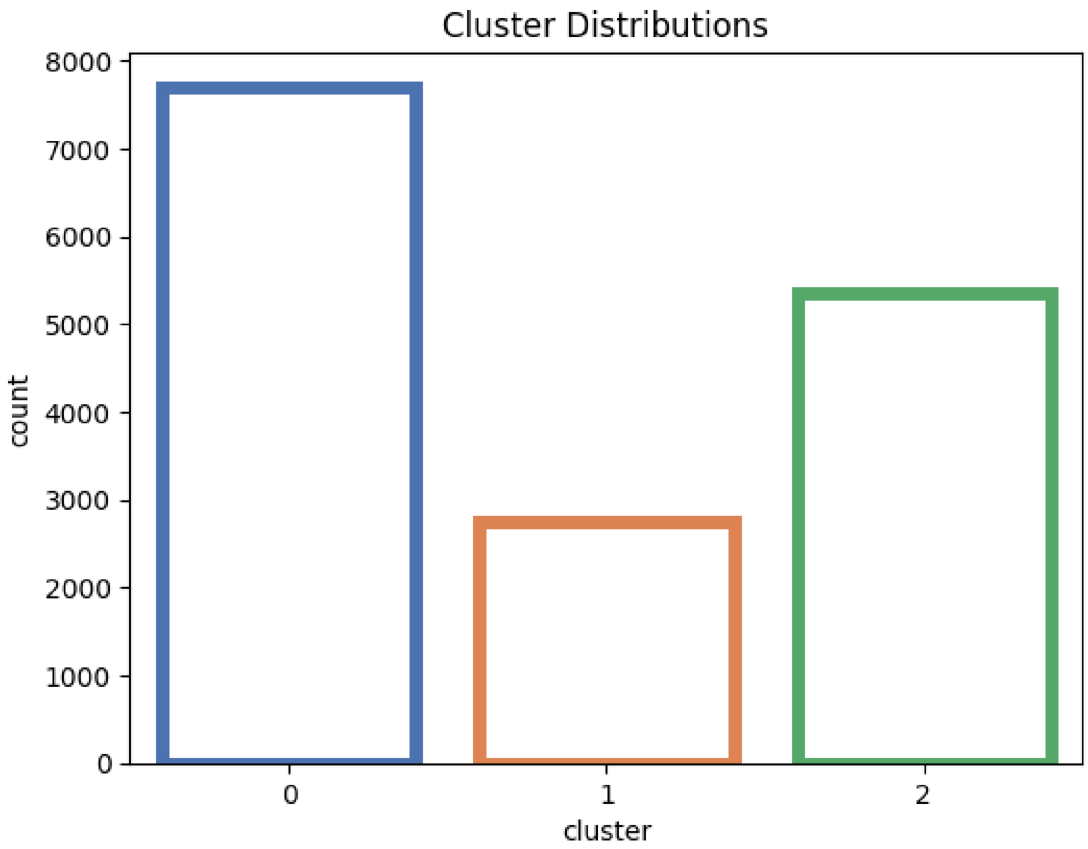


Figure 3.14: Cluster distribution for Hierarchical Clustering

3.3.5 Implementation of DBSCAN Algorithm

Since DBSCAN algorithm is more compatible with small data sets, it was not evaluated in this study because it found the cluster count as 1. Based on the research made due to this study, it can be said that it has been successful in studies with 2–3-dimensional data sets under 10.000 in real life problems in different sectors.

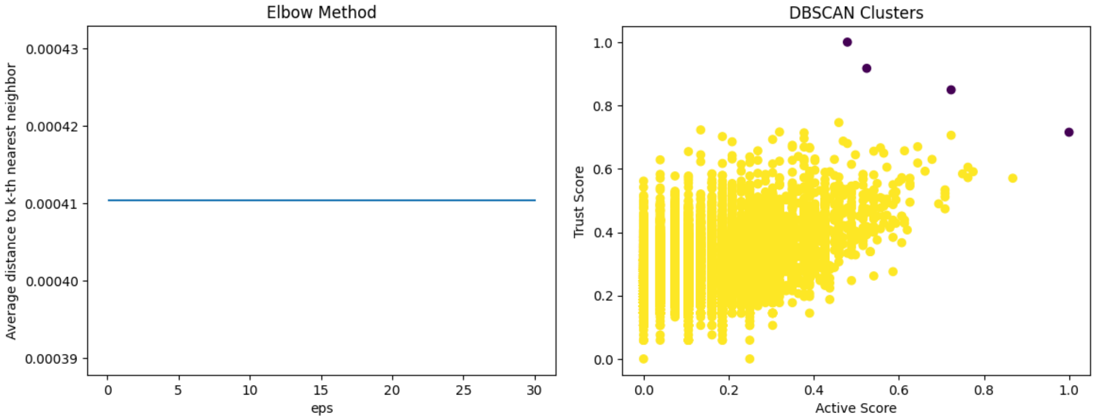


Figure 3.15: Finding the epsilon value with the k-nearest neighbor algorithm

In the figure above, it is observed that there is no separation in the cluster distribution.

3.3.6 Evaluation of the K - Means Algorithm Results

Since GMM and DBSCAN algorithms are not compatible with the current problem and could not produce meaningful outputs, a conclusion could not be reached for customer segmentation with these algorithms. Hierarchical clustering was the costliest method in terms of time and the number of clusters obtained did not help to produce meaningful output. In the FCM and K - means algorithms, the optimum cluster count was calculated equal and since the cluster distributions were approximately similar, the evaluation of the outputs was continued with the K - means algorithm.

For the clusters composed by the K - means, active score and trust score mean values were calculated for each cluster and a pivot table was obtained. Then, statistical methods were used to test whether there was a real difference between the clusters with ANOVA.

Table 3.3 Pivot Table for K – Means Clusters

Cluster	Active Score	Trust Score
0	1.251593	2.422046
1	1.658552	2.769251
2	1.472016	2.458420
3	1.340414	2.688839

Table 3.4 Hypothesis

Hypothesis	
Ho	cluster 1 = cluster 2 = cluster 3 = cluster4
Ha	Al least one of them is different

Table 3.5 ANOVA Test Results for K – Means Clusters

F - statistics	p - value
88.2109751501878	8.272834009763473e-05

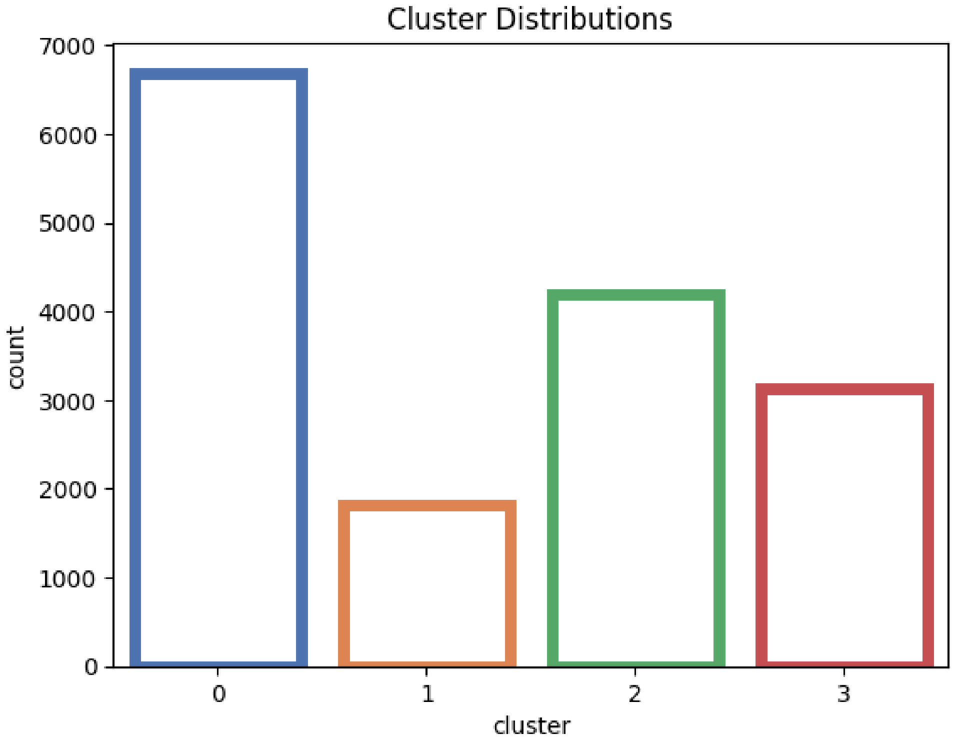
Looking at the result of the ANOVA test, the p value found is a very small value. Statistically, p value less than 0.05 alpha proves that we can reject the null hypothesis. In this case, it can be said that the 4 clusters obtained with K - means are not the same. In the next step, it should be determined which clusters are different from each other. For this purpose, clusters were compared using Bonferroni, one of the pairwise comparison methods.

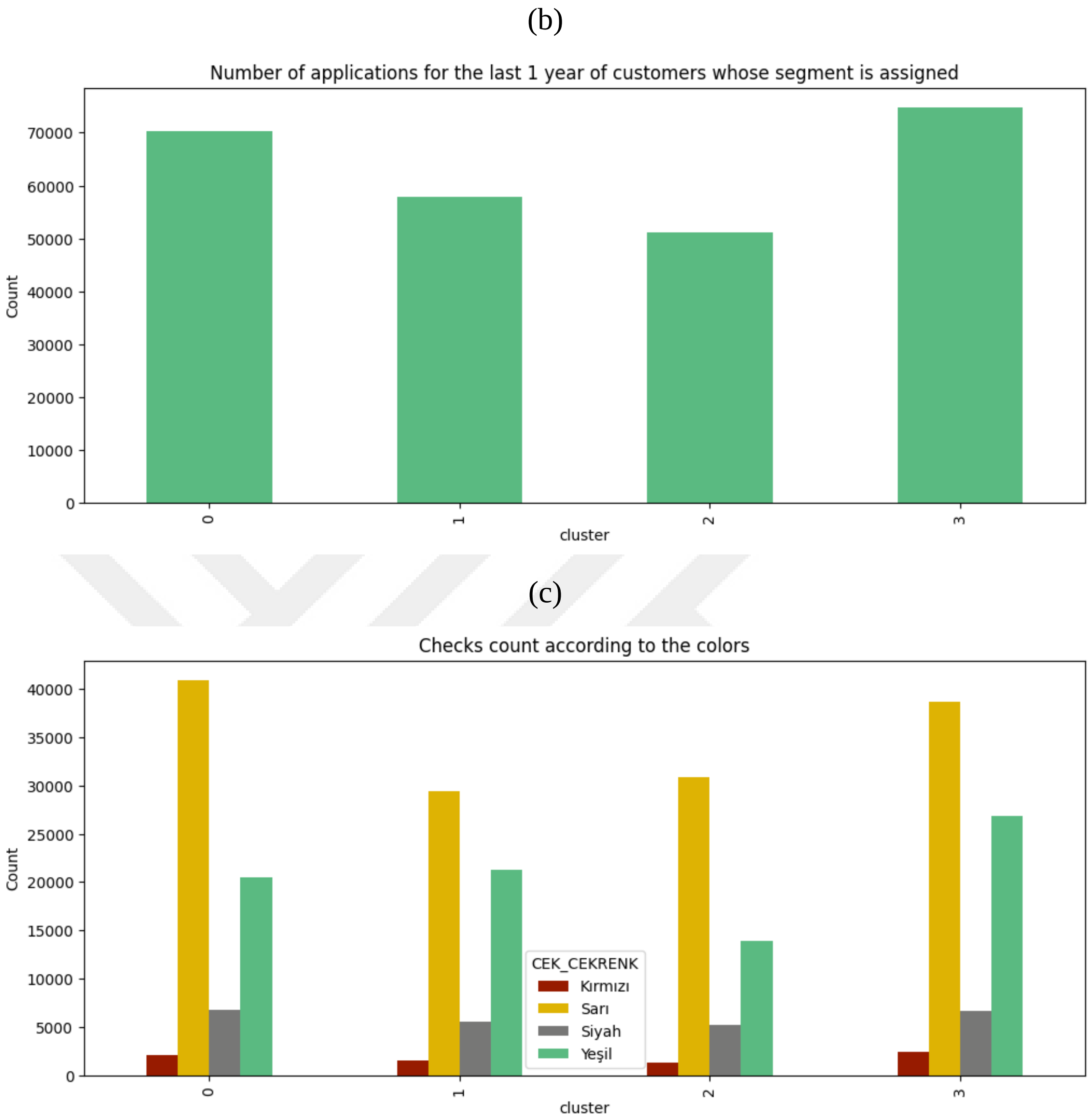
Table 3.6 Pairwise Comparison by Bonferroni Method for the Clusters

<table>
<tr><td colspan="6">Pairwise Comparisons for active_score: Test Multiple Comparison ttest_ind
FWER=0.05 method=bonf - alphacSidak=0.01, alphacBonf=0.008</td></tr>
<tr><th>Group 1</th><th>Group 2</th><th>Stat</th><th>P val</th><th>P val corr</th><th>reject</th></tr>
<tr><td>0</td><td>1</td><td>-208.7339</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>0</td><td>2</td><td>-199.6621</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>0</td><td>3</td><td>-65.6046</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>1</td><td>2</td><td>82.0693</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>1</td><td>3</td><td>114.9379</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>2</td><td>3</td><td>84.8809</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td colspan="6">Pairwise Comparisons for trust_score: Test Multiple Comparison ttest_ind
FWER=0.05 method=bonf - alphacSidak=0.01, alphacBonf=0.008</td></tr>
<tr><th>Group 1</th><th>Group 2</th><th>Stat</th><th>P val</th><th>P val corr</th><th>reject</th></tr>
<tr><td>0</td><td>1</td><td>-120.2896</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>0</td><td>2</td><td>-20.1403</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>0</td><td>3</td><td>-131.0831</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>1</td><td>2</td><td>91.5568</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>1</td><td>3</td><td>21.043</td><td>0.0</td><td>0.0</td><td>true</td></tr>
<tr><td>2</td><td>3</td><td>-97.4976</td><td>0.0</td><td>0.0</td><td>true</td></tr>
</table>

When the data obtained as a result of this comparison were examined, it was seen that the null hypothesis could be rejected for each cluster comparison. As a result, it can be said that each cluster is different from each other. Since the difference between the clusters has been proven statistically, how the clusters are distributed in the data set has been observed and what the 4 clusters mean for the factoring company that owns the data. The charts below are based on the last 1-year data of customers who have been assigned segments. Check applications of customers that were not processed are also included in the charts.

(a)




(b)

(c)

Figure 3.16: Distribution of check applications for the last 1 year

Figure 3.16 in (a) shows the number of customers assigned to clusters. Most customers are in Cluster 0. When the number of assigned customers is compared, it is seen that there is a distribution as Cluster 1 < Cluster 3 < Cluster 2 < Cluster 0. Figure (b) shows the last 1-year distribution of check applications made by customers whose segment is assigned. Based on the application numbers, the equation Cluster 2 < Cluster 1 < Cluster 0 < Cluster 3 can be equated. Figure (c) shows the color distribution of the last 1-year applications. Based on these graphs, it can be said that the average number of applications for cluster 0 customers is low. Cluster 1 customers have the highest average number of applications. Considering customers stability, it is seen that there is a distribution in the form of Cluster 0 < Cluster 2 < Cluster 3 < Cluster 1.

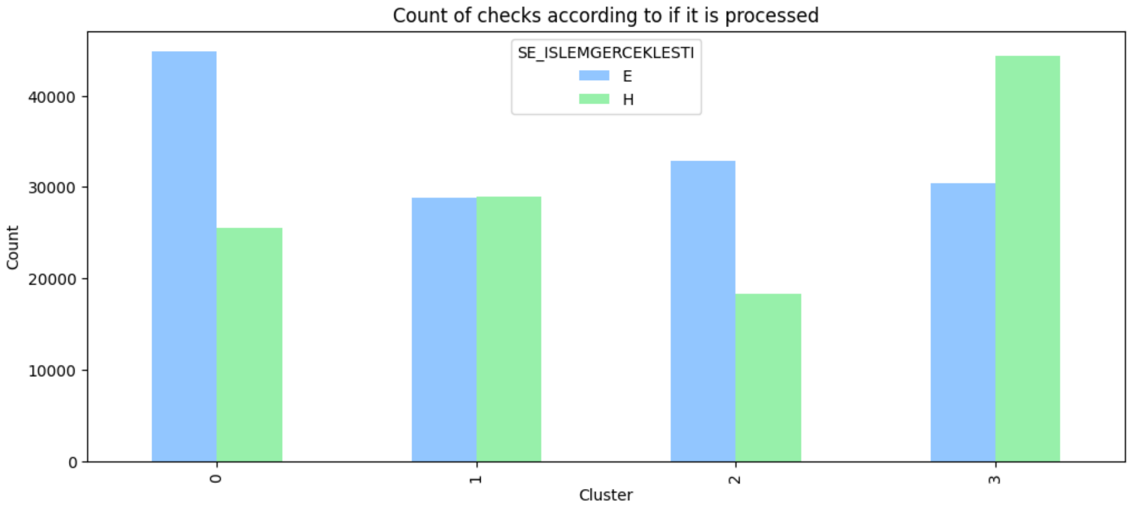


Figure 3.17: Distribution of processed and not processed checks

When the figure above is examined, it is seen that the check applications of cluster 0 customers are mostly processed. Applications of Cluster 1 customers turn into transactions at a rate of approximately 50%. Applications from Cluster 2 customers often turn into transactions, while applications from cluster 3 often do not. The distribution of check applications according to their processed rates is as follows: Cluster 2 > Cluster 0 > Cluster 1 > Cluster3.

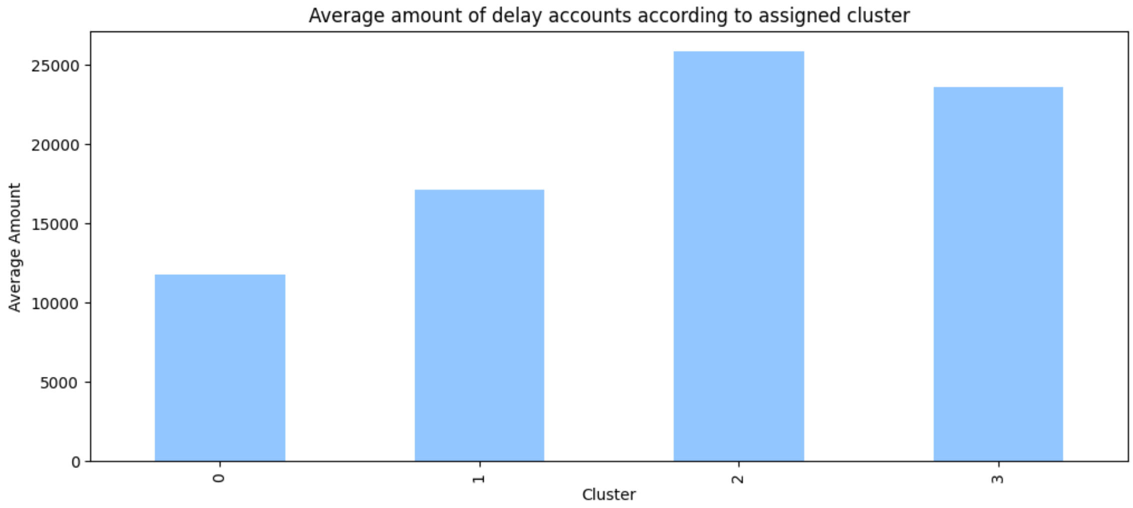


Figure 3.18: Cluster distribution according to arrears averages

The overdue balance indicates a negative situation in terms of financial identity. Cluster 0 customers made a positive impression with a low average overdue balance. Cluster 2 and Cluster 3 customers may represent customers with payment problems.

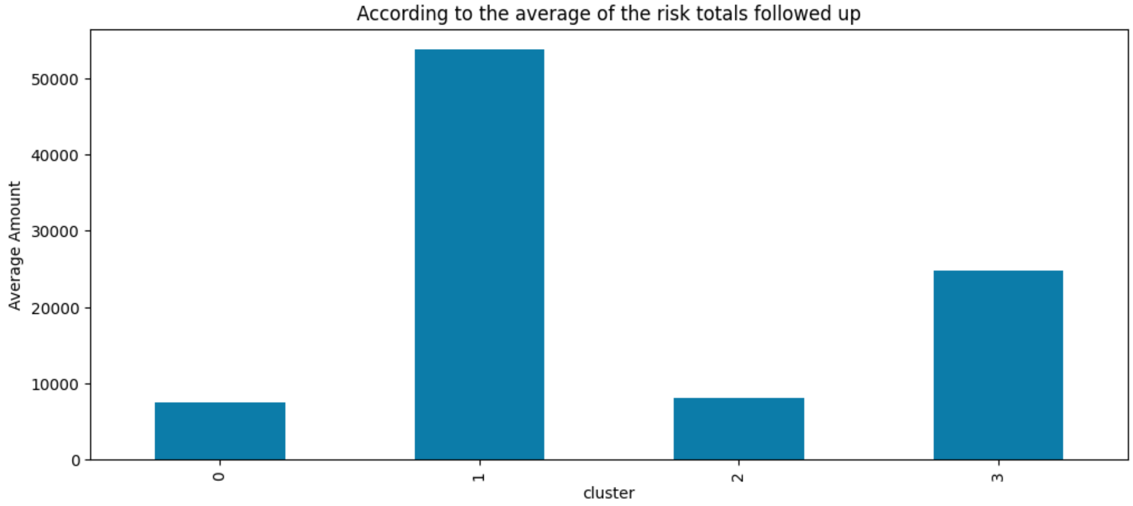


Figure 3.19: Clusters according to their followed-up risks

A high amount of follow-up risk indicates a negative financial situation. If a debt is not paid on time, it is delayed first, and if the delayed debt is still not paid at maturity date which is legally determined period, the customer's case will be followed up. Considering this situation, Cluster 0, and Cluster 2, which have the lowest follow-up risk amount, create a positive impression, while Cluster 1 and Cluster 3 represent financially risky customers.

In addition to this data, the number of institutions where the customer works can also be compared. Working with too many institutions and working with too few institutions may also indicate a negative situation considering the financial information of the customer.

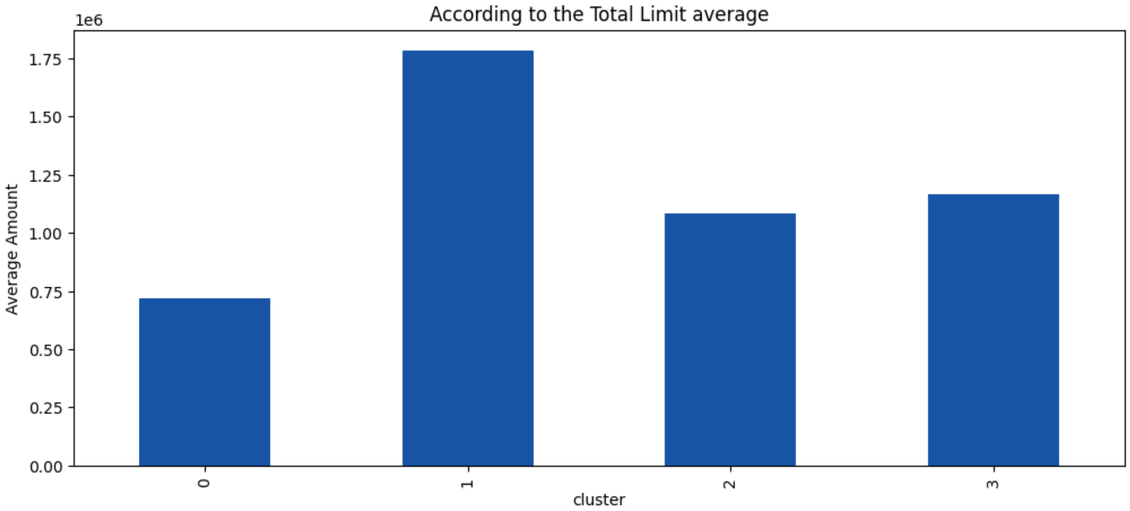


Figure 3.20: Average risk distribution by clusters

It is seen that maximum risk is in cluster 1 customers and the lowest risk is in cluster 0 customers. Cluster 2 has nominal risk, while cluster 3 has high risk. Considering this situation, it is seen that while customers in clusters 0 and 2 create a positive impression, more cautious approach should be taken towards customers belonging to clusters 1 and 3.

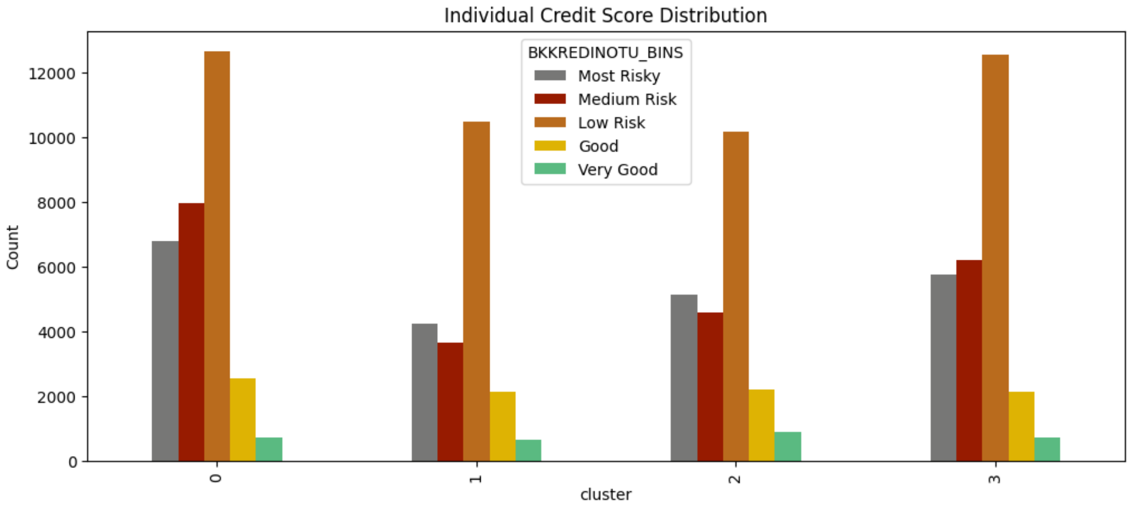


Figure 3.21: Individual credit score distributions

Individual credit score is discussed under 5 main headings. Values of 0 - 699 mean very risky, 700 - 1099 refers medium risk, 1100 - 1499 mean low risk, between 1500 to 1699 good and finally 1700 to 1900 very good. When the check applications made according to the clusters are examined, it is observed that the majority of the applications for each cluster are made by low risk, medium risk and most risky customers. There was any meaningful difference in the distribution among clusters according to individual credit scores.

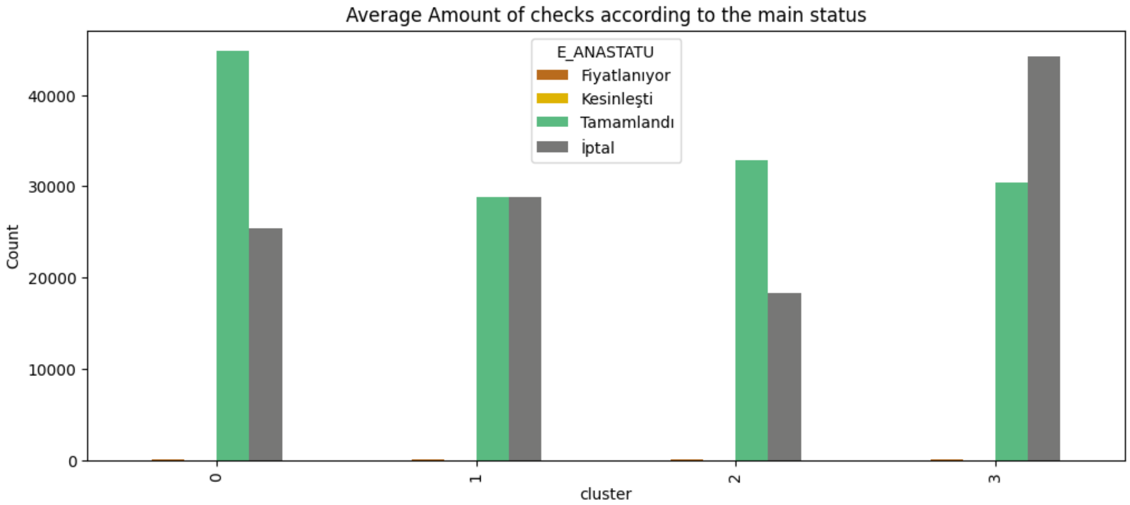


Figure 3.22: Main status distribution of check applications by segment

When check applications are considered according to their main status, it is seen that the cancellation rates of checks from cluster 0 and cluster 2 customers are low. Cluster 3 customers have the highest cancellation rate. The cancellation rate for Cluster 1 customers is approximately half.

4. DISCUSSION

Since the customer is the subject of any shopping, it is the most important factor that adds value to the institutions and can advertise the institutions positively or negatively. So much so that many institutions, including financial institutions, aim to reach their customers correctly through the CRM department within them. Considering that shopping starts with a customer who needs and demands, spending capital for the customer is a very necessary and appropriate approach. CRM departments consist of units dealing with customer segmentation, aiming to reach customers with campaigns or providing customer complaint management. These examples may vary according to the size of the firm and the sectoral needs. In order to reach the right customer at the right time, to win or retain the customer, it is important that the customers are divided into the right segments. Customer segmentation has gained a very important dimension today with developing technology and sectoral competition. Institutions allocate high capital for these studies, both with the teams they have established within their own structure and with the consultancy they receive.

Customer segmentation in financial markets is a slightly more challenging process. Since the spending and payment attitudes of the customer come into play here, for example, giving a credit or credit card to a customer for a bank is a situation that needs to be subject to various checks and approvals. Although the bank carries out these transactions by guaranteeing itself, problem loans and customers also cause extra time and cost. In order to avoid these costs and risks, customer segments are periodically recalculated in order to monitor the changing behavior of the customer.

The fact that the number of parties is higher in the factoring sector has made customer segmentation critical. Even if the customer's financial behavior is risk-free, whether it will be paid on time by the check writer is a matter of risk. Considering that the number of parties in the factoring industry is also high, segmentation studies were carried out on the basis of customer transactions and different algorithms were compared. In this study, 5 different algorithms were used as it is mentioned before. As the final decision, the outputs produced with the K - means algorithm were used and the customers were divided into clusters. Each cluster was examined by considering metrics such as check amount, customer's risk status, transaction realization rate, and rejection rate of check applications. Summary information and suggested segment names of the 4 clusters obtained are given in the table below.

Table 4.1 Factoring Customers Segments

Cluster	Description	Segment
0	The average check amounts → the lowest (- -) Being processed rate → high (+) Overdue balance → the lowest (++) Follow up risk → the lowest (++) Total risk → low (+) Rejection rate → the lowest (++)	OPPORTUNITY
1	The average check amounts → low (-) Being processed rate → low (-) Overdue balance → low (+) Follow up risk → the highest (- -) Total risk → the highest (- -) Rejection rate → high (-)	HEDGE
2	The average check amounts → the highest(++) Being processed rate → the highest (++) Overdue balance → the highest (- -) Follow up risk → low (+) Total risk → low (+) Rejection rate → low (+)	PROFITABLE
3	The average check amounts → high (+) Being processed rate → the lowest (- -) Overdue balance → high (-) Follow up risk → high (-) Total risk → high (-) Rejection rate → the highest (- -)	RISKY

Cluster 0 was named 'Opportunity' because this group refers to customers with the highest number of customers, where the average amount and number of applications are not very high, but with low-risk rates and high processed rates. This gives customers a reliable impression. The risk of customers in this segment is low, and their transaction-based contribution to the factoring company is not very high. It refers to the customer group that carries out regular transactions and is important to retain.

The high-risk averages, low conversion rate and low application amounts were effective in naming 1 as 'Hedge'. Applications belonging to customers in this segment should be handled meticulously in order to prevent possible risks. Considering that the transaction processed rate is 50%, this segment refers to the customers that should be approached with caution.

Cluster 2 has been named as 'Profitable', considering both the high application rates, low risk and high processed rates. These customers not only apply with high check amounts, but also represent the customer group with the highest number of transactions compared to other segments. Retaining these customers is important for corporate profitability. Considering this situation, relations with the customers of this segment should be kept close. Considering the financial situation of the customers in this segment, it can be thought that they have good financial relations with the customers they have worked with.

Cluster 3 has been named 'Risky'. The return rate of checks from customers in this segment is very low. Overdue balances, non-performing risks and overall risks are high. It is seen that customers with payment problems are more in this segment. This may show us that the customer group in this segment also has problems with the customers they work with. It can be thought that customers who have late payments and therefore have difficulty in keeping up with their current payments are in this group.

CONCLUSIONS AND FURTHER WORK

In this thesis, the theoretical background of machine learning algorithms and the evolution of customer segmentation with the development of technology are mentioned. The main purpose is to categorize factoring customers into clusters by using customers all rejected and actioned check applications.

The most accurate customer segmentation is of great importance in our digitalized age. It has become important for them to know their customers and to offer appropriate services to their customers by making appropriate segmentation according to the dynamics of each sector. Since this study contains financial data, information on sectoral approaches was obtained from people working in departments such as credit and analytical CRM in the finance sector. It has been learned that some companies receive support from international consultancy firms for customer segmentation, while some companies experiment with appropriate algorithms for customer segmentation based on sector-specific variables.

In general, it can be said that there are three important steps in examining unlabeled data with ML algorithms. The first is feature selection and dimension reduction, the second is the clustering method, and the last is the interpretation of the clusters with real-life data and making them meaningful. Different algorithms and applications have been proposed for the first two steps. The most critical point here is to identify variables in a way that avoids the curse of multidimensionality and prevents the algorithm from being exposed to over-learning. For the last item, it is possible to evaluate it in different ways from different perspectives.

Actions are taken according to the check colors determined according to the data obtained from the credit risk systems regarding the customer. The purpose of this study is to create a new perspective in evaluating whether a customer's check will be accepted or whether the customer will be persuaded according to the data obtained after clustering. Segmentation study was carried out by running clustering algorithms in the factoring data set. A total of 5 different algorithms were applied to the data set used for this purpose. It has been determined that GMM, DBSCAN algorithms are not suitable for this data set.

The number of clusters determined by hierarchical clustering did not meet the expectation. Since K - means and Fuzzy c - means produce similar outputs, segmentation study was completed using K - means.

As a result, it has been observed that based on the active and trust scores, which are calculated by using check colors and the check counts in 3, 6, 12-month periods , it gives more interpretable results. According to the results obtained no clear distinction was observed in the clusters based on check colors, and individual credit scores, but it was determined that some types of data were more dominant in some clusters. The result to be obtained here is that it is more accurate to interpret on the combinations of variables formed with each other rather than evaluating on a single variable in complex data such as financial data and in which internal and external factors are very active. For this purpose, the clusters assigned based on the calculated activity score and trust score were examined. In this way, more consistent and interpretable results were produced with the data.

In future studies, different segments can be created by re-examining clusters over other important parameters or combinations of parameters. Customer based segmentation can be re-evaluated using different customer metrics. While new customer metrics implemented, outputs that show better performance and more realistic segments can be produced by negotiating with the units that are working with this data. According to the determined segments, future behavior information that has been removed from the data set as it is not used for this study can be taken into account, and the churn tendencies of the customers can be determined according to the check data. Necessary actions can be taken for the customers to be retained regarding this churn analysis. The change in segment data assigned to customers periodically also provides information for the customer's further behavior. Here, since the period information will be very variable between the sector and even among the institutions in the same sector, the segment changes in the periods to be determined according to the internal dynamics of the institution can be evaluated.

REFERENCES

[1] G. S. A. &. T. E. Livne, “Do Customer Acquisition Cost, Retention and Usage Matter to Firm Performance and Valuation?,” *Journal of Business Finance & Accounting,* pp. 334 - 363, 2011.

[2] F. F. Reichheld, Loyalty Rules How Today's Leaders Lasting Relationships, Boston, Massachusetts: Harvard Business School Press, 2001.

[3] C.-H. Xie, J.-Y. Chang and Y.-J. Liu, “Estimating the number of components in Gaussian mixture models adaptively for medical image,” *Optik,* vol. 124, no. 23, pp. 6216-6221, 2013.

[4] H. Chen, R. H. L. Chi and V. C. Storey, “Business Intelligence and Analytics: From Big Data to Big Impact,” *MIS Quarterly,* vol. 36, no. 4, pp. 1165 - 1188, December 2012.

[5] J. Manyika, M. Chui, B. Brown, J. Bughin, R. C. Dobbs, Roxburgh and A. H. Byers, Big data: The next frontier for innovation, competition, and productivity, 2011.

[6] J. R. Bult and T. Wansbeek, “Optimal Selection For Direct Mail,” *Marketing Science,* pp. 378 - 394, 1995.

[7] G. W. Milligan, “An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms,” *PSYCHOMETRIKA,* 1980.

[8] P. (. Balakrishnan, M. C. Cooper, V. S. Jacob and P. A. Lewis, “Comparative performance of the FSCL neural net and K-means algorithm for market segmentation,” *European Journal of Operational Research,* pp. 346 - 357, 1996.

[9] U. F. P. Bradley, “Refining Initial Points for K-Means Clustering,” in *International Conference on Machine Learning,* 1998.

[10] H. Zhexue, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery,* vol. 2, p. 283–304, 1998.

[11] D. Pelleg and A. Moore, “X-means: Extending k-means with efficient estimation of the number of clusters,” Pittsburgh, 2000.

[12] M. Steinbach, G. Karypis and V. Kumar, “A Comparison of Document Clustering Techniques,” Department of Computer Science and Egineering, University of Minnesota, Minneapolis, USA, 2000.

[13] A. Likas, N. Vlassis and J. J. Verbeek, “The Global k-Means Clustering Algorithm,” *Pattern Recognition,* vol. 36, pp. 451-461, 2003.

[14] H. Hwang, T. Jung and E. Suh, “An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry,” *Expert Systems with Applications,* vol. 26, pp. 181 - 188, 2004.

[15] T. Hong and E. Kim, “Segmenting customers in online stores based on factors that affect the customer’s intention to purchase,” *Expert Systems with Applications,* vol. 39, pp. 2127-2131, 2012.

[16] S. Ghosh and S. K. Dubey, “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms,” *International Journal of Advanced Computer Science and Applications,* vol. 4, no. 4, pp. 35 - 39, 2013.

[17] B. Doğan, A. Buldu, Ö. Demir and B. E. Ceren, “Using Clustering Analysis for Customer Relationship Management in Insurance Sector,” *Karaelmas Fen ve Mühendislik Dergisi,* vol. 8, pp. 11-18, 2018.

[18] M. Waheed, S. Hussain, A. A. Khan, M. Ahmed and B. Ahmad, “A methodology for image annotation of human actions in videos,” Springer, 2020.

[19] İ. Kabasakal, “Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing,” *Bilişim Teknolojileri Dergisi,* vol. 1, January 2020.

[20] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, 1967.

[21] M. Ahmed, R. Seraj and S. M. S. Islam, “The k-means Algorithm: A Comprehensive Survey and Performance Evaluation,” MDPI, 2020.

[22] O. B. Koca, “Determining customer segmentation and behaviour models with database marketing and machine learning,” *Journal of Management, Marketing and Logistics,* vol. 8, no. 2, pp. 89-111, 2021.

[23] Y. Zhang, M. Li, S. Wang, S. Dai, L. Luo, E. Zhu, H. Xu, X. Zhu, C. Yao and H. Zhou, “Gaussian Mixture Model Clustering with Incomplete Data,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, 2021.

[24] G. Baloğlu, “A Model Recommendation For Risk Identification In Audit Of Factoring Transactions,” *Denetişim,* vol. 24, pp. 134 - 158, January 2022.

[25] M. Yazıcı, “Finansal Piyasalarda Risk ve Risk Yönetimi,” in *Finansal Piyasalar ve Kurumlar,* 2020, pp. 189-212.

[26] W. R. Smith, “Product Differentiation and Market Segmentation as Alternative Marketing Strategies,” *Journal of Marketing,* vol. 21, no. 1, pp. 3-8, July 1956.

[27] T. J. Brock, “Pareto Principle,” 25 December 2020. [Online]. Available: https://www.investopedia.com/terms/p/paretoprinciple.asp.

[28] U. Özmen, “Uzaktan CRM Eğitimi,” 12 March 2012. [Online]. Available: https://www.uzaktancrmegitimi.com/4174/musteri-odaklilik-nedir. [Accessed 15 November 2022].

[29] W. Chang, C. Chang and Q. Li, “Customer Lifetime Value: A Review,” *Social Behavior and Personality An International Journal* , vol. 40, no. 7, pp. 1057 - 1064, August 2012.

[30] D. Jackson, “Determining a customer’s lifetime value.,” *Direct Marketing,* vol. 51, pp. 60 - 63, 1989.

[31] J. A. McCarty and M. Hastak, “Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression,” *Journal of Business Research,* vol. 60, pp. 656 - 662, 2007.

[32] E. Suh, K. Noh and C. Suh, “Customer list segmentation using the combined response model,” in *Expert Systems with Applications,* 1999.

[33] C.-H. Cheng and Y.-S. Chen, “Classifying the segmentation of customer value via RFM model and RS theory,” *Expert Systems with Applications,* vol. 36, pp. 4176 - 4184, 2009.

[34] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain and A. J. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” in *Supervised and Unsupervised Learning For Data Science,* Springer, 2019.

[35] J. Delua, “IBM Cloud,” IBM, 12 March 2021. [Online]. Available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning. [Accessed 16 March 2023].

[36] G. C. M.-L. M.-M. Cathy Maugis, “Variable Selection for Clustering with Gaussian Mixture Models,” *Biometrics,* vol. 65, pp. 701-709, September 2009.

[37] J. Morimoto and F. Ponton, “Virtual reality in biology: could we become virtual naturalists?,” Springer, 2021.

[38] Sajeev B. U and K. Thangavel, “Assessment of Financial Status of SHG Members: A,” *International Journal of Computer Applications,* vol. 32, October 2011.

[39] M. S. A. K. V. K. P.-N. Tan, “Cluster Analysis: Basic Concepts and Algorithms,” in *Introduction to Data Mining,* 2018, pp. 535, 525-612.

[40] P. Berkhin, “Survey of Clustering Data Mining Techniques,” Accrue SoftwareInc, San Jose,California, 2002.

[41] H.-H. Bock, “A History of k-Means Algorithms,” in *Clustering Methods*.

[42] C. Maklin, “Towards Data Science,” 19 7 2019. [Online]. Available: https://miro.medium.com/v2/resize:fit:640/format:webp/1*D2nunNrckTdV9n7gTUC4Zg.png;https://miro.medium.com/v2/resize:fit:640/format:webp/1*eTAFs5cTUjb_kt-4RgE3uw.png. [Accessed 05 02 2023].

[43] A. P. Sempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from incomplate data via the EM Algorithm,” Harvard University, 1976.

[44] T. K. Moon, “The Expectation Maximization Algorithm,” *IEEE Signal Processing Magazine,* vol. 13, no. 6, pp. 47 - 60, 11 1996.

[45] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *International Conference on Knowledge Discovery and Data Mining,* 1996.

[46] A. Moreira, M. Y. Santos and S. Carneiro, “Density-based clustering algorithms – DBSCAN and SNN,” University of Minho, Portugal, 2005.

[47] F. Nielsan, “Hierarchical Clustering,” in *Introduction to HPC with MPI for Data Science,* Springer, 2016.

[48] s.-l. developers, “Scikit Learn,” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering.

[49] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, M. Nadler, Ed., Logan, Utah: Springer, 1981.

[50] J. C. Dunn, “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters,” *Journal of Cybernetics,* vol. 3, no. 3, pp. 32 - 57, 1973.

[51] J. V. d. Oliveira and W. Pedrycz, Advances in Fuzzy Clustering and its Applications, John Wiley & Sons Ltd., 2007.

[52] T. A. Tran, “A study of the energy efficiency management for bulk carriers considering navigation environmental impacts,” IOS Press, 2018.

[53] A. Atcılı, “Machine Learning Distances,” 3 January 2022. [Online].

[54] “Diagnostic Cluster Analysis of Mathematics Skills,” *Ieri Monograph Series: Issues And Methodologies In Large-Scale Assessments,* vol. 4, pp. 75-107, 2011.

[55] “Nationat Institute of Standards and Technology,” Nationat Institute of Standards and Technology, 11 02 2019. [Online]. Available: https://xlinux.nist.gov/dads/HTML/manhattanDistance.html. [Accessed 19 03 2023].

[56] “Medium,” [Online]. Available: https://cdn-images-1.medium.com/v2/resize:fit:800/1*-xXnL0liqSl-flWgCTFbiw.png. [Accessed 19 03 2023].

[57] N. Altman and M. Krzywinski, “The curse(s) of dimensionality,” *Nature Methods* |, vol. 15, pp. 397 - 400, 2018.

[58] R. E. A. Bellman, Adaptive Control Processes: A Guided Tour, Princeton, New Jersey: Princeton Univ. Press, 1961.

[59] K. Yıldız, Y. Çamurcu and B. Doğan, “A Comperative Analize of Principal Component Analysis and Non-Negative Matrix Factorization Techniques in Data Mining,” in *Akademik Bilişim*, Muğla, 2010.

[60] H. Guo, H. Ayalew, A. Seethepalli, K. Dhakal, M. Griffiths, X.-F. Ma and L. M. York, “Functional phenomics and genetics of the root economics space in winter

wheat using high-throughput phenotyping of respiration and architecture," Noble Research Institute, Sam Noble Parkway, Ardmore, 2020.

[61] L. v. d. Maaten, "Accelerating t-SNE using Tree-Based Algorithms," *Journal of Machine Learning Research,* vol. 15, pp. 1 - 21, 2014.

[62] D. Marutho, S. H. Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News," in *International Seminar on Application for Technology of Information and Communication,* 2018.

[63] M. Forina, C. Armanino and V. Raggio, "Clustering with dendrograms on interpretation variables," Elsevier, 2001.

[64] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics,* vol. 6, no. 2, pp. 464 - 464, March 1978.

[65] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *Wiley Periodicals,* vol. 4, pp. 199 - 203, March - April 2012.

[66] A. Mehrjou, R. Hosseini and B. N. Araabi, "Improved Bayesian Information Criterion for Mixture Model Selection," *Pattern Recognition Letters,* vol. 69, pp. 22 - 27, 2016.

[67] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *Wires Computational Statistics,* 2019.

APPENDIX A: DATASET VARIABLES

Table A.1 : Lists of variables and their datatypes.

COLUMN_NAME	EXPLANATION	TYPE
se_ttdtarih	DateOfProcess	object
se_tamamlananislemadet	CompletedProcessCount	int64
se_kurulustarihi	FoundingDate	object
se_islemgerceklesti	IsProcessCompleted	object
se_musteriilkislem	IsFirstProcessOfCustomer	object
se_012_limittahsis	IsLimitAllocation	object
se_tkbknakdilimit	CashLimitOfCompany	int64
se_tkgayrinakdirisk	NonCashRiskOfCompany	int64
varlikturu	RealEstateOrPortable	object
krmsonuc	IsCreditRiskCenterInfoFound	object
tkkurumsayisi	CountOfCompany	int64
tkilkkredikullandirimtarihi	FirstCreditUsageDateOfCompany	object
tksonkredikullandirimtarihi	LastCreditUsageDateOfCompany	object
tkgecikmedekihesapsayisi	NumberOfAccountsInDelay	int64
tkgecikmisbakiyetoplami	TotalOverdueBalanceOfCompany	int64
tktoplamlimit	TotalLimitOfCompany	int64
tktoplamrisk	TotalRiskOfCompany	int64
tktoplamnakdilimit	TotatlCashLimitOfCompany	int64
tktoplamnakdirisk	TotatlCashRiskOfCompany	int64
tkgayrinakdilimit	NonCashLimitOfCompany	int64
tkgayrinakdirisk	NonCashRiskOfCompany	int64
tktakhesbilbulfinkrmsay	FollowUpAccountCount	int64
tktakipalrisktoplam	TotalRiskFollowed	int64
tkguncelkredibakiyesitoplami	CurrentLoanBalanceOfCompany	int64
tkenyakintakiptarihi	NearestFollowUpDateofCompany	object
tkeneskitakiptarihi	OldestFollowUpDateOfCompany	object
tkenguncellimittahsistarihi	CurrentLimitAllocationDate	object
krsveribulundumu	IsCreditReferenceSystemInfoFound	object
bksonkredikullandirimtarihi	LastCreditUsageDateOfCustomer	object
bkkurumsayisi	CompanyCountOfCustomer	int64
bkkredilihesapsayisi	CountOfAccountWithCredit	int64
bktoplamlimit	TotalLimitOfCustomer	int64

bktoplamrisk	TotalRiskOfCustomer	int64
bkgecikmedekihesapsayisi	CountOfAccountInDelay	int64
bkgeciktirdigibakiyetoplami	TotalOverdueBalanceOfCustomer	int64
bkmevcutenuzungecikmesuresi	LongestDelayTime	int64
bktakipbilbulkrmsay	FollowUpCompanyCountOfCustomer	int64
bkkredinotu	CreditScoreOfCustomer	float64
tktokfaktoringkredilimiti	FactoringCreditLimit	float64
tktokfaktoringcalbanvedigkrm	FactoringBankOtherInstutionCount	int64
bksonkanunitakibealinmatarihi	LastLegalProceedingDate	object
cek_istihbaratsonuc	CheckInformationResult	object
cek_tutar	CheckAmount	float64
cek_cekvade	CheckMaturity	object
cek_cekvadegunsayisi	DateToCheckMaturity	int64
cek_cekortvadegunsayisi	AverageMaturityDayOfCheck	float64
cek_kararsonuc	CheckDecisionResult	object
cek_kararstatu	CheckDecisionStatu	object
cek_cekskor	CheckScore	float64
cek_cekrenk	CheckColor	object
cek_pre_cekrenk	CheckPreColor	object
cek_cutoff_cekrenk	CheckCutOffColor	object
e_cekstatu	CheckStatu	object
e_istdurum	InformationStatu	object
e_istonay	InformationApproval	object
e_istsononay	InformationLastApproval	object
e_islemsegment	ProcessSegment	object
e_anastatu	MainStatu	object
e_faizoran	InterestRate	object
ksd_kesidecitip	DrawerType	object
stc_gercektuzel	IndividualOrCorporate	object
stc_ktutar	SellerCreditAmount	float64
stc_kararfirmasegmenttxt	SellerSegmentDecision	object
stc_musterilimiti	SellerLimit	float64
stc_musteririski	SellerCustomerRisk	float64
stc_risk	SellerRisk	float64
cek_pre_cekskor	CheckPreSkor	float64
cek_kioscekrenk	CheckKioskColor	object

cek_katki	CheckConribution	float64
toplam_katki	TotalContribution	float64
son_12_ay_katki	Last12MonthContribution	float64
toplam_ttd	TotalCheckCountOfCustomer	float64
ilk_tddtarih	FirstProcessDate	object
son_ttdtarih	LastProcessDate	object
gecmistarihli_ttd_tum	TotalPastCheckCount	float64
toplam_islem	TotalProcessCount	float64
ilk_islemtarih	FirstProcessDate	object
son_islemtarih	LastProcessDate	object
gecmistarihli_islem_tum	TotalPastProcessedCheckCount	float64
toplam_ttd_k	TotalProcessCountOfDrawer	int64
ilk_tddtarih_k	FirstProcessDateOfDrawer	object
son_ttdtarih_k	LastProcessDateOfDrawer	object
gecmistarihli_ttd_tum_k	TotalPastCheckCountOfDrawer	int64
toplam_islem_k	TotalProcessedCheckCountOfDrawer	int64
ilk_islemtarih_k	FirstProcessedDateOfDrawer	object
son_islemtarih_k	LastProcessedDateOfDrawer	object
gecmistarihli_islem_tum_k	TotalPastProcessedCheckCountOfDrawer	int64
banka6_ayodemetutar	Bank6MonthPaymentAmount	float64
banka6_limitfark	Bank6MonthLimitDifference	float64
banka6_limitartis	Bank6MonthLimitIncrease	float64
banka6_limitdusus	Bank6MonthLimitDecrease	float64
faktoring6_ayodemetutar	Factoring6MonthPaymentAmount	float64
faktoring6_limitfark	Factoring6MonthLimitDifference	float64
faktoring6_limitartis	Factoring6MonthLimitIncrease	float64
faktoring6_limitdusus	Factoring6MonthLimitDecrease	float64
sorunlu6_ayodemetutar	BadLoan6MonthPaymentAmount	float64
sorunlu6_limitfark	BadLoan6MonthLimitDifference	float64
sorunlu6_limitartis	BadLoan6MonthLimitIncrease	float64
sorunlu6_limitdusus	BadLoan6MonthLimitDecrease	float64
yapilandirma6_limitdusus	Restructuring6MonthLimitDecrease	float64
diger6_ayodemetutar	Other6MonthPaymentAmount	array
diger6_limitfark	Other6MonthLimitDifference	array
diger6_limitartis	Other6MonthLimitIncrease	array

diger6_limitdusus	Other6MonthLimitDecrease	array
gecenyilodenen	PaymentOfLastYear	array