



AN APPLICATION FOR THE EVALUATION OF CLUSTERING ANALYSIS IN DATA MINING

TANZER AKTAŞ

Master's Thesis

Graduate School

Izmir University of Economics

İzmir

2020

**AN APPLICATION FOR THE EVALUATION OF
CLUSTERING ANALYSIS IN DATA MINING**



TANZER AKTAŞ

A Thesis Submitted to

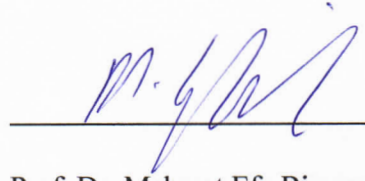
The Graduate School of Izmir University of Economics

Master Program in Industrial Engineering

İzmir

2020

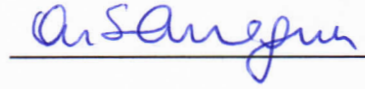
Approval of the Graduate School



Prof. Dr. Mehmet Efe Biresselioğlu

Director

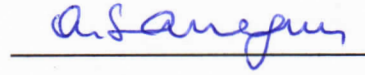
I certify that this thesis satisfies all the requirements as a thesis for a Master's degree.



Prof. Dr. Ahmet Sermet Anagün

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for a Master's degree.



Prof. Dr. Ahmet Sermet Anagün

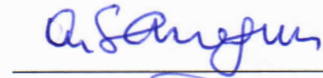
Supervisor

Master's Exam Jury Members

Prof. Dr. Ahmet Sermet Anagün

Asst. Prof. Dr. İlker Gölcük

Prof. Dr. Gözde Yazgı Tütüncü



ABSTRACT

AN APPLICATION FOR THE EVALUATION OF CLUSTERING ANALYSIS IN DATA MINING

Aktaş, Tanzer

M.S. in Industrial Engineering

Supervisor: Prof. Dr. Ahmet Sermet ANAGÜN

Ocak 2020

Data mining techniques have been developed recently and are being used in many fields. Developing information technology tools, both software and hardware, have an important role to play. Data mining used to reveal confidential, valuable, usable information from a large amount of data with developing technology tools and provide strategic decision support; has been able to find answers to problem areas related to large amounts of data.

The use of clustering analysis, one of the data mining methods, has increased in recent years. With the increase in marketing, biology, banking, insurance, stock exchange, retailing, telecommunications, genetics, health, science and engineering, criminology, health, industry, intelligence, education and so on. applications are seen in many branches. In this study, with the help of item analysis, a clustering analysis application with the data of an educational institution was made by using K - Means Algorithm and Hierarchical Clustering Methods.

Keywords: Data Mining, Cluster Analysis, K-Means Algorithm, Hierarchical Clustering, Item Analysis



ÖZET

VERİ MADENCİLİĞİNDE KÜMELEME ANALİZİNİN DEĞERLENDİRİLMESİ İÇİN BİR UYGULAMA

Aktaş, Tanzer

Endüstri Mühendisliği Yüksek Lisans Programı (Tezli)

Tez Yöneticisi: Prof. Dr. Ahmet Sermet ANAGÜN

Ocak 2020

Veri madenciliği teknikleri son zamanlarda gelişmiş ve bir çok alanda kullanılmaya başlanmıştır. Gerek yazılım gerek donanım olsun, gelişmekte olan bilişim teknolojisi araçlarının bu konuda önemli bir rolü vardır. Gelişen teknoloji araçlarıyla büyük miktarda veri içerisinde, saklı kalmış, değerli, kullanılabilir bilgileri ortaya çıkarmak ve stratejik kararlara destek sağlamak amacıyla kullanılan

veri madenciliđi; büyük miktarda verilerle ilgili sorun alanlarına yanıt bulmayı başarmıştır.

Veri madenciliđi yöntemlerinden biri olan kümeleme analizinin de son yıllarda kullanımı artmıştır. Artıřla birlikte pazarlama, biyoloji, bankacılık, sigortacılık, borsa, perakendecilik, telekomünikasyon, genetik, sađlık, bilim ve mühendislik, kriminoloji, sađlık, endüstri, istihbarat, eđitim vb. birçok dalda uygulamaları görölmektedir. Özellikle K-means algoritması ve Hiyerarřik kümeleme yöntemleri en çok kullanılan kümeleme analizi yöntemlerinden olmuřlardır. Bu çalışmada madde analizi yardımı ile kümeleme analizi yöntemlerinden K-Means Algoritması ve Hiyerarřik Kümeleme Yöntemleri kullanılarak bir eđitim kurumunun verileri ile bir kümeleme analizi uygulaması yapılmıştır.

Anahtar Kelimeler: Veri Madenciliđi, Kümeleme Analizi, K-Ortalamalar Algoritması, Hiyerarřik Kümeleme, Madde Analizi



To My Parents

ACKNOWLEDGEMENTS

First of all, I am most grateful to my supervisor Prof. Dr. Ahmet Sermet Anagün for his endless guidance, insights, encouragement, and kindness during the course of this thesis.

I would like to thanks my friends for their support during the course of this thesis.

Lastly, I would like to thanks my family for always being there to support me, making everything possible and their endless love.



TABLE OF CONTENTS

ABSTRACT.....	i
ÖZET.....	iii
ACKNOWLEDGEMENTS.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiii
CHAPTER 1 : INTRODUCTION	1
1.1. What is Data Mining?.....	2
1.2. Methods of Data Mining	3
1.2.1. <i>Predictive Models</i>	4
1.2.1.1. <i>Classification</i>	4
1.2.1.2. <i>Decision Trees</i>	4
1.2.1.3. <i>CART</i>	4
1.2.1.4. <i>Artificial Neural Networks</i>	4
1.2.1.5. <i>Support Vector Machines</i>	5
1.2.1.6. <i>Time Series Analysis</i>	5
1.2.1.7. <i>K-Nearest Neighbor Algorithm</i>	5
1.2.2. <i>Descriptive Models</i>	5
1.2.2.1. <i>Association Rules</i>	5
1.2.2.2. <i>Sequential Discovery</i>	6
1.2.2.3. <i>Clustering Analysis</i>	6
1.3. Usage Areas of Data Mining	6
1.4. Data Mining Process	7
1.4.1. <i>Identifying the Problem</i>	8
1.4.2. <i>Data Preparation</i>	8
1.4.3. <i>Collecting and Harmonizing</i>	8
1.4.4. <i>Combining and Cleaning</i>	9

1.4.5. <i>Selecting</i>	9
CHAPTER 2 : METHODOLOGY	10
2.1. Cluster Analysis	10
2.1.1. <i>Clustering Methods</i>	11
2.1.2. <i>Hierarchical Clustering Methods</i>	11
2.1.2.1. <i>Single-Linkage Clustering</i>	12
2.1.2.2. <i>Complete-Linkage Clustering</i>	13
2.1.2.3. <i>Average-Link Algorithm</i>	13
2.1.2.4. <i>Weighted Average-Link Algorithm</i>	13
2.1.2.5. <i>Central Connection Method</i>	14
2.1.2.6. <i>Median Connection Method</i>	14
2.1.2.7. <i>Ward Method</i>	14
2.1.3. <i>Non-Hierarchical Clustering Methods</i>	14
2.1.3.1. <i>K-Means Algorithm</i>	15
2.1.4. <i>The Comparison Between Methods</i>	16
2.1.5. <i>Determination of Number Of Clusters</i>	16
2.1.6. <i>Measures of The Hierarchical Clustering Methods</i>	17
2.1.6.1. <i>Euclidean Distance</i>	17
2.1.6.2. <i>Squared Euclidean Distance</i>	18
2.1.6.3. <i>Chebyshev Distance</i>	20
2.1.6.4. <i>Minkowski Distance</i>	20
2.1.6.5. <i>City-Block Distance</i>	21
2.2. Related Works	21
CHAPTER 3: ITEM ANALYSIS	25
CHAPTER 4: APPLICATION	29
4.1. Steps Of the Application	29
4.1.1. <i>Identifying the Problem</i>	29
4.1.2. <i>Data Preparation</i>	30
4.1.3. <i>Collecting and Harmonizing</i>	30
4.1.4. <i>Combining and Cleaning</i>	30

4.1.5. <i>Selecting</i>	31
4.2. Clustering Analysis of Raw Data	31
4.2.1. <i>K-Means Clustering on Raw Data</i>	31
4.2.2. <i>Hierarchical Clustering on Raw Data</i>	32
4.3. Clustering Analysis of Weighted Data	32
4.3.1. <i>K-Means Algorithm</i>	33
4.3.2. <i>Hierarchical Clustering</i>	33
4.3.2.1. <i>Within Group Distance</i>	34
4.3.2.2. <i>Furthest Neighbor</i>	35
4.3.2.3. <i>Median Clustering</i>	36
4.3.2.4. <i>Ward's Method</i>	38
4.4. Clustering analysis on Weighted Data With Item Difficulty and Discrimination Indexes.....	39
4.4.1. <i>K-Means Algorithm</i>	39
4.4.2. <i>Hierarchical Clustering</i>	39
4.4.2.1. <i>Within Group Distance</i>	40
4.4.2.2. <i>Furthest Neighbor</i>	41
4.4.2.3. <i>Median Clustering</i>	43
4.4.2.4. <i>Ward's Method</i>	44
CHAPTER 5:COMPARISON of the RESULTS of the METHODS	46
5.1. Comparison of the Results of the K-Means Algorithm.....	46
5.2. Comparison of the Results Obtained by Applying Hierarchical Clustering Methods	48
5.2.1. <i>Comparison of the Results of the Squared Euclidean Distance Measure of the Within Group Distance Clustering Method</i>	48
5.2.2. <i>Comparison of the Results of the Block and Minkowski Measures of the Within Group Distance Clustering Method</i>	50
5.2.3. <i>Comparison of the Results of the Measures of the Furthest Neighbor Clustering Method</i>	52
5.2.4. <i>Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Median Clustering Method</i>	54
5.2.5. <i>Comparison of the Results of the Squared Euclidean Distance Measure of the Median Clustering Method</i>	56

5.2.6. <i>Comparison of the Results of the Measures of the Ward's Method</i>	58
CHAPTER 6: CONCLUSION and SUGGESTIONS	61
REFERENCES.....	63
APPENDIX.....	68



LIST OF TABLES

Table 3.1. Application of a Sample Item Analysis	26
Table 3.2. Item difficulty index values and interpretation	27
Table 3.3. Item discrimination index values and interpretation table (Ebel & Frisbie, 1986).	27
Table 3.4. Item difficulty index and Item discrimination index Interpretation and weighting table (Ebel & Frisbie, 1986).....	28
Table 4.1. Comparison Of Current State and K-Means Algorithm	32
Table 4.2. Hierarchical clustering methods and Measure's cluster values	32
Table 4.3. The Score Range and Number of Student of K-Means Algorithm.....	33
Table 4.4. Number of Cluster of Hierarchical Clustering Methods and Measures....	34
Table 4.5. The Score Range and Student Number of Within Group Distance's Measures	34
Table 4.6. The Score Range and Student Number of Furthest Neighbor Measures ..	35
Table 4.7. The Score Range and Student Number of Median Clustering's Measures	37
Table 4.8. The Score Range and Student Number of Ward's Method's Measures ...	38
Table 4.9. The Score Range and Number of Students of K-Means Algorithm	39
Table 4.10. Number of Cluster of Hierarchical Method's and Measures	40
Table 4.11. The Score Range and Student Number of Within Group Distance's Measures	40
Table 4.12. The Score Range and Student Number of Furthest Neighbor Measures	42
Table 4.13. The Score Range and Student Number of Median Clustering's Measures	43
Table 4.14. The Score Range and Student Number of Ward's Method's Measures .	44
Table 5.1. Comparison of the K-Means Algorithm's Analyzes Results.....	46
Table 5.2. Comparison of the Results of the Squared Euclidean Distance Measure of the Within Group Distance Clustering Method	49
Table 5.3. Comparison of the results of the Block and Minkowski Measures of the Within Group Distance Clustering Method	51
Table 5.4. Comparison of the Results of the 5 Measures of the Furthest Neighbor Clustering Method.....	53

Table 5.5. Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Median Clustering Method..... 56

Table 5.6. Comparison of the Results of the Squared Euclidean Distance Measure of the Median Clustering Method..... 58

Table 5.7. Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Ward’s Method Clustering Method..... 59



LIST OF FIGURES

Figure 1.1. Methods Of Data Mining (Source: Şekeroğlu, 2010).....	3
Figure 1.2. Data Mining Processes	7
Figure 2.1. Clustering Methods.....	11
Figure 2.2. Single Linkage Clustering Process	12
Figure 2.3. Complete Linkage Clustering Process.....	13
Figure 4.1. K-Means Clustering Wilk's Lambda Values Graph.....	31



CHAPTER 1 : INTRODUCTION

The data produced by computer systems are meaningless in their own right because it does not make sense to the naked eye. When this data is processed for a definite goal, it becomes meaningful (Kalikov, 2006). Information is processed data for a purpose. It is not possible to make decisions based on raw data or information, which is merely an illustration of what happened in the past. It is also not possible to prevent loss from past experience. It is important to discover confidential information about past events, to adopt a management approach that enables us to take precautionary measures with predictive situational predictions and to predict possible losses (İnan, 2003). Therefore, it is very important to use techniques that can process large amounts of data. The process of converting this raw data into information or making sense can be done with data mining (Kalikov, 2006).

Data mining is the search for correlations between large-scale data, which can enable us to make estimations about the future from major data stacks, information mining, information mining. Data mining is used to extract substance, unclear, beforehand unknown but potentially useful information from the existing data. This includes some technical approximations such as clustering, data summarizing, analysis of variables, and determination of deviations. It is mainly concerned with the envisagement of patterns or layouts between data sets, the analysis of data and the discovery of relationships between them using software techniques. The purpose of data mining is to detect previously unnoticed data patterns. In particular, the K-means algorithm and Hierarchical Clustering Methods are the most commonly used clustering analysis methods. In the second part, the K-means algorithm and Hierarchical Clustering Methods are clarified in detail. In the third part, substance analysis is clarified and an exemplary application is shown. In the fourth section, the application is explained and the results of the analysis are interpreted. In the last part, the results are discussed and suggestions are made.

In this study, it was aimed to determine the classes by clustering methods instead of the total score, and to observe the effects of different algorithms on the clustering and to compare them with the current practice. In addition, it has been investigated that the positive and negative effects of item analysis on the formation of classes and

applicability and whether the created groups are homogeneous according to the current situation.

1.1. What is Data Mining?

The increasing use of database systems and the remarkable increase in the volumes of data storage units have led to the ineffectiveness of traditional query and reporting tools against huge masses of data. As a result, new searches have emerged in the databases under the name of knowledge discovery (VTBK) (KDD Knowledge Discovery in Databases) (Dinçer, 2006). To make a simple description, data mining is the job of accessing knowledge from large-scale data and mining knowledge. In other sayings, data mining is a collection of processes that contain the use of advanced data analysis tools such as statistics, artificial intelligence, and machine learning to uncover confidential designs and relationships within data that do not mean anything by itself. Data mining implementations are mainly used in marketing, banking, medicine, engineering, industry, stock market analysis and national security fields. Much commercial software has been produced for data mining studies. Oracle DM, Microsoft SQL Server Analysis Services, SPSS Clementine, SAS Enterprise Miner are just a few of these products. (Bozkır et al., 2009).

Data mining is a continuum in which a large number of advanced data analysis methods based on statistics and artificial intelligence are used, preferably through a visual programming interface, to reveal patterns and relationships hidden in large data stacks. (Dolgun et al., 2009). Data Mining provides critical information from very large data stacks. Thus, under normal conditions, the information obtained from the researches that take a long time with the accuracy of the data is obtained in a short time and precisely with data mining. This information is used to make objective evaluations or to make strategic decisions. This information helps to analyze corporate data sources well and make predictions about business approaches. Briefly, data mining allows companies to analyze strategic data by extracting critical data from a huge mass of data to guide them. (Alpaydın, 2000). Basically, data mining is relevant to the use of software techniques and analysis of design or designs between data sets. The computer is liable for determining the relationship, orders, and characteristics between the data. The purpose is to detect beforehand unnoticed data designs. (Arslan, 2008).

1.2. Methods of Data Mining

Data mining methods are separated into predictive models and descriptive models. Predictive models; are used to improve a model from known data and to estimate results for unknown data by using an established model. For example, students who take a grade on the passing grade pass the course. Course passing grade; The course instructor depends on the difficulty level of the exams and the number of students taking the course. The instructor, the degree of difficulty of the exams and the number of students taking the course were independent variables; and the passing grade is a dependent variable. It is estimated whether the student has passed the course according to the grade of the student and other variables.

In descriptive models, it is provided to define the patterns in the data to help decision-making. For example, a descriptive model results in relationships such as “A customer who buys child food is three times more likely to receive diapers.”. Data mining models are basically; classification, clustering and association rules. In addition to 3 models, there are models such as estimation, prediction, time series analysis, and sequence discovery.

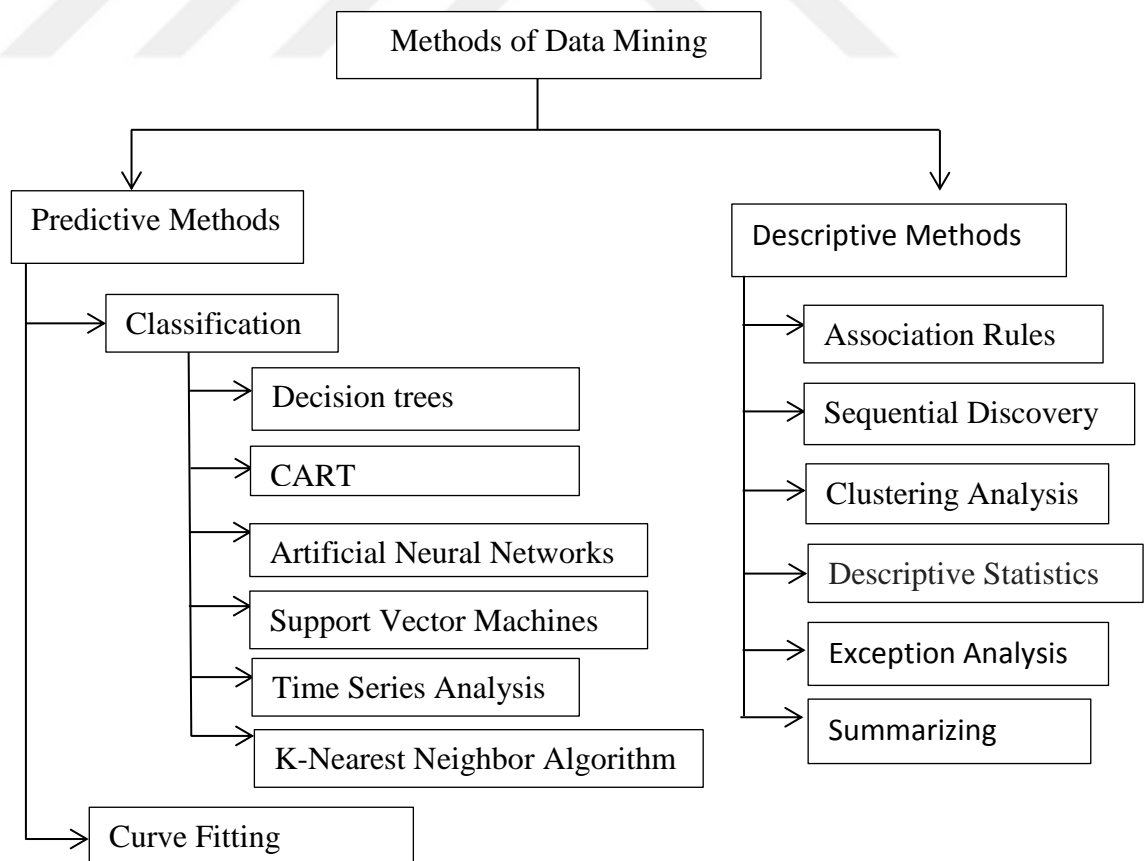


Figure 1.1. Methods Of Data Mining (Source: Şekeroğlu, 2010)

1.2.1. Predictive Models

1.2.1.1. Classification

Classification is one of the most well-known data mining models, including painting, pattern recognition, disease diagnostics, fraud detection, quality control studies, and marketing. Classification is a predictive model; estimating how the weather will be the next day or how much blue ball is in a box is actually a classification process (Dunham, 2003).

1.2.1.2. Decision Trees

Decision trees, which is one of the data classification methods, have been applied in many ways under the name of machine learning in applied statistics. Several learning methods enable the decision tree to be created using a database of examples. Decision trees are one of the most used algorithms in classifying problems. It can be said that decision trees are easier to construct and understand compared to other methods.

1.2.1.3. CART

The classification and regression trees method was introduced by Breiman in 1984. The CART decision tree is based on the principle that the tree is divided into two branches from each decision node.

1.2.1.4. Artificial Neural Networks

Artificial Neural Networks (ANN) is an information processing system inspired by biological neural networks. The history of artificial neural networks goes back to 1942. In 1942 McCulloch and Pitts developed the first cell model and is therefore considered the beginning of artificial neural networks. In 1949 Hebb proposed the first learning rule to adjust cell connections. In 1958, Rosenblatt developed the sensor model and the learning rule, revealing the basis of the rules used today. In 1969, Minsky and Papert conducted a precise analysis of the sensor and proved that it could not be used for complex logic functions. Between 1982-1984 Kohonen described the self-organizing map. He developed a network of uncontrolled learners, named after him. In 1986, Rumelhart reestablished the spread backward. In 1988, Chua and Yang developed cellular neural networks.

Artificial neural networks are formed by the merging of artificial neural cells in various ways and are arranged in layers. The most prominent features of artificial neural networks are the interconnected neurons, the determination of the intervals between connections and the ignition function.

1.2.1.5. Support Vector Machines

Support Vector Machines is a controlled classifying algorithm based on statistical learning theory. The mathematical algorithms of support vector machines were initially designed for the problem of classification of two-class linear data, then generalized for the classification of multi-class and non-linear data. The working principle of the support vector machines is based on the prediction of the most convenient decision function that can separate the two classes, in other words, the description of the hyperplane that can separate the two classes from each other in the most convenient way. (Vapnik, 1995; Vapnik, 2000).

1.2.1.6. Time Series Analysis

Data obtained by observing a response variable at specific times is called time series. Data are obtained from equal interval time points. In summary, time series analysis is defined as investigating the probabilistic structure of a time series and predicting its future status. Time series are analyzed for long term planning and forecasting of future operations. In general, time series analysis is to make predictions for the future from past records. There are 4 elements in the composition of the time series. These are trend constituent, seasonal constituent, cyclical constituent, and a random constituent.

1.2.1.7. K-Nearest Neighbor Algorithm

The K-Nearest Neighbor Algorithm method is one of the data mining methods in hydrology (Brath et al. 2002), energy sector (Sorjamaa et al. 2007), (Lora et al, 2007), meteorology (Dragomir, 2010), (Singh and Ganju, 2006) and medicine (Lowsky et al, 2013). it is also used to model the nonlinear dynamics of series in finance.

1.2.2. Descriptive Models

1.2.2.1. Association Rules

Analyzing the co-occurring events is covered by the topics of data mining. Data mining methods that analyze the co-occurrence of events are called association rules. Determining which goods or services the customer is inclined to purchase during a purchase or in successive purchases are one way to ensure that more products are sold to the customer. Association rules that enable the definition of purchasing trends are commonly used in data mining under the name of Market Basket Analysis for marketing goals. (Göral, 2007).

1.2.2.2. Sequential Discovery

Sequential discovery is used to determine sequential time patterns in the data. Sequence discovery is similar to association analysis, but the relationship is based on time. There is a need to purchase products at the same time in the market basket analysis, but products can be purchased in any order over time.

1.2.2.3. Clustering Analysis

Clustering analysis is the operation of grouping the data as it is in the classification. In the classification process, classes are predetermined, while clustering classes are not predetermined. Unlike classification, it is not clear how many groups will be formed in the cluster analysis.

1.3. Usage Areas of Data Mining

It is possible to use data mining wherever there is a large volume of data. Today, data mining applications are widely used in many areas where decision-making is needed. For example, marketing, biology, banking, insurance, stock exchange, retailing, telecommunications, genetics, health, science and engineering, criminology, health, industry, intelligence, etc. successful applications are seen in many branches. (İnan, 2003; Albayrak, 2008; Akgöbek and Çakır, 2009).

For the last 20 years, various data mining algorithms have been used in the United States for a variety of applications, from confidential listening to the discovery of tax evasion. When the sources are examined, medicine, biology, and genetics are seen as the most used fields of data mining. Nowadays, the fields of data mining can be summarized as follows:

- In the field of marketing; customer classification, demographic characteristics of customers, the establishment of marketing strategies to be developed to retain existing customers in various marketing campaigns, market basket analysis, cross-sale analysis, customer valuation, customer relationship management, various customer analysis, sales forecasts,
- In the field of banking and finance; finding hidden relationships between different financial indicators, estimating financial failures, identifying credit card irregularities, classifying customers, evaluating credit demands, risk analysis, and risk management,
- In the field of insurance; estimating the customers who will demand new policies, detecting insurance frauds, determining the type of risky customers,
- In the field of the stock market; stock price estimation, general market analysis,
- In the field of retailing; the point of sale data analysis, shopping cart analysis, selection of supply and store placement,
- In the field of medicine and medical; estimation of test results, product development, medical diagnosis, determination of treatment process,
- In the field of industry; quality control analysis and optimization of production processes.

Here are a few examples of sectoral applications of data mining methods. International financial institutions such as Merrill Lynch, Citibank, and World Bank use data mining methods to analyze financial forecasting and credit risks. Bank and credit card companies such as American Express, Mellon Bank and First USA Bank use data mining methods to prevent potential fraud and abuse and identify customers' handwriting and signatures. (Elmas, 2011).

1.4.Data Mining Process

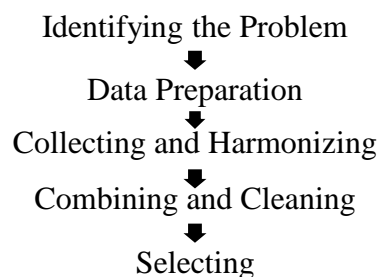


Figure 1.2. Data Mining Processes

1.4.1. Identifying the Problem

The first condition to succeed in data mining studies is to clearly define the purpose of the organization for which the application will be carried out. The aim of the organization should be focused on the problem and denoted in a clear language. A data mining study that does not fully coincide with the problem will not be enough to solve the problem, but may also lead to other problems. Also, the costs to be incurred in the wrong decisions and the predictions of the benefits to be made in the right decisions should be included at this stage.

1.4.2. Data Preparation

Problems that will arise during the establishment of the model will lead to an often return to this stage and reorganization of the data. This makes it possible for a decision-maker to spend 50% to 85% of the energy and time in the data discovery process for the preparation of the data and the installation of the model. The data preparation step consists of collecting and harmonizing, combining and cleaning and selecting steps

1.4.3. Collecting and Harmonizing

It is the step of identifying the data that is thought to be necessary for the identified problem and the data sources from which it will be collected. It is an important decision which data sources will be used. Because too few data sources will leave the data mining work incomplete, too many data sources may lead to data pollution that may cause the process to be prolonged. In addition to the organization's own data sources, various databases just as census, weather, central bank blacklist or databases of data marketing organizations can be used for data collection.

The collection of data to be used in data mining from different sources will naturally lead to data mismatches. The most important of these discrepancies are different times, update errors, different data formats, coding differences (for example, the gender feature m / f in one database, 0/1 in another database), different measurement units and assumption differences. It is also important how, where and under what conditions data is collected. The use of unreliable data sources will affect the reliability of the entire data mining process.

For this reason, since the data mining studies that will be good results can only be built on good data, the compatibility of the collected data should be examined and evaluated in this step.

1.4.4. Combining and Cleaning

In this step, the problems and discrepancies found in the data collected from different sources and identified in the previous step are eliminated as much as possible and the data is collected in a single database. However, it should be kept in mind that simple troubleshooting and trivial troubleshooting will be the source of greater problems in the future.

1.4.5. Selecting

In this step, data selection is made depending on the model to be installed. For example, for a predictive model, this step means selecting dependent and independent variables and the data set to be used in the model.

CHAPTER 2 : METHODOLOGY

2.1. Cluster Analysis

Cluster analysis is a widely used method in social studies such as medicine, sociology, psychology, economics, marketing, grouping them by taking into account the basic characteristics of units and providing summary information. Clustering analysis, which is one of the multivariate statistical analysis methods, is a method used to divide the units or variables into similar and meaningful subsets. In this analysis, the units in the clusters are similar in themselves, but they are significantly different from the units in the other clusters. In the clustering analysis, the measures calculated with the help of similarities or differences between variables are used to determine subsets.

Clustering analysis aims to divide the n number of units or objects into homogeneous, heterogeneous clusters among themselves according to the p number of variables, to sub-clustered p number of variables to reveal common factor structures, and to take both units and variables at the same time, to n to subset with common properties according to the variable. (Kaya and Türkmen, 2013).

In the cluster analysis, firstly, the data matrix is obtained according to the observation values of n number of units p number of variables. The distance/similarity of units or variables is then calculated with a distance/similarity measure showing the similarities or differences between units or variables. Using clustering methods, units or variables are divided into the appropriate number of clusters according to the similarity or difference matrix. Analytical methods are used for the interpretation of these sets (Aaker et al.,1997).

Some alternative criteria and methods can be utilized by using similar distances in the use of cluster analysis. Euclidian, Standardized Euclidian, Manhattan Mahalanobis, Square Euclidian, Minkowski or Canberra measurements can be used for distances between units. This makes it necessary to be careful in using cluster analysis in practice. The clustering algorithm subdivides the database. The elements in each cluster have prevalent characteristics that make different the group from the other groups. In clustering models, the aim is to find clusters whose cluster members are very similar to each other, but whose properties are very different from each

other, and the records in the database are separated into these different clusters. (Arslan, 2008).

2.1.1. Clustering Methods

The methods used in cluster analysis are divided into two groups as hierarchical and non-hierarchical methods.

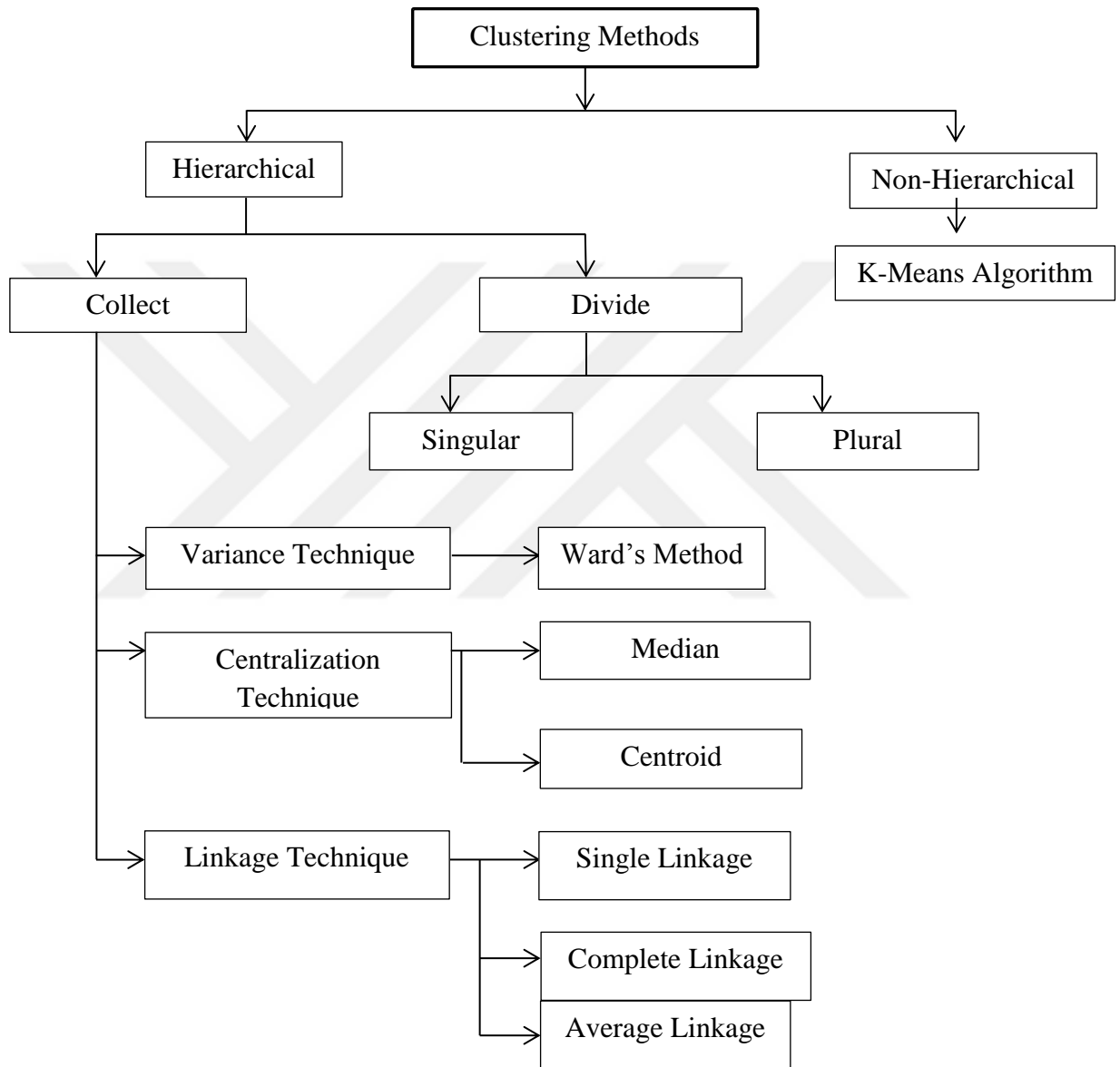


Figure 2.1. Clustering Methods

2.1.2. Hierarchical Clustering Methods

Hierarchical methods are methods of sequentially identifying clusters by uniting units together to merge them according to their similarities. In this method, the

clustering process is shown with the help of two-dimensional graphs called dendrograms (Turanli and Basar, 2011). The number of cluster is not known in hierarchical methods. Hierarchical methods are divided into two as combining and seperating methods (Uzgören et al., 2013). In combinatorial methods, first, starting with n clusters, each with a single observation, then clustering similar clusters is combined to reduce the number of clusters. In other words, it initially assumes that each unit constitutes a set, n places be unit in steps $n, n-1, n-2, \dots, n-r$. Seperating methods, in contrast to combinatorial methods are initially considered to be a set of all units, and so on until a single set of units is created from single set. Hierarchical methods are divided into five groups as the central method, single connection method, full connection method, average connection method and Ward's method (Orhunbilge, 2010).

2.1.2.1. Single-Linkage Clustering

This technique was first applied by Florek (1951) et al. And then by Sneath (1957) and Johnson (1967), respectively. Using the distance or similarity matrix, the two closest observations or clusters are combined and the merging process is repeated (Senturk, 1995; Firat, 1995).

The results of the single connection technique can be shown in a tree diagram or dendrogram. Tree structure branches, clusters (Everitt and Dunn, 2001; Senturk, 1995).

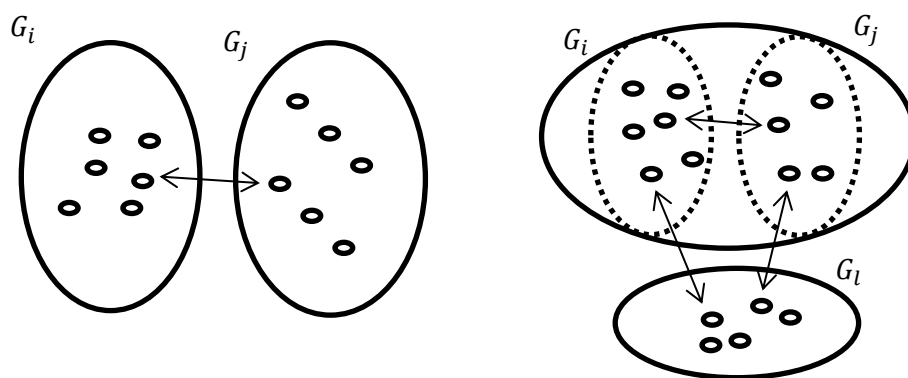


Figure 2.2. Single Linkage Clustering Process

2.1.2.2. Complete-Linkage Clustering

This technique is the exact opposite of a single connection technique. In this technique, the two closest clusters or observations are combined using the resulting distance or similarity matrix.

The full connection technique cannot guarantee that all clusters can be formed healthily if the distances of the observations in the same cluster are smaller than a certain value (Tatlidil, 1992).

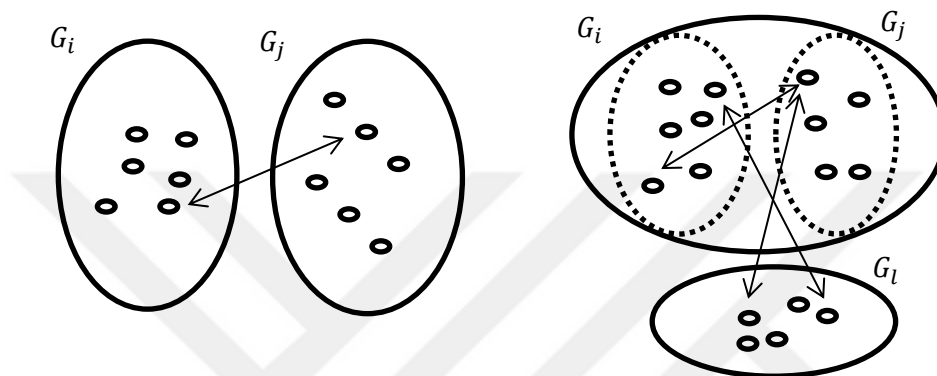


Figure 2.3. Complete Linkage Clustering Process

2.1.2.3. Average-Link Algorithm

The average connection technique was proposed by Sokal and Michener. In this technique, the difference between the two sets is taken as the average difference between element pairs between one set and element pairs in another set (Everitt, 1981). There are modified types of this technique. In the most widely used type, the arithmetic mean of the distance between the observation pairs is calculated. The average linkage technique is widely used in biology, but its use in social sciences is increasing. Similar dendrograms usually occur in full linkage and average linkage techniques. However, since the distance is defined differently in each method, the joins can occur at different levels (Firat, 1995).

2.1.2.4. Weighted Average-Link Algorithm

To find the distance between the two clusters, a distance is calculated as in the average connection technique. In this technique, apart from the average connection, the distance between the newly formed cluster and the other clusters is weighted by the number of observations in each cluster.

2.1.2.5. Central Connection Method

In this method, the average value of individuals or objects in each cluster is the center of that cluster. First, the squares of the Euclidean distances are calculated. Thus, the similarity-proximity matrix is obtained. According to the Euclidean distance, the units which are close to each other are located in the same cluster. The averages of the clusters thus obtained are calculated. This process is continued until all units have been allocated to clusters. Naturally, the average of the units within the cluster will change in each clustering process. Although this situation can lead to complex results, the fact that it is less affected by extreme values causes the central method to outperform other methods.

2.1.2.6. Median Connection Method

Unlike the central connection method, the distance between the two sets in the median connection method is obtained by calculating the distance between the centers of the two sets with equal weight (Gower, 1967).

2.1.2.7. Ward Method

This approach, also known as the minimum variance method, takes the average distance of the observation falling in the middle of a cluster from the observations within the same cluster and makes use of the total deviation squares. It is a commonly used hierarchical clustering method.

2.1.3. Non-Hierarchical Clustering Methods

If there is preliminary information about the number of clusters at the beginning or if the number of clusters is specified by the researcher before the analysis, non-hierarchical methods are used instead of hierarchical methods (Anderberg, 1973). In these methods, the desired number of clusters is specified as k . Then the cluster averages are determined. Each unit is included in the closest cluster. The most commonly used method is the k-means method. K-means method is a method developed by Mac Quinn. In this method, cluster averages are determined first and each unit is assigned to the nearest cluster considering the distance to the center. For new clusters, the cluster averages are calculated again. All units are reclassified according to the new cluster averages. Following this sequence, the process is continued until the cluster averages are almost the same (Mooi and Sarstadt, 2011).

In the K-means method, it is not necessary to calculate the distance matrix or similarity matrix for clustering. Just determining the number of clusters is enough. There are also several methods for detecting the number of clusters. The starting points can be determined at random. Clustering is performed by increasing the number of clusters ($k = 2, 3, 4, \dots$), separation analysis is applied for each clustering model and Wilk's Lambda values are found. The number of clusters with the most appropriate Wilk's Lambda value is assumed.

2.1.3.1. K-means Algorithm

The K-means algorithm is to separate a data set consisting of N data objects into K sets given as input parameters. The purpose is to ensure that the clusters acquired at the end of the partitioning process have a maximum of intra-cluster similarities and a minimum of inter-cluster similarities.

K-means is one of the most commonly used clustering algorithms. It is easy to apply. It can cluster large-scale data quickly and efficiently. "K" refers to the number of fixed sets required before starting the algorithm. The K-means algorithm with its repetitive divisor structure reduces the sum of the distances of each data to the set to which it belongs. The K-means algorithm tries to identify the K sets that will make the least squared error.

As long as K-means and intra-cluster similarity are big and the similarity between clusters is small, the accuracy of the cluster can be mentioned. Although the problem is NP-hard, the means algorithm generally provides a good solution with the iterative approach.

Let each data be an n -dimensional real vector, a data set $\{x_1, x_2, x_3, \dots, x_N\}$ and K as the number of sets divide. K-means clustering purposes to partition N data into K clusters to minimize quadratic error. In other words;

$$\mu_i = \frac{1}{|S_j|} + \sum_{x_i \in S_j} x_i \quad (1)$$

μ_i , the average of points in S_j ;

$$\arg \min_s \sum_{j=1}^K \sum_{x_i \in j} \|x_i - x_j\|^2 \quad (2)$$

It can be found.

According to the working mechanism of the K-means algorithm, first, K objects are randomly selected to represent the center point or average of each set. The remaining objects are included in the clusters they are most similar to, taking into account their distance from average values of the clusters. Then, the average value of each cluster is computed and the new cluster centers are specified and the distance of the objects to the centers are specified and the distance of the objects to the center is examined again. The algorithm continues to repeat until there are no changes.

The algorithm consists of 4 stages:

1. Determination of cluster centers
2. Clustering off-center data by distance
3. Determination of new centers
4. Repeat steps 2 and 3 until the stable state is reached

2.1.4. The Comparison Between Methods

There is no definite judgment as to which of these methods should be chosen. In general, hierarchical methods, the average connection, and the Ward method are widely used. Non-hierarchical methods are also widely used because the similarity matrix reaches large dimensions due to a large number of units, the difference in distance measurements and the effect of outliers. In some cases, both groups of methods are applied together. The number of clusters is specified by hierarchical methods, the cluster center values are determined, and non-hierarchical methods can be used with initial values. In other words, if the number of clusters is determined at the beginning of the research, non-hierarchical clustering methods are preferred, and if the number of clusters is not decided, the hierarchical clustering method is preferred (Burmaoğlu, 2011)

2.1.5. Determination of Number Of Clusters

As mentioned before, in the hierarchical methods, the number of clusters is determined according to the method used, while in the non-hierarchical methods the number of clusters is determined by the researcher before the analysis. Equality commonly used to determine the number of clusters,

$$\mathbf{k} = \left(\frac{n}{2}\right)^{1/2} \quad (3)$$

It is expressed in the form (Tatlıdil, 2002). However, this equation is used for small samples and does not give good results as the sample grows.

$$\mathbf{M} = \mathbf{k}^2 |\mathbf{W}| \quad (4)$$

It is equality. W in the equation is the matrix of the sum of squares within the group. Accordingly, the “ k ” value that makes the minimum “ M ” values determines the number of clusters (Cengiz and Ozturk, 2012).

If the number of units is above 30, it is appropriate to use the Wilk’s Lambda criterion. Wilks likelihood ratio statistics,

$$\Delta = \frac{|\mathbf{W}|}{|\mathbf{W}+\mathbf{B}|} = \frac{|\mathbf{W}|}{|\mathbf{T}|} \quad (5)$$

The sum of the squares matrix between the groups, T is the sum of squares matrix. This criterion takes a value between zero and one. When this values is below 0.01, the appropriate number of clusters is determined (Cengiz and Ozturk, 2012).

2.1.6. Measures of The Hierarchical Clustering Methods

2.1.6.1. Euclidean Distance

In mathematics, Euclidean distance of Euclidean metric is the “ordinary” straight line distance between two points in Euclidean space. With this distance, the Euclidean space becomes a metric space. The related norm is called the Euclidean norm. It is also referred to as the Pythagorean metric. A generalized terms of the Euclidean norm is the L^2 norm or the L^2 distance. The Euclidean distance between points P and q is the length of the line segment joining them \overline{pq} . In Cartesian coordinates, if $p = (p_1, p_2, p_3, \dots, p_n)$ and $q = (q_1, q_2, q_3, \dots, q_n)$ are two points in the Euclidean n -space, then the distance form (d) to p is given from q or q to p by Pythagorean formula (Anton and Rorres, 1994).

$$\begin{aligned} \mathbf{d}(\mathbf{p}, \mathbf{q}) = \mathbf{d}(\mathbf{q}, \mathbf{p}) &= \sqrt{(\mathbf{q}_1 - \mathbf{p}_1)^2 + (\mathbf{q}_2 - \mathbf{p}_2)^2 + \dots + (\mathbf{q}_n - \mathbf{p}_n)^2} \\ &= \sqrt{\sum_{i=1}^n (\mathbf{q}_i - \mathbf{p}_i)^2} \end{aligned} \quad (6)$$

The location of the point in the Euclidean space is the Euclidean vector. Thus, p and q can be represented as Euclidean vectors with ends (terminal points) ending at both

ends starting from the source of the cavity (starting points). Euclidean norm or Euclidean length or the size of a vector measure the length of the vector (Anton and Rorres, 1994).

$$\|\mathbf{p}\| = \sqrt{\mathbf{p}_1^2 + \mathbf{p}_2^2 + \dots + \mathbf{p}_n^2} = \sqrt{\mathbf{p} \cdot \mathbf{p}} \quad (7)$$

When the final expression contains the spot product.

It defines a vector as a line segment directed from the origin of the Euclidean space (vector tail) to a point (vector end) in that area. The length is actually the distance from the tail to the end. It appears that a vector is the only Euclidean distance between the tail and the tip of the Euclidean norm.

The relationship between points P and q may include a direction (for example, from p to q), so this relationship can be represented by a vector given by it.

$$\mathbf{q} - \mathbf{p} = (\mathbf{q}_1 - \mathbf{p}_1, \mathbf{q}_2 - \mathbf{p}_2, \dots, \mathbf{q}_n - \mathbf{p}_n) \quad (8)$$

In a two- or three-dimensional space ($n = 2, 3$), this can be visually represented as an arrow from p to q. In any field, it can pass for the position of q relative to p. It can also be named a displacement vector if p and q represent the two positions of some moving points. The Euclidean distance between P and q is only the Euclidean length of this displacement vector.

$$\|\mathbf{q} - \mathbf{p}\| = \sqrt{(\mathbf{q} - \mathbf{p}) \cdot (\mathbf{q} - \mathbf{p})} \quad (9)$$

Equation is equivalent to 1 and also:

$$\|\mathbf{q} - \mathbf{p}\| = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p} \cdot \mathbf{q}} \quad (10)$$

(Anton and Rorres, 1994).

2.1.6.2. Squared Euclidean Distance

The square of the standard Euclidean distance known as (SED) is also interesting; In the equation:

$$d^2(\mathbf{p}, \mathbf{q}) = (\mathbf{p}_1 - \mathbf{q}_1)^2 + (\mathbf{p}_2 - \mathbf{q}_2)^2 + (\mathbf{p}_i - \mathbf{q}_i)^2 + \dots + (\mathbf{p}_n - \mathbf{q}_n)^2 \quad (11)$$

Squared Euclidean distance is central to the estimation of the parameters of statistical models using the least-squares method, which is a standard approximation in

regression analysis. The corresponding loss function is squared error loss (SEL) and gives greater weight to larger errors. The corresponding risk function is the mean square error (MSE).

Squared Euclidean distance is not a measure since it does not provide triangular inequality. It is, however, a more general concept of distance, ie it can be used as a difference (in particular a Bregman deviation) and as a statistical distance. The Pythagorean theorem is simpler from the point of square distance (since there is no square root); if $pq \perp qr$, then;

$$\mathbf{d}^2(\mathbf{p}, \mathbf{r}) = \mathbf{d}^2(\mathbf{p}, \mathbf{q}) + \mathbf{d}^2(\mathbf{q}, \mathbf{r}) \quad (12)$$

Pythagorean identity in information geometry can be generalized to other Bregman deviations from the SED, including relative entropy (Kullback - Leibler deviation), allowing the use of generalized forms of least squares to solve nonlinear problems.

The SED is a smooth, absolutely convex function of the two points, unlike the distance where the two points are equal and not completely convex (non-smooth because they are smooth). SED is therefore preferred in optimization theory because it allows the use of convex analysis. Since squaring is a uniform function of non-negative values, minimizing SED is equivalent to minimizing Euclidean distance, so the optimization problem is equivalent in both respects, but is easier to solve using SED.

If one of the points is fixed, the SED can be explicated as a potential function, in which case a semi-normalization factor is used, and the sign can be changed depending on the contract. In detail, two points were given p, q . The vector p and q are proportional to the Euclidean distance. If one corrects p , $X_p(q): p - q$ and "p" can define a smooth vector field indicating. This is the gradient of a scalar-valued function. The half SED from "p", where power cancels the two in power. When writing half of the square distance $D_p(q) := \frac{1}{2} \sum_i (p_i - q_i)^2$ from P to p, $p - q = \text{grad}_q D_p$ alternatively, the vector field pointing to the p field can be considered and minus sign.

In information geometry, the concept of a vector field "pointing from one point to another" can be generalized to statistical manifolds - one can use tangent vectors at different points and an affine connection to flow the exponential map from one point

to another and in a statistical manifold, this is reversible by defining a unique "difference vector" from any point to another. In this context, the SED (gradient produces the standard difference vector) is generalized to a decomposition that produces the information geometry of the manifold; A uniform structure (geometric structure) of such a decomposition is called canonical decomposition.

In the field of rational trigonometry, SED is called quads (Ay and Amari, 2015).

2.1.6.3. Chebyshev Distance

The distance Chebyshev between two vectors or points is x and y , with standard coordinates x_i and y_i , respectively.

$$D_{\text{Chebyshev}}(\mathbf{x}, \mathbf{y}) := \max_i (|x_i - y_i|) \quad (13)$$

This equals the limit of the L_p metrics.

$$\lim_{p \rightarrow \infty} (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (14)$$

therefore also known as the L_∞ metric.

Mathematically, Chebyshev distance is a metric induced by the supremum norm or uniform norm. This is an example of an injective metric. Chebyshev distance in two dimensions, ie plane geometry, if the points x and y have x_1, y_1 cartesian coordinates x_1, y_1 and x_2, y_2

$$D_{\text{Chebyshev}}(\mathbf{x}, \mathbf{y}) := \max (|x_2 - x_1|, |y_2 - y_1|) \quad (15)$$

Below this metric, a radius circle r , which is a set of points with a distance Chebyshev from a center point, is a square with sides $2r$ in length and parallel to the coordinate axes. In a chessboard in which one uses a separate Chebyshev distance rather than a continuous one, the circle r of radius r is a square of $2r$ side length; for example, in a chessboard, the radius 1 circle is 3×3 squares (Xu et al., 2013).

2.1.6.4. Minkowski Distance

p distance between two points Minkowski distance

$$\mathbf{P} = (x_1, x_2, x_3 \dots x_n) \text{ and } \mathbf{Q} = (y_1, y_2, y_3 \dots y_n) \in \mathbf{R}^n$$

It is defined as follows:

$$\sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p} \quad (16)$$

Minkowski distance for $P \geq 1$ is a result of a metric Minkowski inequality. For $p < 1$, the distance between (0,0) and (1,1) is $2^{1/p} > 2$, but the point is (0,1), a distance between these two points is 1. Therefore, it violates the inequality of triangles.

Minkowski distance is typically 1 or 2 in p with typical use. the second is the Euclidean distance, the former is sometimes known as the Manhattan distance. Chebyshev distance of P 's limit when reaching infinity:

$$\lim_{p \rightarrow \infty} (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} = \max_{i=1}^n |x_i - y_i| \quad (17)$$

Similarly, we have

$$\lim_{p \rightarrow -\infty} (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} = \max_{i=1}^n |x_i - y_i| \quad (18)$$

for the negative infinity of p .

The Minkowski distance can also be seen as the difference between P and Q intelligent-component of a plural force mean (Xu et al., 2013).

2.1.6.5. City-Block Distance

The city-block distance measure calculated by the sum of the absolute values of the distances between the units is expressed as:

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (19)$$

(Ienco et al., 2012).

2.2.Related Works

In his study, Sarıman tried to reveal the differences between partitioning algorithms based on the clustering algorithms used in data mining. Using partitioned clustering algorithms, the flags data set was obtained from the 30 features of 194 country flags, which would make the most suitable clustering (country area, country population, country religion) and similar countries were aimed to be in the same clusters (Sarıman, 2011).

In their study, Çelik tried to determine the provincial groups showing the same structure with 10 health variables belonging to 81. For this reason, it is deemed

appropriate to use the so-called Cluster Analysis method. Using TURKSTAT health statistics for 2010, they were clustered according to 81 provincial health structures. When clustering analysis of 81 provinces was divided into 7, 10 and 15 clusters, the number of clusters was determined and the results were examined. As a result of the analysis, the worst provinces were specified according to the health data (Çelik, 2013).

Karabulut, Gürbüz and Sandal, socio-economic variables of the work 54 with the aid of Turkey's 81 provinces, showing the same structure have attempted to identify the homogeneous first group. In this study, it is considered appropriate to use the statistical method called Hierarchical Cluster Analysis. 81 cities were firstly divided into 7, 10 and 15 groups and tested. According to the results, it was determined that the most meaningful clustering was obtained as a result of 15 classifications. The results of the analysis were supported by Dendrogram and Agglomerative Chart. According to the Euclidean and Pearson Proximity Matrix measure used in the analysis, Bitlis and Mardin were the most similar cities, while the least similar was Istanbul and Kars. With this method, different socio-economic regions of the provinces were determined with the help of the variables that cause the separation and homogeneous structure of the provinces (Karabulut et al., 2004).

In their study, Gevrekçi, Ataç, Takma, Akbaş, and Taşkın examined comparatively the structure of sheep breeding in 11 provinces in Western Anatolia. In this research, the number of sheep obtained from TurkStat, the number of sheep milked, sheep milk yield, slaughtered sheep-lamb number, sheep-lamb meat production (ton), number of sheep and sheep production were used for the years 2003-2008. Multi-Dimensional Scaling (MDS) and Clustering analyses were applied and the provinces were classified in terms of sheep breeding. As a result of MDS and clustering, the Western Anatolian provinces in terms of sheep breeding constituted four main groups. These groups are Afyonkarahisar-Balıkesir; Izmir-Manisa; Bursa-Çanakkale-Denizli-Kütahya-Uşak and Aydın-Muğla (Gevrekçi et al., 2011).

In their study, Cengiz and Öztürk tried to determine the educational levels of the provinces by using clustering analysis, with the help of the rates of illiterate, illiterate but not graduating, primary school graduates, secondary school graduates, higher

education graduates, graduate graduates, doctorate graduates, and unknown educational levels (Cengiz and Ozturk, 2012).

In their study, Wolfram, Wang, and Zhang examined search session models using the clustering technique on transaction records representing three different types of Web-based information retrieval systems. The results revealed that search behaviors can be clustered into distinct groups based on session characteristics and show similarities, even if different systems exist. The session-based analysis is significant for understanding user search action and can help system designers develop systems that better meet the needs of several user groups (Wolfram et al., 2009).

Černohorská, Černohorský ve Teplý, aims to develop a stability model for the banking sector in the Czech Republic by taking into account the data for the period 1995-2005. According to the model presented in the study, the stability of banks can be easily evaluated by clustering and discriminant analysis. In the application made in 38 banks of the Czech Republic, 17 banks were found to be at the qualification level according to the model (Černohorská et al., 2007).

Kuo, Lin and Shih's work aimed to propose a new data mining framework that first clustered data and then followed merger rules mining. In the first stage, the ant system based clustering algorithm and ant mileage are used in the clustering database, while ant colony system based association rules are applied to find practical rules for each group of mining. The results of the evaluation showed that the proposed method can not only make the rules faster but also discover the more important rules (Kuo et al., 2007).

The purpose of this study is to analyze trophic changes in fish types caused by the swamp of rivers. The trophic data collected before and after the dam construction, and it was used to conduct the study using Clustering methods. The methodology used consisted of data analysis, and then allocating clusters for subsequent information of the application. The description of the number of clusters, the usage of particular types of clustering distances and the usage of validation indexes are discussed. The clustering approximations were applied individually in both stages and in both circumstances, five large clusters of fish were specified. This assessment could be used by biologists so as to utilize environmental influences and managers

can improve strategies to address the social and economic effects caused to the societies that depend on fishing (Almeida et al., 2019).

Capece, Cricelli, Di Pillo and Levialdi's paper concentrates on the changes in performance in the natural gas retail market by analyzing the profit and financial position of the companies interested over the first three years following the market deregulation. The balance sheets of 105 Italian companies in the industry are analyzed, after which cluster analysis is performed operating the most important performance indexes. The companies are then analyzed within each cluster compared to age, size, geographical position and business alteration. The results of this analysis demonstrate that the generality of companies gained a high level of performance, although this positive outcome was appeased by the graded decrease of the average values of performance indicators during the period interested. The companies that attain the best performances belong to longstanding business groups, are medium-large sized and located in the north of the country. Concerning business diversification, in the first two years, the specialized companies outperformed the distributed companies (Capece et al., 2010).

Negnevitsky's paper focuses on the experimental results of cluster analysis using self-regulating neural networks to identify failed banks. The report initially defines the main causes and probabilities of bank failures. Then an application of a self-regulating neural network is shown and the results of the study are presented. The findings of the paper show that a self-organizing neural network is a powerful tool for identifying potentially failed banks. Finally, the paper argues some of the restrictions of cluster analysis related to understanding the full meaning of each cluster (Negnevitsky, 2017).

CHAPTER 3: ITEM ANALYSIS

Item analysis; is the process of making the necessary corrections in order to understand whether the results gained from the application of the substances included in a test work according to the selected criteria, and if not, to understand the possible reasons for this and to serve the purpose. Item analysis is mainly concerned with the selection of substances to be tested or whether the substances are qualified or not. In order to analyze the substances to be taken into the final form of the test to be used for a specific purpose, first of all, pre-application and application results must be obtained (Thompson and Levitov, 1985).

The purpose of using item analysis in this study is to observe the effect of the difficulty levels of the questions on clustering algorithms and to observe the effect of clusters emerging as a result of clustering analysis on the homogeneity conditions. In addition, it is aimed to compare the clusters that emerged as a result of the previous analyzes with the clusters produced by the item analysis and to show the effect of the item analysis with the similarities of the members of the clusters. As a result of the item analysis, the weights of the questions are given in Appendix-1.

The following two main factors are effective in determining the method to be used in item analysis:

- Test scoring method
- Whether the substance analysis group is similar to the final form of the test

The item analysis process operates as follows:

- Exam results are collected and ranked from highest to lowest. 27% of the results are taken from the highest and lowest slice of the results. Excluded parts are not included in the analysis (Kelley, 1939).
- In the upper and lower groups, the answers given to those items are considered as inaccessible and unresponsive. The result of the count is shown on a table.
- The percentages of the correct answer in the upper and lower groups include the difficulty of item (p) and the discriminatory power of the substance (r) (Crocker and Algina, 1986).

$$\text{The Difficulty of Item (p)} = \frac{D_U + D_L}{2N}$$

$$\text{The Discriminatory Power of the Substance (r)} = \frac{D_U - D_L}{N}$$

- The values (p) and (r) found to provide information about how the substance works with the answers given. Substances with values of (p) and (r) of 0.5 and around are good substances (Wiersma and Jurs, 1990).

An example table for substance analysis is given below.

As can be seen in Table 3.1, a test applied to 100 people was examined as an example. For the 100 students, the upper group and the lower group were taken as 27% and there were 27 students in both groups. Other students were excluded. 1. 25 people from the upper group and 15 people from the lower group answered the question correctly. When the item difficulty and discriminant power formulas were applied, the item difficulty value was found to be 0.74 and item discrimination power was 0.36 for Question 1. Interpretation of substance difficulty and substance discrimination power is given in Table 3.1.

Table 3.1. Application of a Sample Item Analysis

The test applied to 100 students	Question 1 Number of correct answers	Question 2 Number of correct answers
The Upper Group (D_U) (The First %27)	25	20
The Lower Group (D_L) (The Last %27)	15	15
The Difficulty of Item(p)	$p = \frac{25+15}{54} = 0,74$	$p = \frac{20+15}{54} = 0,64$
The Discriminatory Power of the Item (r)	$r = \frac{25-15}{54} = 0,36$	$r = \frac{20-15}{54} = 0,19$

When Table 3.2 is examined, item difficulty index ranges and equivalents are seen.

- Very difficult if substance difficulty index value is between 0.00 and 0.20
- Hard if the item difficulty index value is between 0.20 and 0.40

- Medium If the item difficulty index value is between 0.40 and 0.60
- Easy if the item difficulty index value is between 0.60 and 0.80
- If the item difficulty index value is between 0.80 and 1.00, it is interpreted as very easy.

Table 3.2. Item difficulty index values and interpretation

The Difficulty Index Of Item	Interpretation
0,00-0,20	Very Hard
0,20-0,40	Hard
0,40-0,60	Medium
0,60-0,80	Easy
0,80-1,00	Very Easy

Table 3.3. Item discrimination index values and interpretation table (Ebel and Frisbie, 1986).

The Discriminatory Power of the Item	Evaluation of Item
0,40 and bigger than 0,40	Very good item
0,30-0,39	Pretty good item
0,20-0,29	An item to be studied
0,19 and less than 0,19	Very weak item

When Table 3.3 is examined, item difficulty index ranges and equivalents are seen.

- If the item discrimination index value is 0.40 or greater, it is a very good item
- If the substance discrimination index value is between 0.30 and 0.39, it is a pretty good item.
- If the substance discrimination index value is between 0.20 and 0.29, it is an item to be studied
- If the item discrimination index value is between 0.19 and 0.00, it is interpreted as a very weak item

When Table 3.4 is examined, it is seen that the item difficulty index and item discrimination index are interpreted together.

- If the item difficulty index has a value greater than 0.90 and the item discrimination index does not have a value, it is preferred if the item has effective training. The weight of the item is 0.20.
- If the item difficulty index is between 0.60 and 0.90 and the item discrimination index is less than 0.20, the item is a typical good ingredient. The weight of the item is 0.40.
- If the item difficulty index is between 0.60 and 0.90 and the item discrimination index is less than 0.20, it is an item that needs to be studied. The weight of the item is 0.60.
- If the item difficulty index is less than 0.60 and the item discrimination index is greater than 0.20, the item is a difficult but discriminating item. The weight of the item is 0.80.
- If the item difficulty index is less than 0,60 and the item discrimination index is less than 0.20, the item is difficult and non-discriminatory. The weight of the item is 1.00.

Table 3.4. Item difficulty index and Item discrimination index Interpretation and weighting table (Ebel and Frisbie, 1986)

The Difficulty Of Item	The Discriminatory Power of the Substance	Interpretation	Weight
More than 0,90	Does not have a value	It is preferred if there is effective instruction.	0,20
0,60-0,90	$r > 0,20$	Typical good ingredient	0,40
0,60-0,90	$r < 0,20$	An item to be studied	0,60
$p < 0,60$	$r > 0,20$	Difficult but distinctive item	0,80
$p < 0,60$	$r < 0,20$	Difficult and non-distinguishing item	1,00

CHAPTER 4: APPLICATION

4.1.Steps Of the Application

4.1.1. Identifying the Problem

In educational institutions that offer foreign language or preparatory classes for foreign languages, students are taken an exemption exam before being admitted to preparatory classes. Educational institutions and especially universities may exempt students from preparatory education as a result of their exams. Many educational institutions and universities use national and international exams such as YDS, YÖKDİL, TOEFL, IELTS in the evaluation of success. Students who get a grade from any of these exams can register directly to the first year without reading the foreign language preparatory class.

Students who do not reach the success criteria determined by the mentioned exams are taught in foreign language preparatory classes of the institutions. These students are below a certain level in terms of foreign language knowledge level. However, not all students may be at the same level of knowledge and skills. Therefore, it is necessary to gather these students in separate classes according to their foreign language knowledge and skill levels. For this purpose, the institutions determine the level of foreign language knowledge and skills of the students by using the exams called leveling or placement exams. In these placement exams, questions of different difficulty levels are asked to determine the knowledge and skill levels of the students. In this exam, which will determine the foreign language proficiency and level of the student at the same time, the questions are asked such that both the level of the student and the level of proficiency are measured at the same time.

As a result of this exam conducted by the School of Foreign Languages, students are divided into courses. When the clustering process is examined from the data mining window, the process is nothing but partitioning. The students whose grades are closest to each other, ie the students most closely related to each other, are divided into one cluster and the other cluster.

When the students in the same cluster are considered in terms of classical clustering, it is assumed that they know the same thing and do not. However, it is seen that the students in these clusters are often not alike. The reason for this is the following error

in the classical cluster: For example, students who get 30 out of a 100-point exam with 75 questions may not have scored 30 points by doing the same questions. Some students may have reached this score by correcting the questions about listening and some students from the reading-related questions correctly. Classical clustering recognizes that all students with 30 points know the same things and do not know the same things.

The aim of this study is to divide students into similar clusters based on item difficulty index and item discrimination index after the item analysis of the questions, rather than the total score they received, and to ensure that packages that will be equipped with more dynamic and more complementary subjects are used instead of the old program prepared during the preparatory education.

4.1.2. Data Preparation

Exam papers of 1681 students who took the placement exam were evaluated by the preparatory school and transferred to the computer environment. While transferring to the computer environment, the correct answers given by each student were coded as 1 and the incorrect answers as 0. The number of variables is equal to the number of questions and is 75. Therefore, the size to be used in data mining is 75. A matrix of 1681 x 75 was created with this number of variables. Those who score 70 or more at the end of the exam are exempted from the preparatory exam. Therefore, the data of 1078 students who scored 69 or less from this exam were used in the study.

4.1.3. Collecting and Harmonizing

The data set used in the study was taken from the School of Foreign Languages. A total of 75 questions were asked in the exam which was attended by 1681 students enrolled in undergraduate and associate degree programs. The exam covers grammar, vocabulary and reading comprehension. The questions that the students answered correctly were coded as “1” and the questions that they left wrong and empty were coded as “0”.

4.1.4. Combining and Cleaning

In the matrix consisting of the information of the students who took the placement exam, the students who left all questions blank were removed from the matrix and there were 1078 students. As written in the previous section, the questions left by other students were coded as 0.

4.1.5. Selecting

Clustering analysis was conducted by using the level examination exam data obtained from foreign language education and preparatory class institution. After the above-mentioned operations are applied, the data are; K-means and hierarchical clustering methods were applied. As the results of the study will be usable, 3 cluster constraints were set by the institution. Because of this condition, the analysis was continued with clustering methods having 3 clusters from the results obtained from the study conducted on raw data. All analyses were performed using SPSS software. The analyzes are summarized in the following subheadings.

4.2. Clustering Analysis of Raw Data

4.2.1. K-Means Clustering on Raw Data

In this study, the data set of 1078 students' data was used. Analyzes were performed using SPSS. The graph obtained as a result of the analysis is given below.

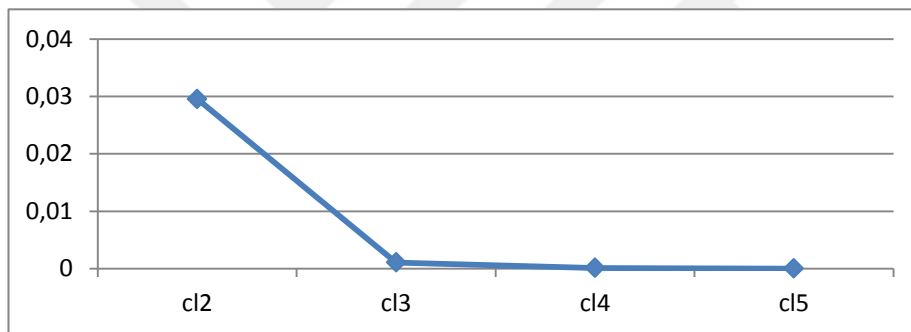


Figure 4.1. K-Means Clustering Wilk's Lambda Values Graph

In cluster studies, Wilk's lambda values are used in deciding the number of clusters. Taking into account the greatest difference between the Wilk's lambda values between the two clusters, the appropriate number of clusters is determined. The Wilk's lambda values calculated for the K-means clustering method are shown in Figure 4-1. As can be seen from Figure 4-1, the greatest difference was experienced in the transition from the second cluster to the third cluster. Subsequent Wilk's lambda values remain stable. As a result, three clusters are selected for the K-means algorithm. In this context, when the K-Means algorithm is used; In the first set, there are 215 students in the 68-42 points range, in the second set there are 441 students in the 41-25 points range and in the third set there are 422 students in the 24-0 points range. When class capacities are assumed to be 20-22 people; 11 classes in the first

cluster, 21 classes in the second cluster and 20 classes in the third cluster. Table 4.1 shows the differences between the obtained results and the current situation.

Table 4.1. Comparison Of Current State and K-Means Algorithm

	Cluster 1 Number of Students (Score Range)	Cluster 2 Number of Students (Score Range)	Cluster 3 Number of Students (Score Range)
Current State	68-60 (19)	59-40 (243)	39-0 (816)
K-Means Algorithm	68-42 (215)	41-25 (441)	24-0 (422)

4.2.2. Hierarchical Clustering on Raw Data

The data set was analyzed by hierarchical clustering methods. The following table shows the number of clusters resulting from the analyzes. The details of the methods are given in the following section, and healthy data could not be obtained according to Cosine and Pearson criteria and they were excluded from the table. Since the Nearest Neighbor clustering method does not provide healthy data, it is excluded from the table. The results obtained are given in Table 4.2.

Table 4.2. Hierarchical clustering methods and Measure's cluster values

	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
Within Group Distance	-	3	-	3	3
Furthest Neighbor	3	3	3	3	3
Median Clustering	3	3	3	3	3
Ward's Method	3	-	3	3	3

4.3. Clustering Analysis of Weighted Data

As a result of the cluster analysis based on raw scores, the tables given above were obtained. Since the study brought 3 cluster constraints for use in the prep classes, the

rest of the study was continued with clustering methods and criteria that gave 3 clusters for interpretation and suggestion. There are a total of 75 questions in the exam administered by the institution. According to the information received from the institution, in order to complete the total score to 100, 1-50 questions were weighted with 1 point and 51-75 questions were weighted with 2. According to the new scores obtained after the weighting process, students who scored 70 or more were excluded from the matrix. The total number of students decreased from 1078 to 1018 after subtraction. The resulting new matrix is 1018x75 in size. The analysis with the new matrix is given below.

4.3.1. K-Means Algorithm

A newly formed matrix clustering analysis was applied with the K-Means algorithm. The results of the analysis are given in the table below.

Table 4.3. The Score Range and Number of Student of K-Means Algorithm

K-Means	Range of Points
1	69-48 (327)
2	47-28 (397)
3	27-0 (294)

When the above table 4.3 is examined in detail:

The K-means algorithm has 3 clusters and the score ranges are 68-48, 47-28 and 27-0 respectively. The number of students in the score ranges is 327, 397 and 294, respectively. When the average number of students in a class is 20-22, the number of classes is 16, 20 and 15, respectively.

4.3.2. Hierarchical Clustering

Within the newly formed matrix due to constraints within Group Distance, Furthest Neighbor, Median Clustering and Ward's Method clustering methods and Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski criteria were applied. The number of clusters obtained as a result of the application is given in the table below. In the following table 4.4, the application of clustering methods together with the criteria and the results are given.

Table 4.4. Number of Cluster of Hierarchical Clustering Methods and Measures

	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
Within Group Distance	-	3	-	3	3
Furthest Neighbor	3	3	3	3	3
Median Clustering	3	3	3	3	3
Ward's Method	3	-	3	3	3

4.3.2.1. Within Group Distance

The within-group distance clustering method was applied to the newly formed matrix with 3 measures. The scores obtained and the number of students is given in the table below.

Table 4.5. The Score Range and Student Number of Within Group Distance's Measures

Within Group Distance	Squared Euclidean Distance	Block	Minkowski
1	69-45 (386)	69-45 (386)	69-45 (386.9)
2	44-19 (497)	44-19 (497)	44-19 (497)
2	18-0 (135)	18-0 (135)	18-0 (135)

When the above table 4.5 is examined in detail:

- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 68-45, 44-19 and 18-0, respectively. The number of students in the score ranges is 386, 487 and 135, respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 25 and 7, respectively.

- Block measure has 3 clusters and the score ranges are 68-45, 44-19 and 18-0 respectively. The number of students in the score ranges is 386, 487 and 135, respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 25 and 7, respectively.
- Minkowski measure has 3 clusters and the score ranges are 68-45, 44-19 and 18-0 respectively. The number of students in the score ranges is 386, 487 and 135, respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 25 and 7, respectively.

When the results are examined, Squared Euclidean Distance, Block and Minkowski measures show similarities in terms of score ranges and number of students.

4.3.2.2. Furthest Neighbor

Furthest Neighbor clustering method was applied to the newly formed matrix with 5 measures. The scores obtained and the number of students is given in the table below.

Table 4.6. The Score Range and Student Number of Furthest Neighbor Measures

Furthest Neighbor	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
1	69-48 (387)	69-48 (387)	69-48 (387)	69-48 (387)	69-48 (387)
2	47-32 (304)	47-32 (304)	47-32 (304)	47-32 (304)	47-32 (304)
3	31-0 (387)	31-0 (387)	31-0 (387)	31-0 (387)	31-0 (387)

When the above table 4.6 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.

- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0, respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- The Chebychev criterion has 3 measure and the score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- Block measure has 3 clusters and score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- The Minkowski measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0, respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.

When the results are examined, Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski criteria show similarities in terms of score ranges and student numbers.

4.3.2.3. Median Clustering

The median Clustering clustering method was applied to the newly formed matrix with 5 measures. The scores obtained and the number of students is given in the table below.

When the above table 4.7 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.

- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0, respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- The Chebychev measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- Block measure has 3 clusters and score ranges are 68-48, 47-32 and 31-0 respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.
- The Minkowski measure has 3 clusters and the score ranges are 68-48, 47-32 and 31-0, respectively. The number of students in the score ranges is 387, 304 and 387 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 19, 15 and 19 for each cluster, respectively.

When the results are examined, Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski measures show similarities in terms of score ranges and student numbers.

Table 4.7. The Score Range and Student Number of Median Clustering's Measures

Median Clustering	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
1	69-48 (387)	69-48 (387)	69-48 (387)	69-48 (387)	69-48 (387)
2	47-32 (304)	47-32 (304)	47-32 (304)	47-32 (304)	47-32 (304)
3	31-0 (387)	31-0 (387)	31-0 (387)	31-0 (387)	31-0 (387)

4.3.2.4. Ward's Method

Ward's Method clustering method was applied to the newly formed matrix with 4 measures. The scores obtained and the number of students is given in the table below.

Table 4.8. The Score Range and Student Number of Ward's Method's Measures

Ward's Method	Euclidean Distance	Chebychev	Block	Minkowski
1	69-39 (490)	69-39 (490)	69-39 (490)	69-39 (490)
2	38-20 (386)	38-20 (386)	38-20 (386)	38-20 (386)
3	19-0 (142)	19-0 (142)	19-0 (142)	19-0 (142)

When the above table 4.8 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 68-39, 38-20 and 19-0, respectively. The number of students in the score ranges is 490, 386 and 142, respectively. Considering that there are 20-22 students in a class, the number of classes is 25, 19 and 7, respectively.
- The Chebychev measure has 3 clusters and the score ranges are 68-39, 38-20 and 19-0, respectively. The number of students in the score ranges is 490, 386 and 142, respectively. Considering that there are 20-22 students in a class, the number of classes is 25, 19 and 7, respectively.
- Block measure has 3 clusters and score ranges are 68-39, 38-20 and 19-0 respectively. The number of students in the score ranges is 490, 386 and 142, respectively. Considering that there are 20-22 students in a class, the number of classes is 25, 19 and 7, respectively.
- The Minkowski measure has 3 clusters and the score ranges are 68-39, 38-20 and 19-0, respectively. The number of students in the score ranges is 490, 386 and 142, respectively. Considering that there are 20-22 students in a class, the number of classes is 25, 19 and 7, respectively.

When the results are examined, Euclidean Distance, Chebychev, Block and Minkowski measures show similarities in terms of score ranges and number of students.

4.4. Clustering analysis on Weighted Data With Item Difficulty and Discrimination Indexes

As a result of item analysis, difficulty and discrimination indexes of the questions were calculated. As a result of the difficulty and discriminative indices, the questions were given values in the range of 0-1. The questions were evaluated according to the p and r values and weighting was made with reference to the interpretations in the third chapter, table-4. Item difficulty and discrimination indices were made using Excel. Weighting and interpretation table is given in Appendix-1. The size of the new matrix remained unchanged and remained at 1018x75. The analysis with the new matrix is given below.

4.4.1. K-Means Algorithm

The newly formed matrix clustering analysis was applied with the K-Means algorithm. The results of the analysis are given in the table below.

Table 4.9. The Score Range and Number of Students of K-Means Algorithm

K-Means	Range of Points
1	49,8-33,2 (298)
2	33-19,6 (371)
3	19,4-0 (349)

When the above table 4.9 is examined in detail:

The K-means algorithm has 3 clusters and the score ranges are 49.8-33.2, 33-19.6 and 19.4-0, respectively. The number of students in the score ranges is 298, 371 and 349, respectively. When the average number of students in a class is 20-22, the number of classes is 15, 18 and 17, respectively.

4.4.2. Hierarchical Clustering

Within the newly formed matrix due to constraints within Group Distance, Furthest Neighbor, Median Clustering and Ward's Method clustering methods and Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski criteria were applied. The number of clusters obtained as a result of the application is given

in the table below. In the following table, the application of clustering methods together with the criteria and the results are given.

Table 4.10. Number of Cluster of Hierarchical Method's and Measures

	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
Within Group Distance	-	3	-	3	3
Furthest Neighbor	3	3	3	3	3
Median Clustering	3	3	3	3	3
Ward's Method	3	-	3	3	3

4.4.2.1. Within Group Distance

The within-group distance clustering method was applied to the newly formed matrix with 3 measures. The scores obtained and the number of students is given in the table below.

Table 4.11. The Score Range and Student Number of Within Group Distance's Measures

Within Group Distance	Squared Euclidean Distance	Block	Minkowski
1	49,8-28,2 (422)	49,8-36,6 (215)	49,8-36,6 (215)
2	28-9,8 (492)	36,2-22,2 (368)	36,2-22,2 (368)
3	9,4-0 (104)	22-0 (435)	22-0 (435)

When the above table 4.11 is examined in detail:

- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 49.8-28.2, 28-9.8 and 9.4-0, respectively. The number of students in the score ranges is 422, 492 and 104, respectively. Considering that there are 20-22 students in a class, the number of classes is 21, 25 and 5, respectively.
- Block measure has 3 clusters and the score ranges are 49.8-36.6, 36.2-22.2 and 22-0 respectively. The number of students in the score ranges is 215, 368 and 435, respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 18 and 22 respectively for each cluster.
- Minkowski measure has 3 clusters and the score ranges are 49.8-36.6, 36.2-22.2 and 22-0 respectively. The number of students in the score ranges is 215, 368 and 435, respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 18 and 22 respectively for each cluster.

When the results are examined, Squared Euclidean Distance Block and Minkowski measures differ in terms of score ranges and the number of students. Block and Minkowski measures are similar to each other in terms of score ranges and the number of students.

4.4.2.2. Furthest Neighbor

Furthest Neighbor clustering method was applied to the newly formed matrix with 5 measures. The scores obtained and the number of students is given in the table below.

When the above table 4.12 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-24,6 and 24,4-0, respectively. The number of students in the score ranges is 215, 303 and 500 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 15 and 25 for each cluster, respectively.
- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-24,6 and 24,4-0, respectively. The number of students in the score ranges is 215, 303 and 500 respectively. Considering

that there is an average of 20-22 students in a class, the number of classes is 11, 15 and 25 for each cluster, respectively.

- The Chebychev measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-24,6 and 24,4-0, respectively. The number of students in the score ranges is 215, 303 and 500 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 15 and 25 for each cluster, respectively.
- The Block measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-24,6 and 24,4-0, respectively. The number of students in the score ranges is 215, 303 and 500 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 15 and 25 for each cluster, respectively.
- The Minkowski has 3 clusters and the score ranges are 49,8-36,6, 36,2-24,6 and 24,4-0, respectively. The number of students in the score ranges is 215, 303 and 500 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 15 and 25 for each cluster, respectively.

When the results are examined, Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski measures show similarities in terms of score ranges and student numbers.

Table 4.12. The Score Range and Student Number of Furthest Neighbor Measures

Furthest Neighbor	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
1	49,8-36,6 (215)	49,8-36,6 (215)	49,8-36,6 (215)	49,8-36,6 (215)	49,8-36,6 (215)
2	36,2-24,6 (303)	36,2-24,6 (303)	36,2-24,6 (303)	36,2-24,6 (303)	36,2-24,6 (303)
3	24,4-0 (500)	24,4-0 (500)	24,4-0 (500)	24,4-0 (500)	24,4-0 (500)

4.4.2.3. Median Clustering

The median Clustering clustering method was applied to the newly formed matrix with 5 measures. The scores obtained and the number of students is given in the table below.

Table 4.13. The Score Range and Student Number of Median Clustering's Measures

Median Clustering	Euclidean Distance	Squared Euclidean Distance	Chebychev	Block	Minkowski
1	49,8-36,6 (215)	49,8-31,6 (342)	49,8-36,6 (215)	49,8-36,6 (215)	49,8-36,6 (215)
2	36,2-15,8 (575)	31,4-15,8 (448)	36,2-15,8 (575)	36,2-15,8 (575)	36,2-15,8 (575)
3	15,4-0 (228)	15,4-0 (228)	15,4-0 (228)	15,4-0 (228)	15,4-0 (228)

When the above table 4.13 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-15,8 and 15,4-0, respectively. The number of students in the score ranges is 215, 575 and 228 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 29 and 11 for each cluster, respectively.
- The Squared Euclidean Distance measure has 3 clusters and the score ranges are 49,8-31,6, 31,4-15,8 and 15,4-0, respectively. The number of students in the score ranges is 342, 448 and 228 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 17, 22 and 11 for each cluster, respectively.
- The Chebychev measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-15,8 and 15,4-0, respectively. The number of students in the score ranges is 215, 575 and 228 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 29 and 11 for each cluster, respectively.

- Block measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-15,8 and 15,4-0, respectively. The number of students in the score ranges is 215, 575 and 228 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 29 and 11 for each cluster, respectively.
- The Minkowski measure has 3 clusters and the score ranges are 49,8-36,6, 36,2-15,8 and 15,4-0, respectively. The number of students in the score ranges is 215, 575 and 228 respectively. Considering that there is an average of 20-22 students in a class, the number of classes is 11, 29 and 11 for each cluster, respectively.

When the results are examined, Euclidean Distance, Chebychev, Block and Minkowski measures show similarities in terms of score ranges and student numbers. But Squared Euclidean Distance is not similar to the other four measures in terms of score ranges and student numbers.

4.4.2.4. Ward's Method

Ward's Method clustering method was applied to the newly formed matrix with 4 measures. The scores obtained and the number of students is given in the table below.

Table 4.14. The Score Range and Student Number of Ward's Method's Measures

Ward's Method	Euclidean Distance	Chebychev	Block	Minkowski
1	49,8-37,2 (193)	49,8-37,2 (193)	49,8-37,2 (193)	49,8-37,2 (193)
2	37-22,2 (390)	37-22,2 (390)	37-22,2 (390)	37-22,2 (390)
3	22-0 (435)	22-0 (435)	22-0 (435)	22-0 (435)

When the above table 4.14 is examined in detail:

- The Euclidean Distance measure has 3 clusters and the score ranges are 49,8-37,2, 37-22,2 and 22-0, respectively. The number of students in the score ranges is 193, 390 and 435, respectively. Considering that there are

20-22 students in a class, the number of classes is 9, 19 and 22, respectively.

- Chebychev measure has 3 clusters and the score ranges are 49,8-37,2, 37-22,2 and 22-0, respectively. The number of students in the score ranges is 193, 390 and 435, respectively. Considering that there are 20-22 students in a class, the number of classes is 9, 19 and 22, respectively.
- Block measure has 3 clusters and the score ranges are 49,8-37,2, 37-22,2 and 22-0, respectively. The number of students in the score ranges is 193, 390 and 435, respectively. Considering that there are 20-22 students in a class, the number of classes is 9, 19 and 22, respectively.
- The Minkowski measure has 3 clusters and the score ranges are 49,8-37,2, 37-22,2 and 22-0, respectively. The number of students in the score ranges is 193, 390 and 435, respectively. Considering that there are 20-22 students in a class, the number of classes is 9, 19 and 22, respectively.

When the results are examined, Euclidean Distance, Chebychev, Block and Minkowski measures show similarities in terms of score ranges and number of students.

CHAPTER 5: COMPARISON of the RESULTS of the METHODS

5.1. Comparison of the Results of the K-Means Algorithm

The results obtained by applying the K-Means algorithm to the data sets are shown in Table 5.1.

Table 5.1. Comparison of the K-Means Algorithm's Analyzes Results

K-Means Algorithm	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-42(215) 11	69-48 (327) 16	49,8-33,2 (298) 15
2	59-40 (243) 12	41-25 (441) 22	47-28 (397) 20	33-19,6 (371) 18
3	39-0 (816) 41	24-0 (422) 21	27-0 (294) 15	19,4-0 (349) 17

When Table 5.1 is analyzed, the K-Means algorithm is applied to Raw Data, Weighted Data, and Item difficulty index. The score ranges obtained as a result of the application, the number of students and the number of classes are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the analysis made with raw data, the score range was 26 and the number of students was 215 while the number of classes increased to 11. As a result of the analysis with weighted data, the number of students increased to 327 and the number of classes increased to 16, although the score range decreased to 21. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 16.6, the number of students to 298 and the number of classes decreased to 15.
- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis made with raw data, the score range decreased to 16, but the number of students increased to 441 and the number of classes increased to

22. As a result of the analysis with weighted data, the score range increased to 19, but the number of students decreased to 397 and the number of classes decreased to 20. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 13.4, the number of students decreased to 371 and the number of classes decreased to 18.

- In the current situation, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis made with raw data, the score range decreased to 24, the number of students decreased to 422 and the number of classes decreased to 21. As a result of the analysis with weighted data, the score range increased to 27, but the number of students decreased to 294 and the number of classes decreased to 15. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 19.4, but the number of students increased to 349 and the number of classes increased to 17.

When the K-Means Algorithm was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution became more stable and the clusters became more homogeneous. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 9%, the similarity rate of the raw data to the weighted data was 46.97% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 99%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 10%, the similarity rate of the raw data to the weighted data was 67.7%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 91.3%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 51.7%, the similarity ratio of the raw data to the weighted data was 99.6%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 83.6%. The effect of weighting and item analysis on cluster analysis was clearly seen.

5.2. Comparison of the Results Obtained by Applying Hierarchical Clustering Methods

The results obtained by applying Hierarchical Clustering Methods are shown and interpreted in tables in this section. Different criteria that give similar results for the same hierarchical clustering method are interpreted under the same table and title.

5.2.1. Comparison of the Results of the Squared Euclidean Distance Measure of the Within Group Distance Clustering Method

The results obtained by applying the Squared Euclidean Distance measure of the Within Group Distance clustering method to the data sets are shown in Table 5.2.

When the Table 5.2 is examined, the Squared Euclidean Distance measure of Within Group Distance clustering method was applied to the raw data, weighted data and item difficulty index weighted data set, and the resultant scores, student numbers, and class numbers were given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the raw data analysis, the score range was 27 and the number of students increased to 240 while the number of classes increased to 12. As a result of the analysis with weighted data, the score range has increased to 24, the number of students has increased to 386 and the number of classes has increased to 29. As a result of the analysis made with the data set obtained from item analysis, the score range increased to 21.6, the number of students increased to 422 and the number of classes increased to 21.
- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis with raw data, the range of points decreased to 16, but the number of students increased to 443 and the number of classes increased to 22. As a result of the analysis with weighted data, the score range increased to 25, the number of students increased to 497 and the number of classes increased to 25. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 18.2, the number of students decreased to 492, but the number of classes remained 25.

- In the current situation, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis made with raw data, the score range decreased to 23, the number of students to 395 and the number of classes decreased to 20. As a result of the analysis with weighted data, the score range decreased to 18, the number of students to 135 and the number of classes decreased to 7. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 9.4, the number of students to 104 and the number of classes decreased to 5.

Table 5.2. Comparison of the Results of the Squared Euclidean Distance Measure of the Within Group Distance Clustering Method

Within Group Distance with Squared Euclidean Distance	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-41(240) 12	69-45 (386) 19	49,8-28,2 (422) 21
2	59-40 (243) 12	40-24 (443) 22	44-19 (497) 25	28-9,8 (492) 25
3	39-0 (816) 42	23-0 (395) 20	18-0 (135) 7	9,4-0 (104) 5

When the Within Group Distance was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution is still in the form of a pyramid. However, a decrease in the number of students between clusters was observed. In addition, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 7.9%, the similarity rate of the raw data to the weighted data was 46.6% and the weighted data according to

the difficulty of the item and the data with discriminant index were calculated as 91.46%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 4.9%, the similarity rate of the raw data to the weighted data was 47.6%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 93.5%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 48.4%, the similarity ratio of the raw data to the weighted data was 34.1%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 99%. The effect of weighting and item analysis on cluster analysis was clearly seen.

5.2.2. Comparison of the Results of the Block and Minkowski

Measures of the Within Group Distance Clustering Method

The results obtained by applying the Block and Minkowski measures of the Within Group Distance clustering method to the data sets are shown in Table 5.3.

When the Table 5.3 is examined, the score ranges, the number of students and the number of classes obtained as a result of the application of the Block and Minkowski measures of Within Group Distance to Raw data, Weighted Data, and Item Difficulty Index are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the raw data analysis, the score range was 27 and the number of students increased to 240 while the number of classes increased to 12. As a result of the analysis with weighted data, the score range has increased to 24, the number of students has increased to 386 and the number of classes has increased to 29. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 13.2, the number of students to 215 and the number of classes decreased to 11.
- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis with raw data, the range of points decreased to 16, but the number of students increased to 443 and the number of classes increased to 22. As a result of the analysis with weighted data, the score range has increased to 25, the number of students to 497 and the number of classes to

25. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 14, the number of students to 368 and the number of classes decreased to 18.

- In the current situation, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis made with raw data, the score range decreased to 23, the number of students to 395 and the number of classes decreased to 20. As a result of the analysis with weighted data, the score range decreased to 18, the number of students to 135 and the number of classes decreased to 7. As a result of the analysis performed with the data set obtained from item analysis, the score range increased to 22, the number of students to 435 and the number of classes to 22.

Table 5.3. Comparison of the results of the Block and Minkowski Measures of the Within Group Distance Clustering Method

Within Group Distance with Measures	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-41(240) 12	69-45 (386) 19	49,8-36,6 (215) 11
2	59-40 (243) 12	40-24 (443) 22	44-19 (497) 25	36,2-22,2 (368) 18
3	39-0 (816) 41	23-0 (395) 20	18-0 (135) 7	22-0 (435) 22

When the K-Means Algorithm was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution is still in the form of a pyramid. However, a decrease in the number of students between clusters was observed. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw

data for the first cluster was 7.9%, the similarity rate of the raw data to the weighted data was 46.6% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 55.7%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 4.9%, the similarity rate of the raw data to the weighted data was 47.6%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 53.5%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 48.4%, the similarity ratio of the raw data to the weighted data was 34.1%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 31%. The effect of weighting and item analysis on cluster analysis was clearly seen.

5.2.3. Comparison of the Results of the Measures of the Furthest Neighbor Clustering Method

The results obtained by applying the five measures which are Euclidean Distance, Squared Euclidean Distance, Chebychev, Block and Minkowski of Furthest Neighbor clustering method are same. Therefore, the results of the measures are shown and interpreted on the same table. The results are shown in Table 5.4.

When the Table 5.4 is examined, the score ranges, the number of students and the number of classes obtained as a result of the application of the measures of Furthest Neighbor to Raw Data, Weighted Data, and Item Difficulty Index are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the analysis with raw data, the score range was 20 and the number of students increased to 106 while the number of classes increased to 5. As a result of the analysis with weighted data, the score range has increased to 21, the number of students to 327 and the number of classes to 16. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 13.2, the number of students to 215 and the number of classes decreased to 11.
- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis made with raw data, the score range decreased to 15, but the

number of students increased to 372 and the number of classes increased to 18. As a result of the analysis conducted with weighted data, the score range remained 15 while the number of students increased to 304 and the number of classes increased to 15. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 11.6, the number of students decreased to 303, but the number of classes remained as 15.

- At present, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis with raw data, the score range decreased to 31, the number of students to 600 and the number of classes decreased to 30. As a result of the analysis with weighted data, the score range remained as 31, the number of students decreased to 387 and the number of classes decreased to 19. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 24.4, the number of students increased to 500 and the number of classes increased to 25.

Table 5.4. Comparison of the Results of the 5 Measures of the Furthest Neighbor Clustering Method

Furthest Neighbor With Measures	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-48(106) 5	69-48 (327) 16	49,8-36,6 (215) 11
2	59-40 (243) 12	47-32 (372) 18	47-32 (304) 15	36,2-24,6 (303) 15
3	39-0 (816) 41	31-0 (600) 30	31-0 (387) 19	24,4-0 (500) 25

When the Furthest Neighbor clustering method was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item

analysis applications to the data set, the distribution is still in the form of a pyramid. The clustering in the lowest cluster is higher than in other clusters. However, it was observed that there were transitions from the lowest to the other clusters. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 17.9%, the similarity rate of the raw data to the weighted data was 14% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 55.7%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 41.9%, the similarity rate of the raw data to the weighted data was 6.8%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 43.5%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 73.5%, the similarity ratio of the raw data to the weighted data was 22.5%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 27%. The effect of weighting and item analysis on cluster analysis was clearly seen.

5.2.4. Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Median Clustering Method

The results obtained by applying the four measures which are Euclidean Distance, Chebychev, Block and Minkowski of Furthest Neighbor clustering method are the same. Therefore, the results of the measures are shown and interpreted on the same table. The results are shown in Table 5.5.

When the Table 5.5 is examined, the score ranges, the number of students and the number of classes obtained as a result of the application of the four measures of Median Clustering to Raw Data, Weighted Data, and Item Difficulty Index are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the analysis with raw data, the score range was 20 and the number of students increased to 106 while the number of classes increased to 5. As a result of the analysis with weighted data, the score range has increased to 21, the number of students to 327 and the number of classes to 16. As a result of

the analysis performed with the data set obtained from item analysis, the score range decreased to 13.2, the number of students decreased to 215 and the number of classes decreased to 11.

- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis made with raw data, the score range decreased to 15, but the number of students increased to 372 and the number of classes increased to 18. As a result of the analysis conducted with weighted data, the score range remained 15 while the number of students increased to 304 and the number of classes increased to 15. As a result of the analysis performed with the data set obtained from item analysis, the score range increased to 20.4, the number of students increased to 575 and the number of classes increased to 29.
- In the current situation, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis with raw data, the score range decreased to 31, the number of students to 600 and the number of classes decreased to 30. As a result of the analysis with weighted data, the score range remained as 31, the number of students decreased to 387 and the number of classes decreased to 19. As a result of the analysis performed with the data set obtained from item analysis, the score range decreased to 15.4, the number of students decreased to 228 and the number of classes decreased to 11.

When the Median Clustering method was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution is still in the form of a pyramid. The clustering in the lowest cluster is higher than in other clusters. However, it was observed that there were transitions from the lowest to the other clusters. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 7.25%, the similarity rate of the raw data to the weighted data was 61.7% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 65.75%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 41.9%, the similarity rate of the raw data to the weighted data

was 29.95%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 52.8%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 46.3%, the similarity ratio of the raw data to the weighted data was 97.1%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 58.9%. The effect of weighting and item analysis on cluster analysis was clearly seen.

Table 5.5. Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Median Clustering Method

Median Clustering With Measures	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-48(106) 5	69-48 (327) 16	49,8-36,6 (215) 11
2	59-40 (243) 12	47-32 (372) 18	47-32 (304) 15	36,2-15,8 (575) 29
3	39-0 (816) 41	31-0 (600) 30	31-0 (387) 19	15,4-0 (228) 11

5.2.5. Comparison of the Results of the Squared Euclidean Distance Measure of the Median Clustering Method

The results obtained by applying the Squared Euclidean Distance measure of the Median clustering method to the data sets are shown in Table 5.6.

When the Table 5.6 is examined, the score ranges, the number of students and the number of classes obtained as a result of the application of the Squared Euclidean Distance measure of Median to Raw Data, Weighted Data and Item Difficulty Index are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the analysis with raw data, the score range was 20 and the number of

students increased to 106 while the number of classes increased to 5. As a result of the analysis with weighted data, the score range has increased to 21, the number of students to 327 and the number of classes to 16. As a result of the analysis made with the data set obtained from item analysis, the score range decreased to 17.2, the number of students increased to 342 and the number of classes increased to 17.

- At present, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis made with raw data, the score range decreased to 15, but the number of students increased to 372 and the number of classes increased to 18. As a result of the analysis conducted with weighted data, the score range remained 15 while the number of students increased to 304 and the number of classes increased to 15. As a result of the analysis conducted with the data set obtained from item analysis, the score range increased to 15.6, the number of students increased to 448 and the number of classes increased to 22.
- At present, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis with raw data, the score range decreased to 31, the number of students to 600 and the number of classes decreased to 30. As a result of the analysis with weighted data, the score range remained as 31, the number of students decreased to 387 and the number of classes decreased to 19. As a result of the analysis performed with the data set obtained from item analysis, the score range decreased to 15.4, the number of students decreased to 228 and the number of classes decreased to 11.

When the Median Clustering method was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution is not in the form of a pyramid. Transitions from the lowest to the other clusters were observed. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 17.9%, the similarity rate of the raw data to the weighted data was 14% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 95.6%. For the second

cluster, the similarity ratio of the current situation to the raw data was calculated as 41.9%, the similarity rate of the raw data to the weighted data was 29.95%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 64.5%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 73.4%, the similarity ratio of the raw data to the weighted data was 64.5%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 58.9%. The effect of weighting and item analysis on cluster analysis was clearly seen.

Table 5.6. Comparison of the Results of the Squared Euclidean Distance Measure of the Median Clustering Method

Median With Squared Euclidean Distance	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-48(106) 5	69-48 (327) 16	49,8-31,6 (342) 17
2	59-40 (243) 12	47-32 (372) 18	47-32 (304) 15	31,4-15,8 (448) 22
3	39-0 (816) 41	31-0 (600) 30	31-0 (387) 19	15,4-0 (228) 11

5.2.6. Comparison of the Results of the Measures of the Ward's Method

The results obtained by applying the four measures which are Euclidean Distance, Chebychev, Block and Minkowski of Ward's Method clustering method are the same. Therefore, the results of the measures are shown and interpreted on the same table. The results are shown in Table 5.7.

Table 5.7. Comparison of the Results of the Euclidean Distance, Chebychev, Block and Minkowski Measures of the Ward's Method Clustering Method

Ward's Method with Measures	Current State	Raw Data	Weighted Data	P and R
1	68-60 (19) 1	68-40 (262) 13	69-39 (490) 24	49,8-37,2 (193) 9
2	59-40 (243) 12	39-23 (438) 22	38-20 (386) 19	37-22,2 (390) 19
3	39-0 (816) 41	22-0 (378) 19	19-0 (142) 7	22-0 (435) 22

When the Table 5.7 is examined, the score ranges, the number of students and the number of classes obtained as a result of the application of the measures of Ward's Method to Raw Data, Weighted Data, and Item Difficulty Index are given. If we need to interpret the results in terms of courses:

- In the current situation, the score range is given as 9 in the first level and there are 19 students in this range. The number of classes is given as 1. As a result of the analysis with raw data, the score range was 28 and the number of students increased to 262 while the number of classrooms increased to 13. As a result of the analysis with weighted data, the score range has increased to 31, the number of students to 490 and the number of classes to 24. As a result of the analysis performed with the data set obtained from item analysis, the score range decreased to 12.6, the number of students to 193 and the number of classes decreased to 9.
- In the current situation, the score range is given as 19 in the second level and there are 243 students in this range. The number of classes is 12. As a result of the analysis made with raw data, the score range decreased to 16, but the number of students increased to 438 and the number of classes increased to 22. As a result of the analysis with weighted data, the score range increased to 18 while the number of students decreased to 386 and the number of

classes decreased to 19. As a result of the analysis with the data set obtained from item analysis, the score range decreased to 14.8, the number of students increased to 390 and the number of classes remained as 19.

- In the current situation, the score range is 39 in the third level and there are 816 students in this range. The number of classes is 41. As a result of the analysis made with raw data, the score range decreased to 22, the number of students decreased to 378 and the number of classes decreased to 19. As a result of the analysis with weighted data, the score range decreased to 19, the number of students decreased to 142 and the number of classes decreased to 7. As a result of the analysis performed with the data set obtained from item analysis, the score range increased to 22, the number of students increased to 435 and the number of classes increased to 22.

When Ward's Method clustering method was applied to the data set, the results were compared with the current situation. In the present case, the distribution is in the form of a pyramid. In other words, while the number is lower in the upper groups, an increase is observed towards the lower group. After weighting and item analysis applications to the data set, the distribution is still in the form of a pyramid. The clustering in the lowest cluster is higher than in other clusters. However, it was observed that there were transitions from the lowest to the other clusters. Also, when the similarities of the clusters were examined, the similarity rate of the current situation and the raw data for the first cluster was 7.25%, the similarity rate of the raw data to the weighted data was 41.2% and the weighted data according to the difficulty of the item and the data with discriminant index were calculated as 39.4%. For the second cluster, the similarity ratio of the current situation to the raw data was calculated as 0%, the similarity rate of the raw data to the weighted data was 38.85%, and the similarity rate of the weighted data to the data with the item difficulty and discriminant index was calculated as 18.97%. For the third cluster, the similarity ratio of the current situation to the raw data was calculated as 46.3%, the similarity ratio of the raw data to the weighted data was 37.5%, and the similarity ratio of the weighted data to the data with the item difficulty and discriminant index was calculated as 32.65%. The effect of weighting and item analysis on cluster analysis was clearly seen.

CHAPTER 6: CONCLUSION and SUGGESTIONS

In this study, a clustering analysis was conducted with the data obtained from an educational institution. The raw data were weighted with the values obtained from classical weighting and substance difficulty index and 2 new matrices were obtained. Initially, the number of students decreased from 1078 to 1018 after the weighting process. The raw data set and two new data sets were analyzed using the K-Means algorithm and hierarchical clustering methods. The results of the analysis were compared with the current situation of the institution. As a result of the comparison can be said:

- When the current situation is compared with the analysis results, the structure of the clusters is observed to change. This means that when the item weights of the questions change, the number of students of the clusters changes and the students in the clusters change between the clusters.
- After the weighting process and item analysis, the lower-upper score range decreased compared to the current situation. This means that students in clusters and classrooms are more likely to show similarity to each other in terms of knowledge and accumulation than in the current situation.
- The results obtained with the K-Means Algorithm give a more homogeneous distribution than Hierarchical Clustering Methods. With the results obtained by the K-Means Algorithm, the students of the first set who have taken preparatory education have the opportunity to finish their preparatory education earlier.
- The results of hierarchical clustering methods vary according to method and criteria. Similar results as a current state were obtained, as were the more homogeneous results. It presents multiple scenarios for this institution and gives the opportunity to show flexibility according to the training to be implemented.
- With the gender variable to be added to the data set, the number of men and women in the classes can be balanced.
- The results of these analyses are specific to the data set and methods used, and variability may be observed when the data set and methods used to change.

- When the similarity rates of the clusters were examined due to the change of data sets, it was observed that there were changes in clusters. This clearly demonstrates the effect of weighting and item analysis on cluster analysis.



REFERENCES

- Aaker, D. A., Kumar, V. and Day, G. S., (1997) *Marketing Research*, 5th edn., Canada: John Wiley Sons.
- Akgöbek, Ö. and Çakır, F. (2009) '*Veri Madenciliğinde Bir Uzman Sistem Tasarımı*', *Akademik Bilişim*, 9, pp. 801-806.
- Akpınar, H., (2000) '*Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği*', İ.Ü. İşletme Fakültesi Dergisi, 29(1)
- Albayrak, M., (2008) '*EEG Sinyallerindeki Epileptiform Aktivitenin Veri Madenciliği Süreci ile Tespiti*', Basılmamış Doktora Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü.
- Alpaydın E, (2000) '*Zeki veri madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri*', *Bilişim 2000 Eğitim Semineri Bildiriler Kitabı*, 2000;s.5
- Anderberg, M.R. (1973) *Cluster Analysis for Applications*, New York: Academic Press.
- Anton, H. and Rorres, C. (1994) *Elementary linear algebra : applications version* , 7th edn., New York: Wiley.
- Arslan, H. (2008) '*Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi*'. Yüksek Lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya, 80s.
- Ay, N. and Amari, S. (2015) *A Novel Approach to Canonical Divergences within Information Geometry*. Entropy. 17. 8111-8129. 10.3390/e17127866.
- Bozkır, S. A., Gök, B. and Sezer, E. (2009) '*Öğrenci Seçme Sınavında Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti*', 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09), 5(), pp. 13-15 .
- Brath, A., Montanari, A. and Moretti, G. (2006) '*Assessing the effect on flood frequency of land-use change via hydrological simulation (with uncertainty)*', *Journal of Hydrology*, 324(1-4), pp. 141-153.

Burmaoğlu,S., (2009) '*Satınalma Alternatiflerinin Çok Değişkenli İstatistiksel Yöntemlerle Belirlenmesi: Keskin Nişancı Tüfekleri Üzerine bir Uygulama*', Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 15 (1), pp. 369-382.

Capece, G., Cricelli, L., Di Pillo, F. and Levialdi, N. (2010) '*A cluster analysis study based on profitability and financial indicators in the Italian gas retail market*', Energy Policy, 38(7), pp. 3394-3402.

Cengiz, D. and Öztürk, F. (2012) '*Türkiye'de İllerin Eğitim Düzeylerine Göre Kümeleme Analizi ile İncelenmesi*', Trakya Üniversitesi Sosyal Bilimler Dergisi, 14(1), pp. 69-84.

Černohorská, L., Černohorský, J. and Teplý, P. (2007) '*The Banking Stability in The Czech Republic Based on Discriminant and Cluster Analyses*', Anadolu University Journal of Social Sciences, 7(2), pp. .

Crocker, L. and Algina, J. (1986) *Introduction to classical and modern test theory*, New York: Holt, Rinehart and Winston.

Çelik, Ş. (2013) '*Kümeleme Analizi İle Sağlık Göstergelerine Göre Türkiye'deki İllerin Sınıflandırılması*', İstanbul, Dogus University Journal, 14(2), pp. 175-197.

de Almeida, R., Steiner, M. T. A., dos Santos Coelho, L., Francisco, C. A. C. and Neto, P. J. S. (2019) '*A case study on environmental sustainability: A study of the trophic changes in fish species as a result of the damming of rivers through clustering analysis*', Computers & Industrial Engineering, 135(), pp. 1239-1252.

Dinçer, E. (2006) '*Veri Madenciliğinde K-Means Algoritması ve Tıp Alanında Uygulanması*'. Yüksek Lisans Tezi, Kocaeli, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli, 101s.

Dolgun, Ö. M., Özdemir, G. T. and Oğuz, D. (2009) '*Veri Madenciliğinde Yapısal Olmayan Verinin Analizi:Metin ve Web Madenciliği*', İstatikçiler Dergisi, 2(), pp. 48-58.

Dragomir, E.G., (2010) '*Environmentally sensitive disclosures and financial performance in a European setting*'. Journal of Accounting & Organizational Change. 6. 359-388. 10.1108/18325911011075222.

Dunham, M.H. (2003) *Data Mining: Introductory And Advanced Topics*, New Jersey: Prentice-Hall.

Ebel, R. L. and Frisbie, D. A. (1986) *Essentials of educational measurement*, NJ: Prentice-Hall: Englewood Cliffs.

Everitt, B. and Dunn, G. (1992) *Applied Multivariate Data Analysis*, New York: Oxford Uni. Press.

Gevrekçi, Y., Ataç, F. E., Takma, Ç., Akbaş, Y. and Taşkın, T. (2011) '*Koyunculuk açısından Batı Anadolu illerinin sınıflandırılması*', Kafkas Üniversitesi Veteriner Fakültesi Dergisi, 17(5), pp. 755-60.

Gower, J. C. (1967) '*A comparison of some methods of cluster analysis*', Biometrics, 23(4), pp. 623-637.

Ienco, D., Pensa, G. and Meo, R., (2012) '*From context to distance: learning dissimilarity for categorical data clustering*', ACM Transactions on Knowledge Discovery, 6(1), pp. 1-27.

İnan, O., (2003) '*Veri Madenciliği*', Yüksek Lisans Tezi, Konya, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü

Kalıkov, A., (2006) '*Veri Madenciliği ve Bir E-Ticaret Uygulaması*', Yüksek Lisans Tezi, Ankara, Gazi Üniversitesi, Fen Bilimleri Enstitüsü

Karabulut, M., Gürbüz, M. and Sandal, E. K. (2004) '*Hiyerarşik Kluster (Küme) Tekniği Kullanılarak Türkiye'de İllerin Sosyo-Ekonomik Benzerliklerinin Analizi*', Coğrafi Bilimler Dergisi, 2(2), pp. 65-78.

Kaya, V. and Türkmen, A., (2013) '*Küresel Krizin Üst Orta Gelir Grubu Ülkelere Makro Ekonomik Yansımaları*', Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 4(27), pp. 317-338.

Kelley, T. I. (1939) '*The selection of upper and lower groups for the validation of test items*', Journal of Educational psychology, 30(1), pp. 17-24.

Kuo, R. J., Lin, S. Y. and Shih, C. W. (2007) '*Mining association rules through the integration of clustering analysis and ant colony system for health insurance database in Taiwan*', Expert Systems with Applications, 33(3), pp. 794-808.

Lora, A. T., Santos, J. M. R., Expósito, A. G., Ramos, J. L. M. and Santos, J. C. R. (2007) '*Electricity market price forecasting based on weighted nearest-neighbor techniques*', IEEE Transactions on Power Systems, 22(3), pp. 1294-1301.

Lowsky, D. J., Ding, Y., Lee, D. K., McCulloch, C. E., Ross, L. F., Thistlethwaite, J. R. and Zenios, S. A. (2013) '*A K-nearest neighbor survival probability prediction method*', Statistics in medicine, 32(12), pp. 2062-2069.

Mooi, E-Sarstadt, M. (2011) *A Concise Guide to Market Research*, 1 edn., New York: Springer-Verlag Berlin Heidelberg.

Orhunbilge, N. (2010) *Çok değişkenli istatistik yöntemler*, 1 edn., İstanbul: İstanbul Üniversitesi, İşletme Fakültesi Yayını.

Sariman, G. (2011) '*Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: k-means ve k-medoids kümeleme algoritmalarının karşılaştırılması*', Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 11(21), pp. 1-23.

Savaş, S., Topaloğlu, N. and Yılmaz, M. (2012) '*Veri madenciliği ve Türkiye'deki uygulama örnekleri*', İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21), pp. 1-23.

Singh, D. and Ganju, A. (2006) '*Improvement in the nearest neighbor weather forecast model performance while considering the previous day's forecast for drawing forecast for the following day*', Current Science, 91(12), pp. 1686-1691.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y. and Lendasse, A. (2007) *Methodology for long-term prediction of time series. Neurocomputing.* 70. 2861-2869. 10.1016/j.neucom.2006.06.015.

Şekeroğlu, S. (2010) *Hizmet sektöründe bir veri madenciliği uygulaması*, Doctoral dissertation, Fen Bilimleri Enstitüsü.

Tatlıdil, H. (2002) *Uygulamalı çok değişkenli istatistiksel analiz*, Ankara: Ziraat Matbaacılık A.Ş.

Thompson, B. and Levitov, J. E. (1985) '*Using microcomputers to score and evaluate test items*', Collegiate Microcomputer, 3(), pp. 163-168.

Turanlı, M. and Başar, Ö. D. (2011) '*Sağlıkta Dönüşüm Uygulaması Sonrası Hastane Tercihlerindeki Değişimin İncelenmesi*', Trakya Üniversitesi Sosyal Bilimler Dergisi, 13(1), pp. 95-105.

Uzgören,N., Keçek,G. and Uzgören,E. (2013) '*Türkiye'deki İllerin Beşeri Sermayenin Unsuru Olan Temel Eğitim Göstergeleri Bakımından Sınıflandırılması*', TİSK Akademi, 11(), pp. 119-133.

Wiersma, W. and Jurs, S. G. (1990) *Educational measurement and testing* , 2 edn., Boston, MA: Allyn and Bacon.

Xu, G., Zong, Y. and Yang, Z. (2013) *Applied Data Mining*, New York: CRC Press.

APPENDIX

Appendix 1: Results of The Item Analysis and Interpretations and Weights of Questions

Question	Item Difficulty Indexes (p)	Item Discrimination Indexes (r)	Interpretation	Weight Value
1	0,520547945	0,417808219	Difficult but distinctive item	0,8
2	0,623287671	0,554794521	Typical good ingredient	0,4
3	0,650684932	0,520547945	Typical good ingredient	0,4
4	0,453767123	0,517123288	Difficult but distinctive item	0,8
5	0,470890411	0,695205479	Difficult but distinctive item	0,8
6	0,875	0,174657534	An item to be studied	0,6
7	0,686643836	0,448630137	Typical good ingredient	0,4
8	0,616438356	0,438356164	Typical good ingredient	0,4
9	0,58390411	0,626712329	Difficult but distinctive item	0,8
10	0,571917808	0,705479452	Difficult but distinctive item	0,8
11	0,488013699	0,695205479	Difficult but distinctive item	0,8
12	0,292808219	0,27739726	Difficult but distinctive item	0,8
13	0,363013699	0,219178082	Difficult but distinctive item	0,8
14	0,405821918	0,448630137	Difficult but distinctive item	0,8
15	0,491438356	0,76369863	Difficult but distinctive item	0,8
16	0,179794521	0,29109589	Difficult but distinctive item	0,8
17	0,195205479	0,321917808	Difficult but distinctive item	0,8
18	0,414383562	0,753424658	Difficult but distinctive item	0,8
19	0,452054795	0,732876712	Difficult but distinctive item	0,8
20	0,589041096	0,506849315	Difficult but distinctive item	0,8
21	0,251712329	0,256849315	Difficult but distinctive item	0,8
22	0,114726027	0,113013699	Difficult and non-distinguishing item	1
23	0,196917808	0,270547945	Difficult but distinctive item	0,8
24	0,246575342	0,417808219	Difficult but distinctive item	0,8
25	0,349315068	0,465753425	Difficult but distinctive item	0,8
26	0,289383562	0,393835616	Difficult but distinctive item	0,8

27	0,25	0,376712329	Difficult but distinctive item	0,8
28	0,297945205	0,212328767	Difficult but distinctive item	0,8
29	0,260273973	0,342465753	Difficult but distinctive item	0,8
30	0,077054795	0,113013699	Difficult and non-distinguishing item	1
31	0,284246575	0,273972603	Difficult but distinctive item	0,8
32	0,342465753	0,342465753	Difficult but distinctive item	0,8
33	0,128424658	0,037671233	Difficult and non-distinguishing item	1
34	0,474315068	0,619863014	Difficult but distinctive item	0,8
35	0,193493151	0,29109589	Difficult but distinctive item	0,8
36	0,29109589	0,431506849	Difficult but distinctive item	0,8
37	0,114726027	0,181506849	Difficult and non-distinguishing item	1
38	0,058219178	0,034246575	Difficult and non-distinguishing item	1
39	0,337328767	0,537671233	Difficult but distinctive item	0,8
40	0,465753425	0,630136986	Difficult but distinctive item	0,8
41	0,48630137	0,664383562	Difficult but distinctive item	0,8
42	0,76369863	0,397260274	Typical good ingredient	0,4
43	0,498287671	0,743150685	Difficult but distinctive item	0,8
44	0,599315068	0,719178082	Difficult but distinctive item	0,8
45	0,385273973	0,613013699	Difficult but distinctive item	0,8
46	0,383561644	0,616438356	Difficult but distinctive item	0,8
47	0,232876712	0,315068493	Difficult but distinctive item	0,8
48	0,203767123	0,373287671	Difficult but distinctive item	0,8
49	0,393835616	0,609589041	Difficult but distinctive item	0,8
50	0,33390411	0,469178082	Difficult but distinctive item	0,8
51	0,696917808	0,523972603	Typical good ingredient	0,4
52	0,667808219	0,54109589	Typical good ingredient	0,4
53	0,585616438	0,602739726	Difficult but distinctive item	0,8

54	0,470890411	0,647260274	Difficult but distinctive item	0,8
55	0,455479452	0,705479452	Difficult but distinctive item	0,8
56	0,616438356	0,753424658	Typical good ingredient	0,4
57	0,751712329	0,448630137	Typical good ingredient	0,4
58	0,224315068	0,339041096	Difficult but distinctive item	0,8
59	0,510273973	0,616438356	Difficult but distinctive item	0,8
60	0,287671233	0,363013699	Difficult but distinctive item	0,8
61	0,761986301	0,414383562	Typical good ingredient	0,4
62	0,767123288	0,356164384	Typical good ingredient	0,4
63	0,38869863	0,660958904	Difficult but distinctive item	0,8
64	0,385273973	0,462328767	Difficult but distinctive item	0,8
65	0,393835616	0,582191781	Difficult but distinctive item	0,8
66	0,202054795	0,308219178	Difficult but distinctive item	0,8
67	0,386986301	0,547945205	Difficult but distinctive item	0,8
68	0,292808219	0,332191781	Difficult but distinctive item	0,8
69	0,191780822	0,294520548	Difficult but distinctive item	0,8
70	0,082191781	0,116438356	Difficult and non-distinguishing item	1
71	0,428082192	0,383561644	Difficult but distinctive item	0,8
72	0,301369863	0,48630137	Difficult but distinctive item	0,8
73	0,260273973	0,308219178	Difficult but distinctive item	0,8
74	0,196917808	0,243150685	Difficult but distinctive item	0,8
75	0,306506849	0,530821918	Difficult but distinctive item	0,8