SEPTEMBER 2021

M.Sc. in Industrial Engineering

AYSU BORSÖKEN

REPUBLIC OF TURKEY

GAZİANTEP UNIVERSITY

GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES

DATA ANALYSIS FOR CLUSTERING CARPET MANUFACTURING DEFECTS

M.Sc. THESIS

IN

INDUSTRIAL ENGINEERING

BY

AYSU BORSÖKEN

SEPTEMBER 2021

DATA ANALYSIS FOR CLUSTERING CARPET MANUFACTURING DEFECTS

M.Sc. Thesis

in

Industrial Engineering

Gaziantep University

Supervisor

Assoc. Prof. Dr. Zeynep Didem UNUTMAZ DURMUŞOĞLU

by

Aysu BORSÖKEN

September 2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Aysu BORSÖKEN

ABSTRACT

DATA ANALYSIS FOR CLUSTERING CARPET MANUFACTURING DEFECTS

BORSÖKEN, Aysu
M.Sc. in Industrial Engineering
Supervisor: Assoc. Prof. Dr. Zeynep Didem UNUTMAZ DURMUŞOĞLU
September 2021
77 pages

Nowadays, with the latest technology, the accumulation of data has increased, and it has become more important to transform the stored raw data into information. The transformation of raw data into information increases earnings and can more easily be adapted to the competitive environment. Data mining is an approach based on statistical applications and machine learning algorithms in converting raw data into information. Data mining has been an important research area to find the hidden information/knowledge inside huge amounts of data. In this thesis, quality problems in carpet manufacturing are gathered in a database and the data are analyzed and clustering by using data mining algorithms. The main purpose is clustering carpet manufacturing defects by using different clustering algorithms WEKA software and find out which algorithm will be most suitable for the users. Also, this thesis is focusing on the improvement of data quality in databases with the help of current data cleaning methods. Thus, a deeper perspective on carpet quality problems is targeted and the number of customer complaints is expected to reduce in the long term. The data obtained from the database of a carpet producer were pre-processed and adapted to WEKA 3.9.4 software. Clustering algorithms which are used are partitioning based i.e K-means, Farthest First, Expectation maximization and non-Partitioning based i.e Cobweb. The best performing algorithm is Farthest First algorithm. It is taking less time then other clustering algorithm to find similar clusters through weka tool for quality problem in carpet manufacturing dataset.

Key words: Data Mining, Raw Data, Cluster Analysis, Waikato Environment for Knowledge Analysis Software.

ÖZET

HALI ÜRETİM HATALARININ KÜMELENMESİ İÇİN VERİ ANALİZİ

BORSÖKEN, Aysu
Yüksek Lisans, Endüstri Mühendisliği Bölümü
Danışman: Doç. Dr. Zeynep Didem UNUTMAZ DURMUŞOĞLU
Eylül 2021
77 sayfa

Günümüzde son teknoloji ile birlikte veri birikimi artmış ve depolanan ham verinin bilgiye dönüştürülmesi daha önemli hale gelmiştir. Ham verinin bilgiye dönüştürülmesi kazançları artırmakta ve rekabet ortamına daha kolay adapte edilebilmektedir. Veri madenciliği, ham verilerin bilgiye dönüştürülmesinde istatistiksel uygulamalara ve makine öğrenmesi algoritmalarına dayanan bir yaklaşımdır. Veri madenciliği, büyük miktardaki verinin içindeki gizli bilgiyi/bilgiyi bulmak için önemli bir araştırma alanı olmuştur. Bu tezde halı imalatındaki kalite sorunları bir veri tabanında toplanmış ve veriler veri madenciliği algoritmaları kullanılarak analiz edilmiş ve kümelenmiştir. Temel amaç, farklı kümeleme algoritmaları WEKA yazılımı kullanarak halı imalat hatalarını kümelemek ve hangi algoritmanın kullanıcılar için en uygun olacağını bulmaktır. Ayrıca bu tez, mevcut veri temizleme yöntemleri yardımıyla veri tabanlarında veri kalitesinin iyileştirilmesine odaklanmaktadır. Böylece halı kalite sorunlarına daha derin bir bakış açısı hedeflenmekte ve uzun vadede müşteri şikâyetlerinin azalması beklenmektedir. Bir halı üreticisinin veri tabanından elde edilen veriler ön işleme tabi tutulmuş ve WEKA 3.9.4 yazılımına uyarlanmıştır. Kullanılan kümeleme algoritmaları, bölümleme tabanlı, yani K-ortalama, Önce En Uzak, Beklenti maksimizasyonu ve Bölümlemesiz, yani Örümcek Ağı'dır. En iyi performans gösteren algoritma Farthest First algoritmasıdır. Halı üretim veri setinde kalite problemi için WEKA aracı ile benzer kümeleri bulmak diğer kümeleme algoritmalarına göre daha az zaman almaktadır.

Anahtar Kelimeler: Veri Madenciliği, Ham Veri, Kümeleme Analizi, Waikato Environment for Knowledge Analysis Yazılımı.

''Dedicated to my family''

ACKNOWLEDGEMENTS

First and foremost, I would like to express my respect and love all the people who supported me in completing my thesis.

Foremost, I'd like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Zeynep Didem UNUTMAZ DURMUŞOĞLU for her guidance, patience and support.

I am also grateful to the Department of Industrial Engineering for the knowledge and the experience that I gained during my graduate study.

I am deeply grateful to my beloved parents, especially my mother, father; Cemile BORSÖKEN, M. Neşet BORSÖKEN for all the devotions they made for me.

TABLE OF CONTENTS

Page

ABSTRACT ..... **v**

ÖZET ..... **vi**

ACKNOWLEDGEMENTS ..... **viii**

TABLE OF CONTENTS ..... **ix**

LIST OF TABLES ..... **xiii**

LIST OF FIGURES ..... **xiv**

LIST OF ABBREVIATIONS ..... **xvi**

CHAPTER I: INTRODUCTION **1**

1.1 Background and Objective 1

1.2 Textile Industry and Carpet Manufacturing 2

1.2.1 Weaving 2

1.2.2 Finishing 4

1.2.3 Raw Materials 6

1.2.4 Weft Yarns 6

1.2.4.1 Cotton Weft 6

1.2.4.2 Jute 7

1.2.4.3 Chenille 7

1.2.5 Warp Yarn 7

1.2.6 Pile Yarns 7

1.2.6.1 Polype Yarns 8

1.2.6.2 Poliester Yarns 8

1.2.6.3 Acrylic Yarns 8
1.2.6.4 Viscose 8
1.3. Planning 8
1.4 Quality of Carpet 9
1.4.1 Carpet Pile 9
1.4.2 Carpet Density: 9
1.4.3 Carpet Fiber 9
1.4.4 Carpet Color 9
1.4.5 Carpet Backing 10
CHAPTER II: LITERATURE SURVEY OF DATA MINING APPLICATIONS 11
CHAPTER III: DATA MINING 16
3.1 Definition of Data Mining 16
3.2 History of Data Mining 17
3.3 Data Mining Processes 21
3.3.1 Business Understanding 22
3.3.2 Data Understanding 22
3.3.3 Data Preparation 22
3.3.4 Modelling 23
3.3.5 Evaluation 23
3.3.6 Deployment 23
3.4 Data Mining Methods 25
3.4.1 Classification 25
3.4.2 Regression 26
3.4.3 Deviation Detection 27
3.4.4 Clustering 27

3.4.4.1 K-Means Algorithm 29
3.4.4.2 Hierarchical Clustering 29
3.4.4.3 COBWEB Algorithm 30
3.4.4.4 EM Algorithm 30
3.4.4.5 Farthest First Clustering Algorithm 32
3.4.5 Association Rules 33
3.4.6 Sequential Pattern Discovery 33
3.5 Model Performance Criterion 34
3.6 Data Mining Applications 36
3.7 Data Mining with WEKA 38
3.7.1 WEKA Explorer 39
3.7.2 WEKA Classification Test Options 41
3.7.3 WEKA Experimenter 42
3.7.4 WEKA Knowledge Flow 43
CHAPTER IV: EXPERIMENTAL STUDY 45
4.1 Dataset Description 45
4.2 Source of Defects by the Departments 51
4.3 Preparing the Data for Application 52
4.4. Implementation of Methods 52
4.4.1 COBWEB Algorithm 53
4.4.2 EM Algorithm 55
4.4.3 Farthest First Algorithm 56
4.4.4 K-Means Clustering Algorithm 58
4.3.4 Result Analysis 62
CHAPTER V: DISCUSSION AND CONCLUSION 63
REFERENCES 65

APPENDIX 72

APPENDIX A: Pseudocodes of Online Learning Algorithms 72

APPENDIX B: Arff File 76

CIRRICULUM VITAE 77

LIST OF TABLES

Page

Table 4.1 Dataset Attributes 46

Table 4.2 Data encoding of the class 46

Table 4.3 Data encoding of the Problem Department 47

Table 4.4 Data encoding of the Quality 48

Table 4.5 Data encoding of the Customer 50

Table 4.6 Examples of Quality Defects Originating from Departments 52

Table 4.7 Value of error terms for different number of clusters and starting values 61

Table 4.8 Comparison results of different clustering algorithms 62

LIST OF FIGURES

Page

Figure 1.1 An image from the weaving machines, and reed 3

Figure 1.2 An image from the weaving section 3

Figure 1.3 An image from the chemical finish application 4

Figure 1.4 An image from the finishing section 5

Figure 1.5 An image from the shipping department 6

Figure 3.1 Relation of Data Mining to Other Disciplines 18

Figure 3.2 History of Data Mining 19

Figure 3.3 Phases of the CRISP-DM Reference Model 21

Figure 3.4 Data mining as a step in the process of knowledge discover 24

Figure 3.5 Example of Clustering 28

Figure 3.6 Flow chart for EM algorithm 31

Figure 3.7 Farthest-first traversal 32

Figure 3.8 ROC Curve 36

Figure 3.9 Data Mining Application 37

Figure 3.10 A screenshot of WEKA Application Menu 39

Figure 3.11 A screenshot of WEKA Explorer Menu 40

Figure 3.12 A screenshot of Weka Test Options 42

Figure 3.13 A screenshot of Weka Experimenter 43

Figure 3.14 A screenshot of Weka Knowledge Flow 44

Figure 4.1 A screenshoot of interface created as a result of loading data 53

Figure 4.2 Visualize COBWEB Algorithm 54

Figure 4.3 Result of COBWEB in form of graph 54

Figure 4.4 EM algorithm..... 55

Figure 4.5 Result of EM Algorithm..... 56

Figure 4.6 Farthest First Algorithm 57

Figure 4.7 Result of Farther First Algorithm 57

Figure 4.8 A screenshoot of Clustering with the K-means clustering algorithm..... 59

Figure 4.9 A screenshoot of result of 10 clusters..... 60

Figure 4.10 A screenshoot of visualization screen of 10 clusters..... 60

Figure 4.11 A screenshoot of evaluate Classes into Clusters 61

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BCF	Bulked Continuous Flament
CRM	Customer Relationship Management
CART	Classification and Regression Trees
CDC	Centers for Disease Control and Prevention
CHAID	Chi-squared Automatic Interaction Detector
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
EM	Expectation-Maximization
ID3	Iterative Dichotomiser
IBM	International Business Machines
KDD	Knowledge Discovery in Database
KNN	K-Nearest Neighbors Algorithm
NB	Naïve Bayes
OneR	One Rule
OLAP	Online Analytical Processing
RIPPER	Repeated Incremental Prunning to Produce Error Reduction
ROC	Receiver Operating Characteristic
SGI	Silicon Graphics
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER I

INTRODUCTION

1.1 Background and Objective

Nowadays, there are very complex operations in carpet factories. Also, the factories are producing in a highly competitive environment. The principal hardship for factories is to examine detailed their manufacturing, to prove their specialness, and develop a production process for further advancement. If the carpet industry focusses on the creation of high-quality, high-fashion, style, and performance-based products, customer satisfaction increases in the way. In manufacturing, it is important to get all the necessary data to analyze carpet defects at the starting point of production. Thus, managers can decide easily to organize the operation management and work on the bestseller collection.

The leading companies in the carpet sector are located in Gaziantep at Turkey. These companies have very complex manufacturing and produce carpets for both domestic and abroad customers. Carpet factories that compete with each other in Gaziantep have advantages and disadvantages. One of the advantages is the rapid recognition of the new qualities in the competing company. However, one of the most important disadvantages is a lot of manufacturer options. To increase customer satisfaction and avoid possible quality defects by adhering to the quality feature demanded by the customer, it is much more important that the quality control per each stage of the manufacturing process. For these reasons, carpet quality has become very important for carpet companies.

This study is based on detecting remarkable knowledge from the database of a carpet factory by the execution of data mining techniques and methods. The principal purpose of the research is data analysis for clustering carpet manufacturing defects.

The specific purpose of this thesis is carried out in order to determine the similarities in terms of 5 different features (4 numeric values and their classes). Quality defect data in carpet production are included abrage (weaving), abrage (sourced from the supplier of yarn), design defect, vapor stain, warp yarn, overlock defect, etc. The operation manager wants to learn which characteristics in the available data are the strongest forecasters of carpet quality. Also, the data is being taken into account − is the resulting data sufficient to make reliable forecasting, is it necessary to make any changes in the data collection process, what other data should be collected to increase availability.

Thanks to this thesis are identified the methodological analysis for data mining applications initiated in carpet manufacturing. Also, it provides analyzing the performance of different k-means algorithms of the clustering method on the dataset. In addition, The WEKA software is used on this thesis application.

1.2 Textile Industry and Carpet Manufacturing

In this chapter of the thesis, information about the carpet company will be given. The firm's three basis production areas, raw materials, and planning process will be explained. And then, the quality of the carpet industry will be examined.

This carpet manufacturing company is one of the leading companies in the domestic and foreign carpet market. It is activated in a 60.000 m² closed area in Gaziantep Organized Industrial Zone. Also, the company has a wide range of customers both in Turkey and abroad with its well- assorted products.

There are two main processes in the company:

* Weaving

* Finishing

1.2.1 Weaving

This company has 25 weaving machines. These machines differ due to the number of reed. The annual production capacity of the weaving department is approximately 15 million m² as an area rug.

The weft number is the number of tie knots per 10 centimeters in length. For example; a carpet with a 32*48 model (called Casablanca name in this company) means that this carpet is weaved in a machine with 32 reeds. An image example from the weaving machines and reed is shown in figure 1.1.

Figure 1.1 An image from the weaving machines, and reed

A technician can change the reed number of a machine. This process is not done very frequently. As the number of wefts and reeds increase, the density and quality of a carpet increases. The production capacity of each machine varies according to the model (quality) of the produced carpet at that time. For example; a machine with 16 reeds produces carpets faster and in large m². On the other hand, the production of a machine with 48 reeds is slower. An image example from the weaving section is shown in figure 1.2.

Figure 1.2 An image from the weaving section

There are 3 main yarn types used in the weaving area:

* Weft yarn

* Warp yarn

* Pile yarn

1.2.2 Finishing

The finishing department is one of the manufacturing areas where finishing processes of the woven carpets (semi-products) are done. After weaving, the carpets are first passed through an area called a chemical finish and exert to heat and adhesive application. The adhesive aims to avoid the slipping of the carpet and ensure the best better adhesion of the carpet yarns to each other. An image from the chemical finish application is shown in figure 1.3.

Figure 1.3 An image from the chemical finish application

After this application, carpets are divided according to customers, they are completely processed according to the customer requests. For instance, a lot of American customers want cardboard pipes inside the carpet to avoid destroying the carpet shape. On the other hand, some customers want to sew the label on the back of the carpet, while some customers want a single barcode. This section works specifically and

completely according to customer demands. An image example from the finishing section according to customer demands is shown in figure 1.4.

Figure 1.4 An image from the finishing section

Carpets completed in finishing are sent to the shipping department. So, the finished product is set. An image from the shipping department is shown in figure 1.5.

Figure 1.5 An image from the shipping department

1.2.3 Raw Materials

The company has a lot of raw materials depending on the manufacturing department. There are three different types of yarn used in carpet weaving in the below contents. They are weft yarn, warp yarn, and pile yarn.

1.2.4 Weft Yarns

The weft, sometimes known as the woof, are the horizontal threads. They are threaded over and under the warp threads. The single thread of yarn that goes across the warp is known as a pick. Cotton weft, jute, and chenille are used as weft yarns in the factory.

1.2.4.1 Cotton Weft

Cotton is one of the oldest fibers need in the textiles sector. The cotton weft is mostly used that given domestic customer orders. because a cotton weft is much more white than jute color, it is much more is preferred in Turkey. Domestic suppliers provide cotton weft to the carpet factory.

1.2.4.2 Jute

Jute is one of the most affordable natural fibers, and the most produced and variety of uses in the world after cotton. 70 % of the world's production is covered by India. A large part of the jute is used as a weft yarn for carpet manufacturing.

According to different carpets, models are used different types of jute. 10 different types of jute are used in this firm. These jute types are 13/1 LBS (a unit of mass used in the imperial and US customary systems of measurement), 13/2 LBS, 16/1 LBS, 16/2 LBS, 18/1 LBS, 20/1 LBS, 20/2 LBS, 22/1 LBS, 24/1 LBS, and 26/1 LBS. LBS is one of the weight units for jute raw material. For instance; 22/1 in LBS code; 22 is called the number of jute type, and 1 is called the number of yarn layer. The yarn layer is explained the number of flament numbers in the jute.

1.2.4.3 Chenille

Chenille is also called weft yarn. It is not used chenille on every carpet. Chenille is used to create a pattern on the carpet. Since the raw material of chenille is not used too much, and the stock quantity is low, domestic companies produce.

1.2.5 Warp Yarn

Warp yarn is the yarn in the longitudinal direction of a woven carpet. Warp yarn creates the carpet floor with jute.

1.2.6 Pile Yarns

Pile yarn types are the most important raw materials that enlarge the product range in this firm. Types of pile yarn used in the company are as below:

- Polype
- Poliester
- Acrylic
- Viscose

1.2.6.1 Polype Yarns

BCF is an abbreviation of the term Bulked Continuous Flament. Its raw material is polypropylene. Polype carpets are produced from polypropylene raw materials. They are used the most than other yarns in the weaving.

1.2.6.2 Poliester Yarns

Synthetic fibers are not of natural origin but are produced from chemical compounds. These are produced from polymers formed from chemical elements and compounds. It is one of the most used and important synthetic fibers in textile.

1.2.6.3 Acrylic Yarns

Acrylic yarns are known as short fiber yarns. Most of the acrylic yarns are used for orders of foreign customers. domestic suppliers provide the acrylic yarn.

1.2.6.4 Viscose

Viscose yarn is one of the raw materials taken from abroad. It is usually supplied from India. Its purchase price is higher and supply is difficult. Therefore, the stock is held for the viscose yarn.

1.3 Planning

When the customers give the orders, the planning phase starts. First of all, giving orders are assigned to the relevant machines according to the reed numbers of machines. When the orders are assigned to the relevant machines according to the reed numbers of machines, the following criteria should be observed:

Is the workload quantity on the machines suitable for the shipment date wanted by the customer?

When the machine is chosen which order is to be produced, is the setup between different orders set to minimum time?

Are the raw material and finishing materials (label, carton board pipes, barcode, packing, etc.) available and sufficient?

If they are not available and are taken from another supplier, is the order quantity and due date informed?

The planning manager approves after all of these stages are review, and then occurs a plan. The customer is notified about the deadline of the orders.

1.4 Quality of Carpet

How well a product affects its life can change according to a diversity of factors. As a result, carpet qualities depend on the function of the carpets. When carpet quality has been examined, the primary factors to pay attention that are carpet pile, carpet density, carpet fiber, carpet backing, and carpet color.

1.4.1 Carpet Pile

Pile means to the fabric loops of a carpet the soft surface that's made carpet so durable. If a carpet has a "high pile," it means the yarns on the carpet surface are taller and looser. For instance, the Shaggy model is a well-known high pile option according to the customers. But Low pile carpeting has shorter carpet fibers.

1.4.2 Carpet Density:

The density of a carpet is occurred by finding the weight of the pile yarn on the carpet surface. A carpet that is low-density has risks of "misshape" before it frays. A carpet density of 4,000 to 7,000 weight is thought fit for a lot of commercial spaces.

1.4.3 Carpet Fiber

Carpet fiber is the species of material from which the strands of carpet are made. In carpet manufacturing, there are several different fiber types. Each fiber has its properties, containing some strengths and weaknesses specific to that fiber type.

There are two carpet fibers used for commercial aims: olefin and Nylon fibers. Wool fiber is natural than the others. In general, fibers that are synthetic perform greater strength and resistance to dirtying, on the other, wool has states where choice.

1.4.4 Carpet Color

For carpet fibers coloring, there are two major ways: solution-dyed and stock dyed. In solution-dyed, the fiber is exposed to the color pigment during the process, so the yarn

material is colored. This process provides that are the fibers resistant to fading and preserve stability color, making them suitable when the carpet is exposed to bleach, harsh detergents, or sunlight.

1.4.5 Carpet Backing

The materials on the carpet backing provide to evaluate the performance of the carpet by supplying strength. In carpet manufacturing, polypropylene and jute are used as backing materials.

The polypropylene and jute are resistant and bouncy. Polypropylene performs resistance towards mildew, so it is better appropriate for high humidity applications.

In this part of the thesis, information about the company, carpet manufacturing processes, raw materials used in carpet production, and quality of carpet ındustry are given.

This thesis includes five sections. The following outline is followed respectively;

- In Chapter 2, previous studies about Data Mining applications from different science areas are given.
- In Chapter 3, detailed information about Data Mining and WEKA software is given.
- In Chapter 4, the case study is explained in detail (all computations, comparisons of obtained results, and graphs).
- The epilogue is located at Chapter 5, with discussions of experiment, conclusion.

CHAPTER II

LITERATURE SURVEY OF DATA MINING APPLICATIONS

This section aims to examine data mining studies in the manufacturing areas. In the manufacturing sector, data mining applications are used in many areas such as production time and process, quality optimization, research and development studies, cost of production, stock management in the production process, production planning, supplier selection.

The "K-means" algorithm, which is one of the methods frequently used in data mining, is first put forward in 1955 and is still used effectively although there are many algorithms in the same field [1]. In this study, the hierarchical k-means technique is used. However, in the literature review, studies using both hierarchical and non-hierarchical k-means techniques are included.

Data mining studies have been increasingly used in the literature in recent years. Data mining starts to use in manufacturing in the 1990s [2].

One of the first studies is performed by Irani et al. [3]. Researchers generalize an ID3 (ID3 is an algorithm invented by Ross Quinlan [4], and the algorithm used to originate a decision tree from a dataset.) algorithm that forecasted the outcome of future tests under various. The researchers use successfully in various semiconductor manufacturing operations in process modeling.

A combination of self-organizing map neural networks and rule induction was used by [5] to identify the critical poor yield factors from normally collected wafer manufacturing data and thus increase the yield. To increase the yield of semiconductor manufacturing, Chien et al. [6] developed a framework consisting of Kruskal-wallis test, K-means clustering, and variance reduction splitting criteria. Sebzalli and Wang [7] applied principal component analysis and fuzzy c-means clustering to a refinery

catalytic process to identify operational spaces and develop operational strategies for the manufacture of desired products and to minimize the loss of product during system changeover. Four operational zones were discovered, with three for product grade and the fourth region giving high probability of producing off-specification product. Liao et al. [8] presented fuzzy clustering-based techniques for the detection of welding flaws. A comparative study between fuzzy k-nearest neighbors clustering and fuzzy c-means clustering has been performed.

The study of Huang and Wu [9] protect product quality with the decision tree method. The researchers execute a method of quality product development in the ultra-precision production industry by using data mining techniques. According to ultra-precision optical products, significant influences on the product quality are determined via the decision tree method of data mining. In 2006, Harding et al. [10] examine applications of data mining in production processes, operations, faulty products, decision support, and improvement of product quality. In 2006, Wang [11] says that is the importance of using data mining techniques for manufacturing. Also, Rokach and Maimon [12] realize the importance and necessity of data mining a matter of to develop manufacturing quality.

With the increase of automated data generation and gathering in manufacturing, mining interesting patterns is of utmost concern. Kusiak [13] presented various algorithms and models for data analysis including cluster analysis, precedence analysis and other data mining methods. Lee et al. [14] proposed an intelligent in-line measurement sampling method for process excursion monitoring and control in semiconductor manufacturing. Chip locations were clustered within the wafer through SOM (self-organizing map) classifications. It has been shown that the generated sampling method can be very effective for all defect detection despite small sampling size. Crespo and Weber [15] presented a fuzzy c means clustering based methodology for dynamic data mining and applied it on real life data. In the era of web-based design and manufacturing, web mining is now frequently used to mine useful information.

Customer service support is an integral part of big manufacturing companies that manufacture and market expensive machines and electronic equipment. Hui and Jha [16] investigated the application of data mining techniques to extract knowledge from the customer service database for decision support and fault diagnosis. Both structured

and unstructured data are mined over the world wide web. Proposed data mining technique integrates the neural network, case-based reasoning, rule-based reasoning for classification and clustering purposes. Qian et al. [17] applied clustering algorithms for churn detection via customer profile modelling.

Dudas et al. [18] address the benefits of using the data mining method in production applications. They propose two different applications. The first application is how data mining can be used to discover the effect of process timing and settings on the quality outcome in the casting industry. The other application is the result of multi-objective optimization of a camshaft process. The main of this study is to find the most appropriate dispatching rule settings in the buffers on the line.

Neis [19] submits an approach that is using clustering methods to measure the costs of new versions based on specification products. Firstly, products are clustered as product families. Then, specification products are assigned according to the shortest distance to all other products within the cluster. The cost function is calculated according to the distance between a new version and the reference product.

Koonce and Tsai [20] suggest using decision trees to get information about dispatching rules and factors that affect lead times.

Firstly, dispatching rules associate with lead times for a realistic scenario. The researchers have created a decision tree to learn the factors that affect the lead times of different dispatching rules.

Thanks to the Fourth Industrial Revolution, information technologies are widely used in industries areas. Turker et al. [21] create a simulation model of a production system. the model provides collecting data via sensors in work centers in real-time and determination operating conditions. The researchers compare business strategies according to the delay periods of the jobs thus run the simulation model with the best strategy. They evaluate data with data mining classifiers and determine rules for delayed tasks. The aim of the model is, the expert system estimates whether or not there will be a delay for new orders, and makes a decision to outsource the order's production if needed when an order is accepted.

Advances in smart technologies contribute to solving problems in the manufacturing areas. Mangla et al. [22] present a 3-stage model. The models classify products based on defects (defects or non-defects) and defect types. Their objective is to find possible defects to supply superior quality through data mining techniques. The researchers develop a Multilayer Perceptron algorithm classification with 96% accuracy for the quality of the product by using the main features like defect code, product, customer, production line, production volume, sample quantity, country, and data set. Thus, a Multilayer Perceptron algorithm is developed with 98% performance for re-work quantity prediction model by considering the attributes.

In the markets, production must raise to supply all demands depending on the customers ' increased. In the last years, big data and restricted measurements taking from different production processes and are not been sufficient to handle all daily data. Hamza Saad [23] provides to analyzes data and discovers important information on how to make better production and quality. The researcher uses some algorithms in the industry field. K-Nearest Neighbor, Tree, Support Vector Machine, Random Forests, Artificial Neural Networks, Naive Bayes, and AdaBoost are implemented to can classify data according to attributes.

The Decision tree and Random Forest and AdaBoost algorithms provide the best accuracy and area under the curve that is ROC. A decision tree can handle numerical and categorical data. Thus, a decision tree is an appropriate algorithm because it handles production data.

In an attempt to perform a data mining process, it is initially required to understand the data and determine the process. E.Demir and S.Dincer [24] use Cross-industry Standard Process for data mining algorithm. The researchers forecast the production of faulty goods by using modeling and classification algorithms, and in the model, prefer to use a supervised learning model based on machine learning methods. The feasibility of the model has been measured with findings. As the result, this study provides to predict whether there are faulty products that reduce quality and has been aimed to give a signal to the production line in advance.

Zeng et al. [25] evaluated 25 manufacturing sectors in China in terms of industrial sustainability by dividing them into 4 clusters using the hierarchical clustering method.

From this review, the major areas where data mining is used for deriving useful knowledge include customer service support, yield improvement, Fault diagnostics etc.

CHAPTER III

DATA MINING

3.1 Definition of Data Mining

In the literature, data mining has a lot of different definitions. One of the first articles to use the term "data mining" was published by Michael C. Lovell in 1983 [26]. Data mining can be defined as the process of discovering potentially useful information in large information repositories [27]. In general, data mining is the process of converting data into meaningful information by analyzing the data available.

Data mining that is knowledge discovery process in which relationships and patterns in large and complex data sets are exposed. This should not be considered as data retention to obtain certain inferences. Because data mining uses well-defined algorithms that draw associations and rules by looking at the raw state of the data [28]. The increased use of the database has necessitated the emergence of new technologies to digest large volumes of generated data [29]. Traditional methods and human analyze are not sufficient to be able to make big and complex data meaningful. Knowledge Discovery in Database (KDD) is the processes that produce methodologies and tools to make large data meaningful [30].

Actually, data is an important source for manufacturing if it has been analyzed wisely. Data mining provides to reveal of unknown patterns and trends by digging into big data [31]. As the world grows in mass, data mining is our only hope to unveil the underlying patterns [32].

The researchers use data mining methods to develop better marketing strategies, improve performance, or reduce the costs of business. Also, the method provides to find new behavioral patterns for consumers.

For predictive and descriptive analysis, data mining methods are used:

- Derived model in Data Mining helps to better understand customer behavior, which leads to better and productive future decisions.
- Data Mining is used to find hidden facts by approaching the market that is useful for business but not yet accessible.
- Data Mining is also used to identify the market area, achieve marketing objectives and generate a reasonably good return on investment.
- Data Mining helps reduce operating costs by discovering and identifying potential investment areas.

3.2 History of Data Mining

Data mining is a multi-disciplinary field that finds a solution concerning countless technical fields. There is a relationship between data mining and other disciplines such as “Statistics”, “Artificial Intelligence”, “Machine Learning”, “Pattern Recognition”, and “Data Visualization”.

Data mining analysis deals with three techniques. They are classical statistics, artificial intelligence, and machine learning [33]. In general, Classical statistics is used on data relationships, as well as for numeric data in large databases [34]. For statistical problems, Artificial intelligence (AI) applies as human-thought-like [33]. As a result, machine learning is the combination of them [35]. Data mining benefits from these techniques, and provides some differences from others: culling patterns, depicting trends, and forecasting behavior. The application area of data mining is quite wide. As shown in Figure 3.1, database technology, data visualization, information science, statistics, machine learning, etc. disciplines among these fields.

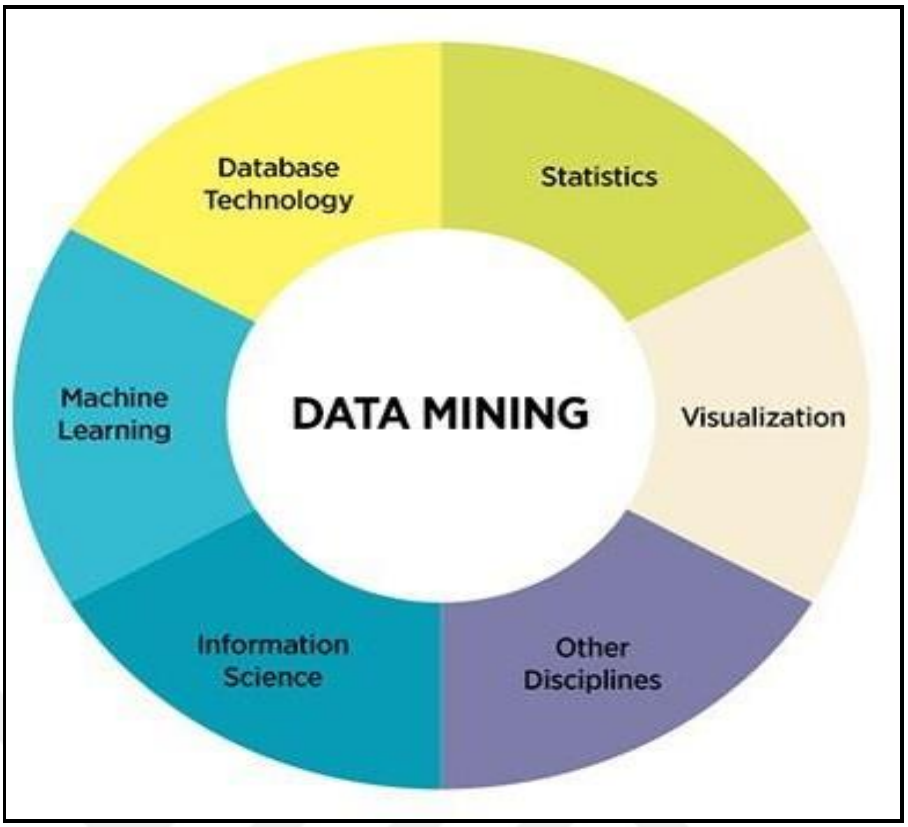

Figure 3.1 Relation of Data Mining to Other Disciplines [36]

In the 1990s, the concept of data mining has been seriously remarked, but the basis actually dates back to the 1700s. In these years, basic statistical methods were used to analyze data. For this reason, it is necessary to search the bases of data mining in classical statistical methods. Major milestones in the history of data mining are shown in figure 3.2.

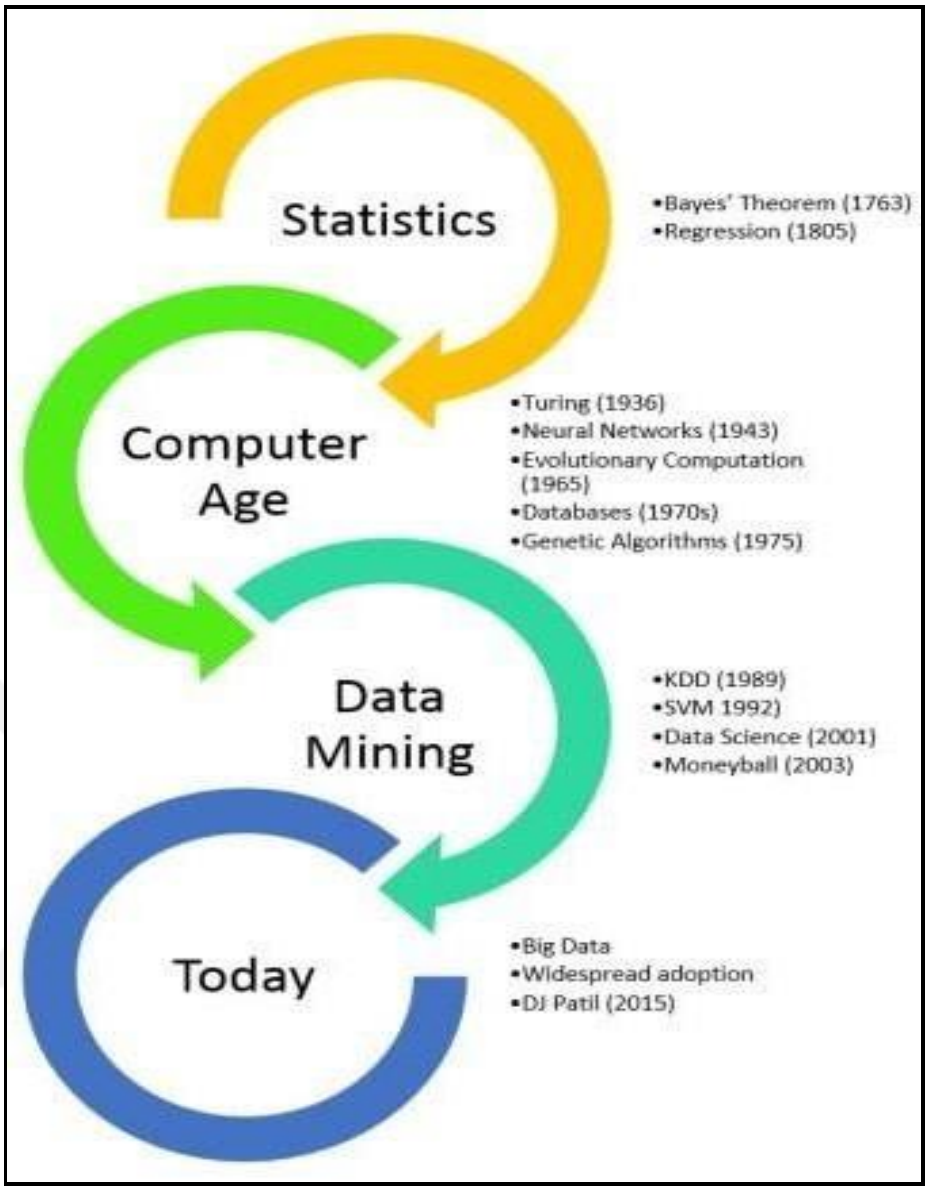

Figure 3.2 History of Data Mining [37]

With increasing data complexity, statistics has become a method of providing services for the evaluation and analysis of data. Statistics forms the basis of data mining.

Statistics is included in study of the process of extracting, analyzing, and using data. Thomas Bayes presented a journal called Bayes' theorem in 1763 [38]. Adrien-Marie Legendre and Carl Friedrich Gauss implemented Regression, which purposes to predict the relationship between variables in 1805 [39].

With the development of computer capabilities, analysis and understanding of data has been easier. Alan Turing published an article called "On Computable Numbers" [40] and submitted a universal machine idea that allowed computers to process data in big quantities. As computers began to improve, concepts of artificial intelligence and machine learning began to be argued. In 1943, Warren McCulloch and Walter Pitts were the first people to produce a theoretical model of the neural network [41]. They have found that the data could be analyzed by computer and artificial intelligence algorithms.

The notion of data mining was called data scanning and data fishing in the 1960s. When the necessary inquiry was made with the help of the computer, it was supposed that the required information could be achieved. Like the tree data structure, a hierarchical data model based on the existence of parent data for each record and the presence of many child records was also established in the 1960s.

The relational database management system emerged towards the end of the 1970s, which is a system that holds the data in rows and columns on tables, and provided data consistency. In 1989, "Knowledge Discovery in Databases" the term is created (KDD) by Gregory Piatetsky-Shapiro. Also, he established the first workshop called KDD [42].

In the 1990s, the term “data mining” occurred in the database community. Also, Bernhard et al. [43] proposed betterment on the original support vector machine. It allows for the creation of nonlinear classifiers.

In 2003, the Oakland Athletics used a statistical, data-driven approach to select players that were cheaper cost. The approach is called Moneyball that is published by Michael Lewis [44]. And then, it changed the way many major league front offices do business. In this case, they assembled a team thanks to bringing them to the 2002 and 2003 playoffs with 1/3 the payroll.

At the White House, DJ Patil was the first Chief Data Scientist in February 2015 [45]. Today, Big Data becomes commonplace with the proliferation of devices capable of collecting data.

3.3 Data Mining Processes

Data mining is a process of decomposing the samples in the information discovery process and making the next step ready is also a part of this process. In order to apply data mining methods, data held in data warehouses must pass through certain stages.

There is a Cross-Industry Standard Process for Data Mining (CRISP-DM) which is commonly used by business supporters [46]. This model involves six stages intended as a cyclical procedure [45- 47]. The cyclical procedure includes business understanding, data understanding, data preparation, Modelling, Evaluation, and deployment as given in figure 3.3.

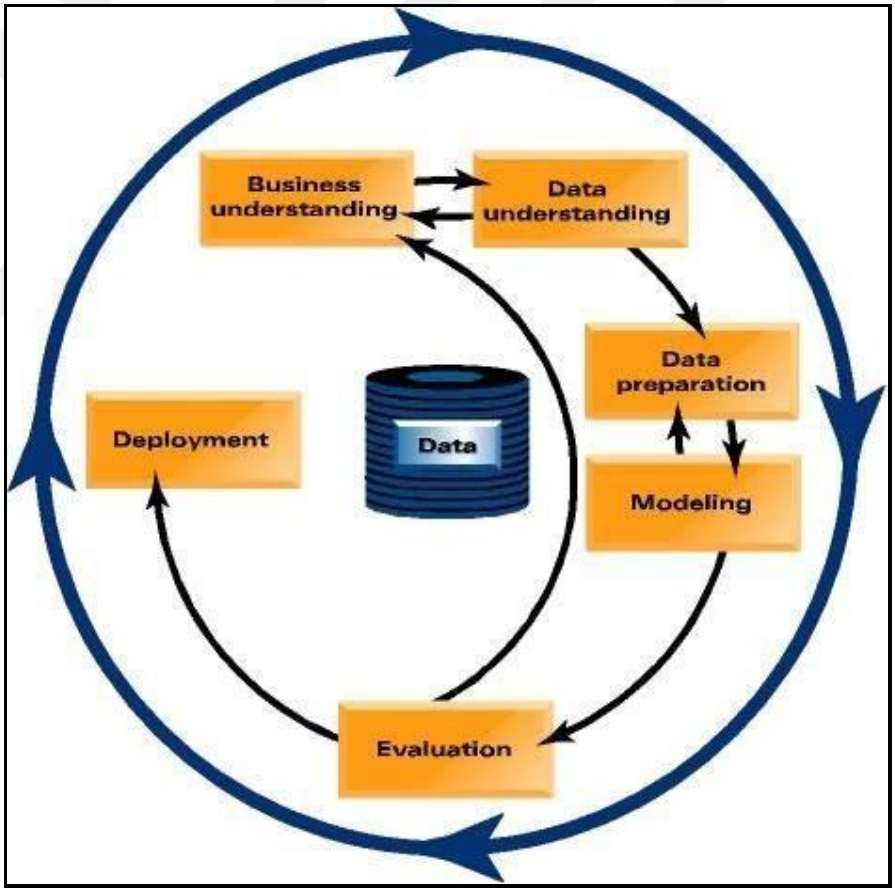

Figure 3.3 Phases of the CRISP-DM Reference Model [48]

3.3.1 Business Understanding

The phase of Business Understanding, which can also be called the stage of understanding the research problem, becomes the first step of the data mining process. Transparently, the objectives of the project as a whole include the process of setting aims and restraints in terms of data mining formulations and establishing a preliminary strategy in this direction. The goal must be focused on the operation problems and must be expressed clearly. As a result, this stage provides agreement on success criteria.

3.3.2 Data Understanding

The phase of data understanding starts with the data collection stage. The data can be acquired from interior and exterior sources. Internal resources are the database of the operation, such as quality defects and orders. External sources are data acquired from outside the enterprise, such as taking information of yarn from suppliers. And then the process continues with the description and the measurement of the data. The goal is able to determine important the data in the study.

3.3.3 Data Preparation

Data preparation phase includes all the arrangements to be processed from the first raw to the final. Determination of necessary conditions and variables for analysis involves processes such as transformation and cleaning of data. The pre-processing stage of data is the most time-consuming phase in the data mining process, because of the size of the data sets.

Collection of Data: Data is collected to explain an analysis. Data collection is a methodical process to able to relevant questions and evaluate the results. Data collection is based on finding out all about a particular subject matter.

Cleaning of Data: This phase provides comprehensibility of data by annihilating erroneous and inconsistent data. It is important and necessary for data mining. In this phase, incorrect entries or exceptional cases are prevented from the resultant defects. In some states, different models can be implemented to complete the missing data, instead of directly extracting the missing data.

Conversion of Data: Data conversion provides the transition of one data format into a new format. It is a technical process that is mostly done by software, although human interference is sometimes used. The goal of the data conversion is to provide interoperability and to maintain all of the data.

Reduction of Data: Big data sets are usually preferred in the data mining works, but this also causes some problems. Especially in some studies, the number of variables is considerably large, and the defect of quality's state and variable numbers have to be reduced to usable numbers. With only certain positions, it is possible to restrict the data by removing certain dimensions from the dataset, or by creating smaller datasets that providing it instead of larger datasets.

3.3.4 Modelling

The modeling phase is the stage in which the appropriate modeling techniques are chosen and implemented several different methods for the problem. When some methods are not appropriate for the problem, the previous stage is returned which is data preparation. Thus, the steps of modeling and data preparation are echoed until the best model is obtained. It is not easy to decide the most useful technique before starting the model establishment studies. It is useful to make studies to get the most appropriate model by using different models and accuracy measures.

3.3.5 Evaluation

This phase is the process of evaluating whether the quality and efficiency of one or more models implemented in the modeling phase and the model conform to the objectives set in the first stage. The evaluation phase includes reviewing the applied data mining processes and provides to evaluate conclusions. It is essential that the model is in line with the specific aims.

3.3.6 Deployment

In the phase of deployment, the information finding at the end of the process is used to solve the problem in the direction of the examined aims. The conclusions are evaluated and interpreted according to real life. The decision model can be either a direct function or a subset of another function. The goal of the model is to provide

knowledge about the given data, but the given data must be organized and shown, and then the data can be used. Also, this phase provides to occur the research report. The reports are substantial for providing that the conclusions of the projects are communicated and, if necessary, repeated by others. If necessary, varies in the systems over time will necessitate monitoring of the established models and re-arrangement. The control is consistently essential to avoid the model from running with incorrect data.

Data mining is described as the discovery of knowledge. The information discovery process is presented in Figure 3.4.

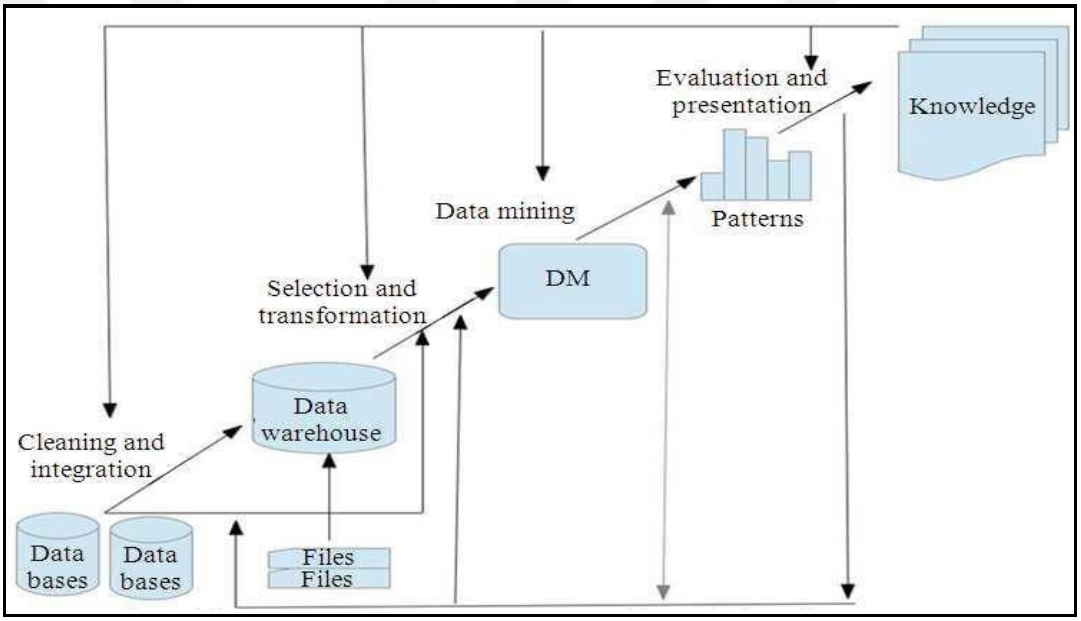

Figure 3.4 Data mining as a step in the process of knowledge discover [49]

- Data cleaning (to eliminate noise and incoherent data)
- Data integration (to correlation numerous data sources)
- Data selection (to take back data related to the examination studies from the database)
- Data transformation (to convert data into forms suitable for mining by

execution summary or combination processes)

- Data mining (The procedure where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to find different patterns representing information based on conditions of interest)
- Knowledge presentation (where information illustration procedures are used to current mined knowledge to users)

3.4 Data Mining Methods

Data mining is the process of analyzing massive volumes of data. The method provides to discover business intelligence so that helps companies solve problems, mitigate risks, and seize new opportunities [50]. Data mining models can be divided into two groups as descriptive and predictive [30, 51, 52]. In the descriptive models, the patterns are described from the given data and the conclusions are displayed to take action. Clustering, Association Rules, and Sequential Pattern Discovery are examples of the descriptive models. In predictive models, the objective is to find out patterns within existing data or to predict a future state of existing data. In predictive models, target groups are specified at the beginning. The available data is analyzed and divided into specified classes and a future situation is predicted. Classification, Regression, and Deviation Detection are predictive models.

3.4.1 Classification

Classification is a method of the data analysis process. It is the process of finding a model that distinguishes data classes. Classification identifies which categories (subpopulations), a new observation belongs to.

Classification is a two-step procedure, involving a learning step (where a classification model is created) and a classification step (where the model is used to forecast class labels for given data) [30, 51].

The initial step is to build a classifier for describing a predetermined set of data classes. Some of the data in the database are chosen randomly and then used as training data. The remaining data is also used as test data. A classification model is established by implementing an algorithm on the training data.

The other step is to try the rules by implementing them to the test data. The model is implemented to the training data if the accuracy of the test data is estimated.

3.4.2 Regression

In the regression analysis, it is important to specify which events are affected. Each variable is a mathematical state of the relationship between one or more different variables. The purpose of the regression is to create a model that can make the most accurate estimate of the relationship between inputs and outputs [32].

The property of the regression is different from the classification is that the estimated dependent variable is a continuous numerical variable. In the analysis, the result is called "dependent variable" and inputs are called "independent variable". When regression analysis is made, variables contained in the model are one or more independent variables and a dependent variable. Dependent and independent variables are measurable. Univariate models are presented as simple linear regression, multiple independent variables, and multiple regression models. Regression analysis is split into linear regression and nonlinear regression according to the equation of the made equation. A linear regression equation is expressed as in equation (3.1):

$$y = a + bx \tag{3.1}$$

The independent variable, x, forms y, and y forms classes. So, the y value is estimated by replacing the x value with the attribute value, i.e., the class to which x belongs is predicted. If the problem has more than one attribute, then the equation called multiple regression is shown as follows in equation (3.2):

$$y = a + b_1 x_1 + b_2 x_2 \tag{3.2}$$

Regression analysis can be used in many areas. Regression analysis can be used to solve many different problems such as operation management, sales prediction, customer assessment, and financial risks.

3.4.3 Deviation Detection

At the most basic of the detection, the data set is split into clusters and the data that does not belong to these clusters are had. So, anomalies are detected. In the anomaly analysis to be applied the dataset that does not have similar behavior characteristics is determined in the dataset [62].

Errors during reading, recording, measuring, application, or calculation lead to deviation data. For example, at the time program is run, a customer's quality that given order may be written as Casablanca quality instead of loft quality. Most of the algorithms of data mining are fronted to minimize or completely remove the effect of abnormal data.

3.4.4 Clustering

Cluster analysis is the method of division a set of data items into subgroups. At each cluster, objects are related to one another, however different from objects in other clusters. A set of clusters subsequent from a cluster analysis is clustering. In this meaning, dissimilar clustering procedures may create different clustering on a similar dataset. The separation is not completed by the researchers, but by the clustering algorithm. Therefore, clustering is beneficial to guide the finding of earlier unidentified groups in the data [51].

Clustering is an important data mining function performed on specified manufacturing data. For instance, in logistics, order picking is routine in distribution center. Before, picking a large set of orders, orders are clustered into batches to accelerate the product movement within the storage zone. Another example is the formation of cell in cellular manufacturing where clustering is used for design of the part families and machine cells simultaneously. A detailed review and study of clustering techniques, application areas are mentioned in [75].

There is a difference between classification models and clustering models. The biggest difference is that the classification models have not been previously described. The clustering model has not a predetermined class. The data taken from the database are grouped according to certain characteristics examined by the researcher who sets up the model. So, clusters are created in different characteristics.

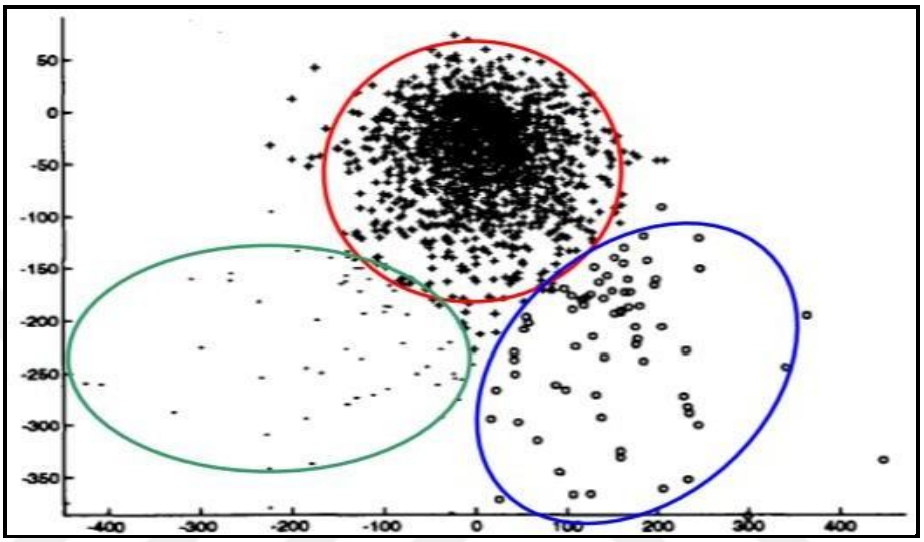

Figure 3.5 Example of Clustering [8]

As you can see in figure 3.5, The goal is that the objects in a very cluster are similar (or related) to 1 different and totally different from (or unrelated to) the objects in different teams.

At sometimes, before other data mining methods are implemented, clustering can be used to narrow or customize records in the database. The aimed data mining method may be implemented to just only one of the clusters. Clustering models are implemented especially when customer characteristics are determined, so it is commonly used in the retail sector. Customers can be grouped based on their profitability, usage rate, or potential in the market. So, in the market, it provides companies to communicate with other customers who have similar characteristics.

3.4.4.1 K-Means Algorithm

This section describes the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance [72]. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step-by-step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering. The pseudocode of the k-means clustering is available in Algorithm 5, Appendix A [73].

The k-means algorithm is the most extensively studied clustering algorithm and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids. The k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

3.4.4.2 Hierarchical Clustering

Hierarchical clustering [74] is one of clustering algorithms and has been used extensively. Hierarchical clustering begins with an initial partition into singleton clusters by successive merging of clusters until all elements belong to the same cluster. All the distances between the items to be clustered have to be calculated in a distance matrix. At each step, joining the two closest items creates a node. Subsequent nodes are created by pairwise joining of items or nodes based on the distance between them, until all the nodes merge into a desired number of clusters. The completeness of

hierarchical clustering is deterministic. A remarkable problem in implementation of the hierarchical clustering is its inability to handle large data sets within a reasonable time and memory resources. The pseudocode of the hierarchical clustering is available in Algorithm 6, Appendix A.

3.4.4.3 COBWEB Algorithm

The COBWEB algorithm was developed by machine learning researchers in the 1980s for clustering objects in an object-attribute data set. The COBWEB algorithm yields a clustering dendrogram called classification tree that characterizes each cluster with a probabilistic description. Cobweb generates hierarchical clustering [77], where clusters are described probabilistically.

COBWEB uses a heuristic evaluation measure called category utility to guide construction of the tree. It incrementally incorporates objects into a classification tree in order to get the highest category utility. And a new class can be created on the fly, which is one of big difference between COBWEB and K-means methods. COBWEB provides merging and splitting of classes based on category utility, this allows COBWEB to be able to do bidirectional search. For example, a merge can undo a previous split. While for K-means, the clustering [78] is usually unidirectional, which means the cluster of a point is determined by the distance to the cluster center. It might be very sensitive to the outliers in the data.

COBWEB has a number of limitations. First, it is based on the assumption that probability distributions on separate attributes are statistically independent of one another. This assumption is, however, not always true because correlation between attributes often exist. Moreover, the probability distribution representation of clusters makes it quite expensive to update and store the clusters. The pseudocode of the cobweb algorithm is available in Algorithm 7, Appendix A.

3.4.4.4 EM Algorithm

The Expectation-Maximization (EM) algorithm [79-80] is a way to find maximum-likelihood estimates for model parameters when your data is incomplete, has missing data points, or has unobserved (hidden) latent variables. It is an iterative way to approximate the maximum likelihood function. While maximum likelihood estimation

can find the “best fit” model for a set of data, it doesn’t work particularly well for incomplete data sets. The more complex EM algorithm can find model parameters even if you have missing data. It works by choosing random values for the missing data points, and using those guesses to estimate a second set of data. The new values are used to create a better guess for the first set, and the process continues until the algorithm converges on a fixed point.

The Expectation-Maximization algorithm aims to use the available observed data of the dataset to estimate the missing data of the latent variables and then using that data to update the values of the parameters in the maximization step.

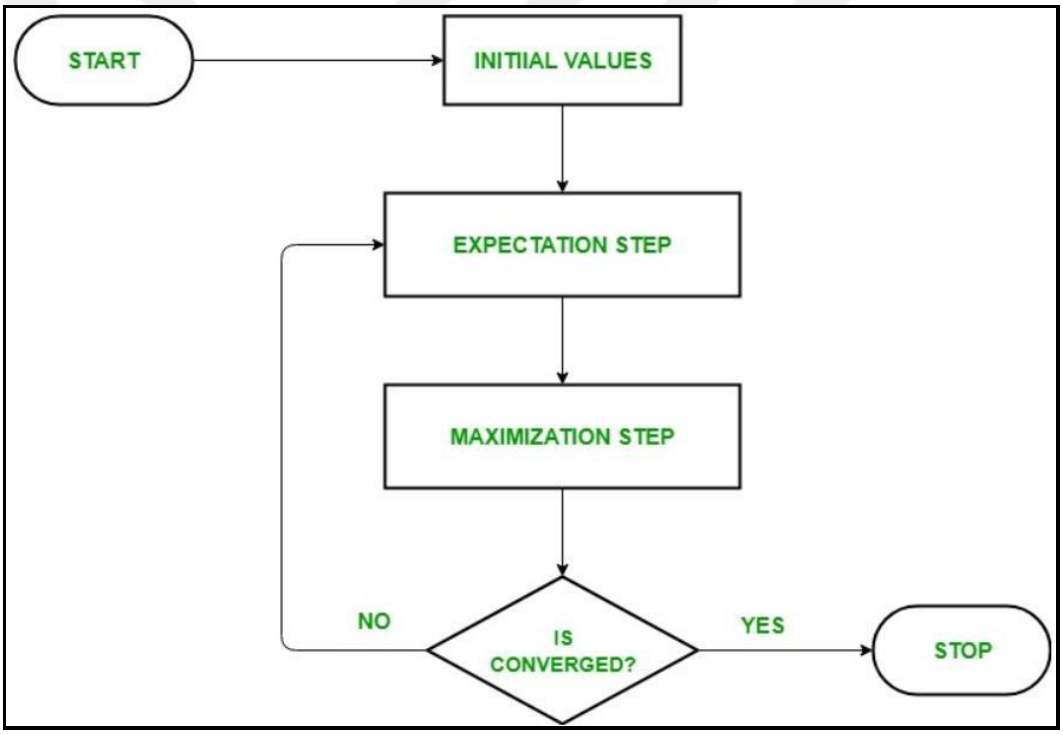

Figure 3.6 Flow chart for EM algorithm

As you can see in figure 3.6, the EM algorithm in a detailed manner:

- **Initialization Step:** In this step, we initialized the parameter values with a set of initial values, then give the set of incomplete observed data to the system with the assumption that the observed data comes from a specific model i.e.,

probability distribution.

- **Expectation Step:** In this step, by using the observed data to estimate or guess the values of the missing or incomplete data. It is used to update the variables.
- **Maximization Step:** In this step, we use the complete data generated in the "Expectation" step to update the values of the parameters i.e., update the hypothesis.
- **Checking of convergence Step:** Now, in this step, we checked whether the values are converging or not, if yes, then stop otherwise repeat these two steps i.e., the "Expectation" step and "Maximization" step until the convergence occurs.

3.4.4.5 Farthest First Clustering Algorithm

Farthest First is a modified of K-Means that places each cluster center in turn at the point further most from the existing cluster center. This point must lie within the data area. This greatly increases the clustering speed in most of the cases since less reassignment and modification is needed [81].

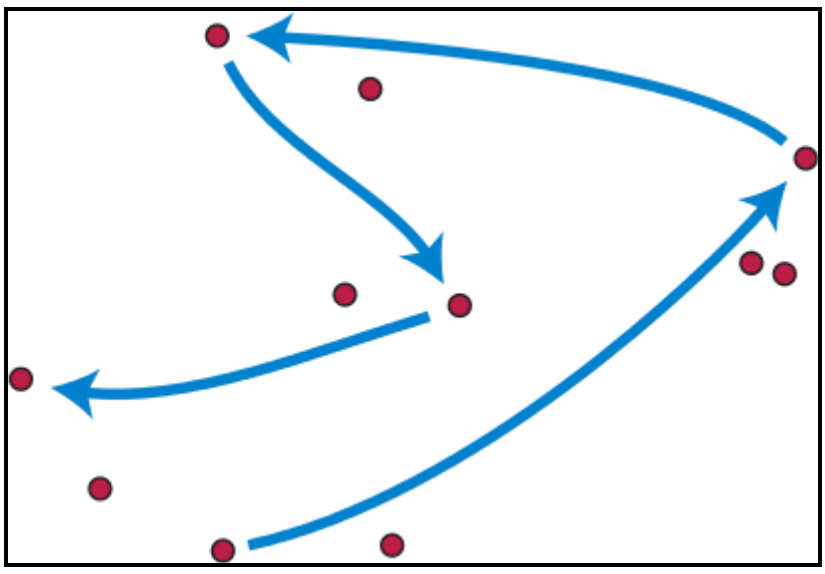

Figure 3.7 Farthest-first traversal

As you can see in the figure 3.7, the first five steps of the farthest-first traversal of a planar point set. The first point is chosen arbitrarily and each successive point is as far as possible from all previously chosen points.

3.4.5 Association Rules

The Association rule can be used to find the relation between the data attributes in the databases. Association rules examine the relations between data attributes and calculate the likelihood that the events occur. Association rules establish links that are not easy to be distinguished in data. These links provide many benefits for companies, ranging from customer satisfaction to strategic decision-making [63].

In general, it uses as a market basket analysis, as its purpose to reveal relationships between data. In this way, many rules may exist in the retail sector data: from placement of products on shelves, creation of advantageous products, preparation of catalogs, and determination of the area to be used.

3.4.6 Sequential Pattern Discovery

Sequence pattern discovery is a data mining technique that explores statistically relevant patterns in sequential data. Connections are constituted by examining the events in certain frequencies within certain time intervals. The fact that ordered sequences are based on events occurring over time distinguishes this algorithm from association rules [51]. An example of a sequential pattern is “Customers who buy a "Canvas" quality carpet are likely to buy a "Brampton" quality carpet within a month.”

Sequence patterns are also a great advantage in a highly competitive industry such as predicting customer behavior. Companies have quite a lot of data utilizing barcode technology. Firms can examine and correlate customer movements using the data which they can achieve via products such as loyalty cards, and turn them into an advantage for themselves by offering customers campaigns. Also, it is possible to use sequential order with these records to analyze the behavior of the customer lost and to implement a lot of ways to keep existing customers.

3.5 Model Performance Criterion

Quality of the model depends on the number of target samples that are properly categorized. Insomuch that, it depends on the number of the target sample that are faultily categorized. The different performance measures used to estimate the performance of model are sensitivity, precision, error rate, F-criterion, and ROC Graphs [64].

In the test conclusion, the performance measures of the results completed can be explained by the confusion matrix. The rows symbolize the real numbers of the samples that used the test set in the confusion matrix, and the columns symbolize the guess of the model [65].

Table 3.1 Confusion Matrix

		Predicted	
		Positive	Negative
Observed	Positive	TP (# of TPs)	FN (# of FNs)
	Negative	FP (# of FPs)	TN (# of TNs)

*TP (True Positive) *FN (False Negative)

*FP (False Positive) *TN (True Negative)

The most common way used of measuring the model performance is checking the accuracy of the models. Accuracy is shown in equation (3.3) can be measured using one or more test data which are different from the training data set [66]. It is that the correct number of classified samples (TP + TN) divided by the total number of samples (TP + TN + FP + FN). The error rate (the number of misclassified samples) is shown in equation (3.4) (FP + FN) is the ratio of the total number of samples (TP + TN + FP + FN).

$$Accuracy = \frac{TP + TN}{TP+FP+FN+TN} \tag{3.3}$$

$$Error\ Rate = \frac{FP + FN}{TP+FP+FN+TN} \tag{3.4}$$

Precision is shown in equation (3.5) is to specify how many of the values we estimated as Positive are actually Positive.

$$Precision = \frac{TP}{TP+FP} \tag{3.5}$$

(3.5) Sensitivity is shown in equation (3.6) is the ratio of the number of correctly estimated positive samples to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP+FN} \tag{3.6}$$

Specificity is shown in equation (3.7) is evidence of the ability of a test to distinguish what is wrong.

$$Specificity = \frac{TN}{TN+FP} \tag{3.7}$$

Precision and sensitivity measures only are not sufficient to make an important evaluation. Predicting both measures collected gives extra precise conclusions. The f-criterion is described for this. The F-measurement is shown in equation (3.8) is the harmonic mean that deal with precision and sensitivity.

$$F - measurement = \frac{2\ x\ Sensitivity\ x\ Precision}{Sensitivity + Precision} \tag{3.8}$$

The ROC (Receiver Operating Characteristic) curve is the standard method to provide the reliability and general accuracy of model tests. In the population, ROC analysis the model in splitting positive groups from negative and the independence of the conclusions from the frequency of positive subjects. ROC analysis is particularly useful in predicting models and assessing other tests since it captures a threshold value between sensitivity and specificity in a continuous range.

ROC curve shows behavior of FPR and TPR at dissimilar classification thresholds. If the classification threshold decreases, more items are classified as positive, so enhancing both False Positives and True Positives. The relation between the true positive rate and the false positive rate is shown in figure 3.8.

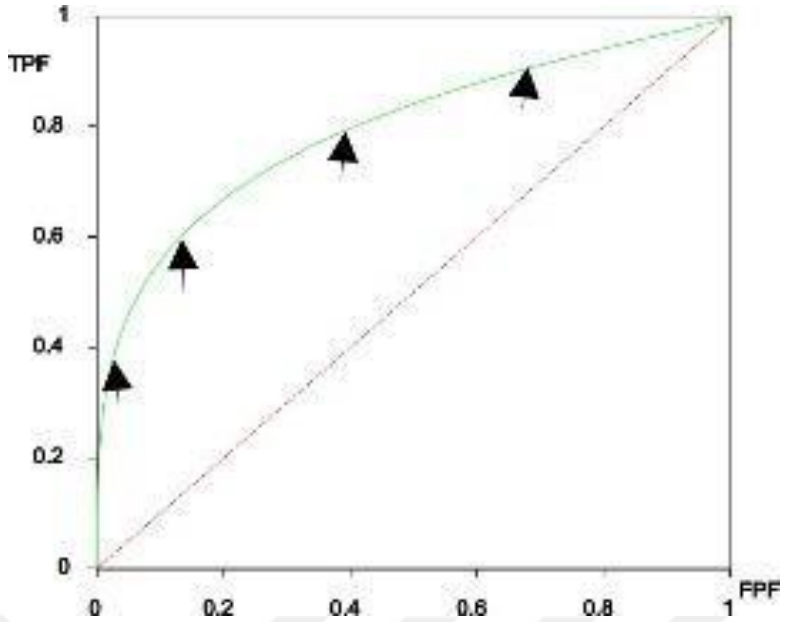

Figure 3.8 ROC Curve [67]

The test performance at all possible threshold values is the total area under the ROC curve. Because the AUC (Area under the ROC Curve) is a part of the area of the unit square, its rate must be between 0 and 1.0. But the diagonal line has an area of 0.5 because chance predicting produces between (0,0) and (1, 1), so no accurate classifier should have an AUC less than 0.5.

3.6 Data Mining Applications

Today, data mining has a wide range of uses from finance to telecommunications. Data mining can be applied in almost any field of application that data warehouses are built. In the process of improving data mining, new and progressively novel applications occur. The implementations of data mining are separated into groups [68].

Information achieved as a conclusion of data mining analysis is used for customer relationship management and customer retention, marketplace research, and analysis, predicting financial, production management and control, and scientific information discovery. The applications in areas shown in Figure 3.9 as follows:

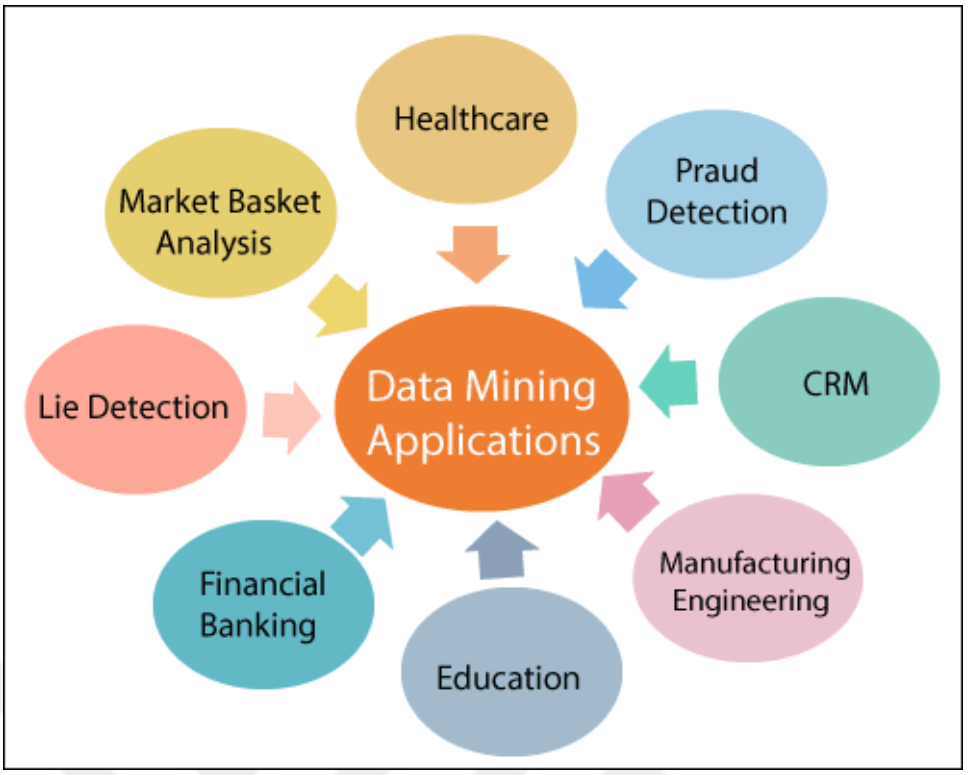

Figure 3.9 Data Mining Application [69]

Marketing: The marketing industry is important for the sales of a product. If marketing can be made accurately, it provides customer sustainability and acquisition of new customers, customer buying habits, to increase market share and profit, and to determine the purpose and campaigns. Many data mining applications are used in the management of customer relationships, such as discovering the most profitable customers.

Retail and Logistics: Data mining is used for different aims there. For example, to estimate the sales of specific points to identify the right amount of inventory, to describe the connections between different products, to organize products in the shop, to improve various promotions, or to examine the consumption levels of different product variants to optimize.

Banking and Insurance: Data mining applications can be used to provide customer groups, fraud and fraud detection, credit requirements, and reimbursement. Also, data mining methods are being used to solve insurance- based problems.

Telecommunication: Such as creating customer profiles, providing appropriate campaigns and offers, and finding lost customers, data mining is being used in the telecommunication sector.

Production: Data mining is used to estimate machine errors with the use of sensory data, to discover anomalies and generalizations in the production management system to optimize the production capacity, and to determine different states to develop the quality of the product.

Medicine: Data mining is used to estimate diseases, predict diseases, to provide early diagnosis of a disease.

Tourism: Data mining is used in such subjects as providing services to identify optimal prices for services of many companies in the tourism sector, to forecast the demand at different locations, to identify the customers with high profitability.

Science and Engineering: In the last years, scientific data has occurred in high quantities in the process of simulating and analyzing systems in a laboratory or computer environment. Data mining provides a very appropriate opportunity to understand the data.

3.7 Data Mining with WEKA

Weka is an open-source data mining program developed as java sources. It was improved by Waikato University in New Zealand. Weka is one of the machine learning algorithms to provide data mining studies. Basically, the algorithms can be applied straight to a dataset. another one is called from Java code. Weka has a lot of tools for lots of function. For example: classification, association rules, regression, and clustering. It is also compatible with emerging new machine learning schemes [70].

WEKA includes data preprocessing tools. In addition to, it was created to apply data mining methods to the database easily and flexibly. WEKA application menu is shown in figure 3.10.

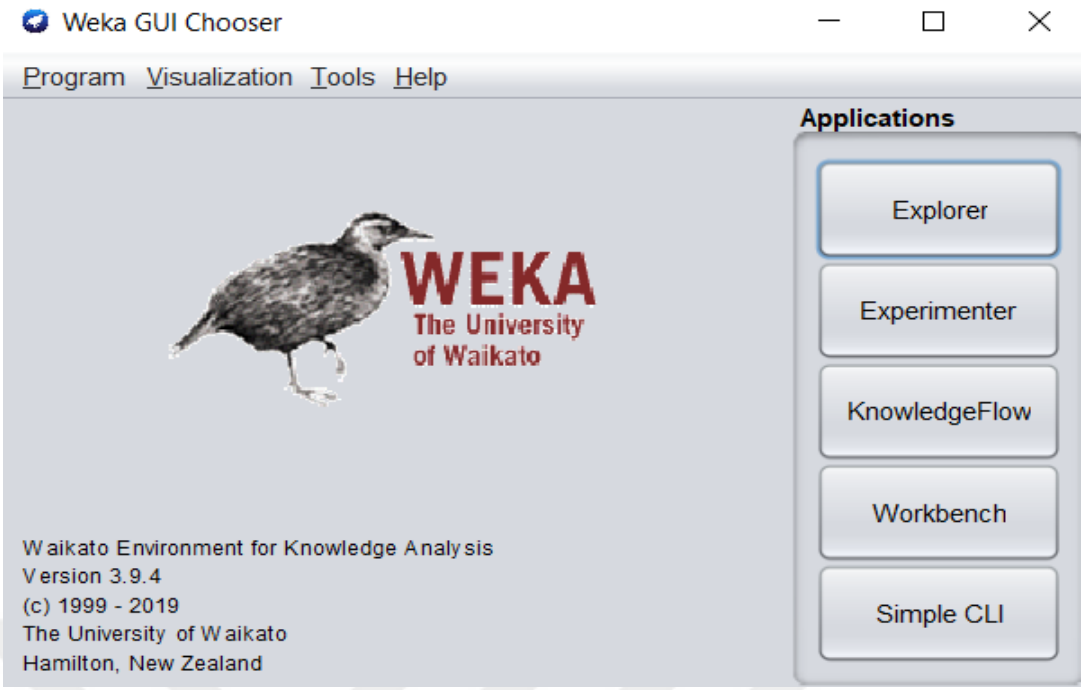

Figure 3.10 A screenshot of WEKA Application Menu

3.7.1 WEKA Explorer

In the WEKA application, is several user interfaces. In the basic, explorer is user interface. Explorer is a panel-based, and includes 6 panels corresponding to the data mining techniques promoted by WEKA. WEKA Explorer Menu is shown in figure 3.11.

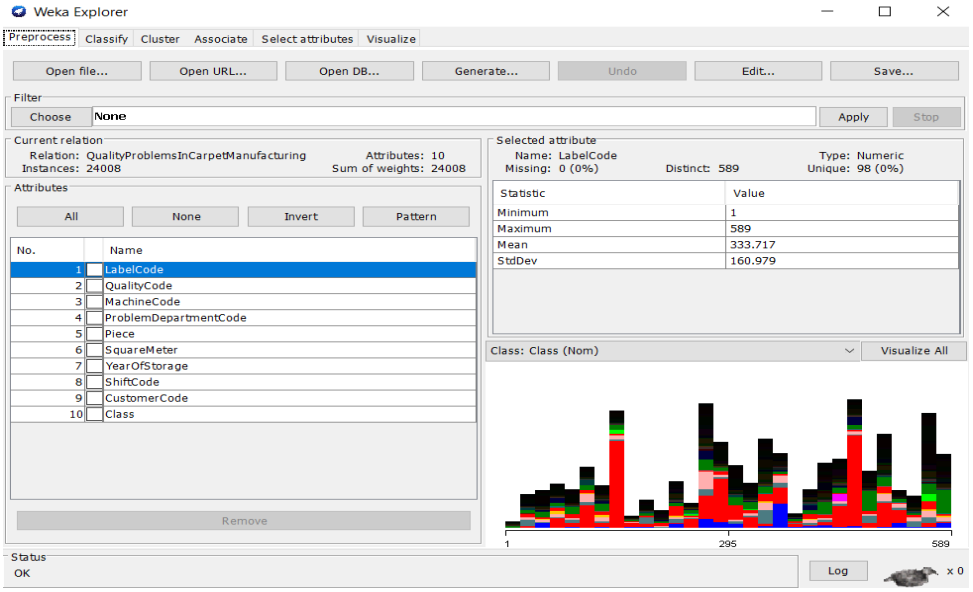

Figure 3.11 A screenshot of WEKA Explorer Menu

Preprocess: In the panel, the data set is chosen and arranged in various ways. There are data editing tools that are presented here as filters. Data can be loaded in 3 ways; database, from file, or URL. Supported data formats contain; CSV, LibSVM.

Classify: The second panel is the panel that contains the classification or regression algorithms. The cause for the paneling classification is that the continuous classes of regression techniques are known as predictors. The panel performs independent cross-validation on the data set organized in the data preparation panel with the chosen learning algorithm to specify the predictive performance. Also, it presents the textual representation of the data set and to provides a graphical representation of the model or decision trees if the conditions for the data are available and visualization of prediction errors in scatter graphs and evaluation of curves such as ROC. At the same time, the model can be permanently registered and reloaded in this panel.

Cluster: It presents the data set loaded into the data preparation panel. The WEKA provides how many clusters it has and the number of instances in each cluster. The panel supplies simple statistics to utilize clustering performance. The model can also be permanently recorded.

Associate: This panel shows learning and evaluating the cohesion rules on data. It is easier to use than the clustering and classification panels.

Select attributes: This panel provides that the data set access to a wide range of algorithms. In this way, it provides to combine different evaluation criteria with different search methods and to construct a wide variety of techniques.

Visualize: This panel provides visualized data set. The point is that the data set itself is visualized in this panel, not the conclusions of the classification or clustering model. Also, it provides a scatter graph of all attribute pairs over a two-dimensional matrix.

3.7.2 WEKA Classification Test Options

The result of classification is tested along with the waisted options by ticking in the test options box. There are four test modes [71]. WEKA Test Options is shown in figure 3.12.

Use training set: In the test option, the classifier is forecasted on how well it foresees the class of the event it is represented on.

Supplied test set: In the test option, the classifier is guessed on how well it estimates the class of a set of examples taken from a file.

Cross-validation: In the test option, the classifier is guessed by cross- validation, using the number of folds that are shown in the Folds text part. To test the success of the applied method, the dataset is splitting into training and test clusters. Once the cross-validation has split the data set by K values, one of the parts will be chosen for testing and used for the rest of the training.

Percentage split: In the test option, the classifier depends on how well it estimates a certain percentage of the data which is holding for testing. Also, the amount of data that is holding depends on the value shown in the % part.

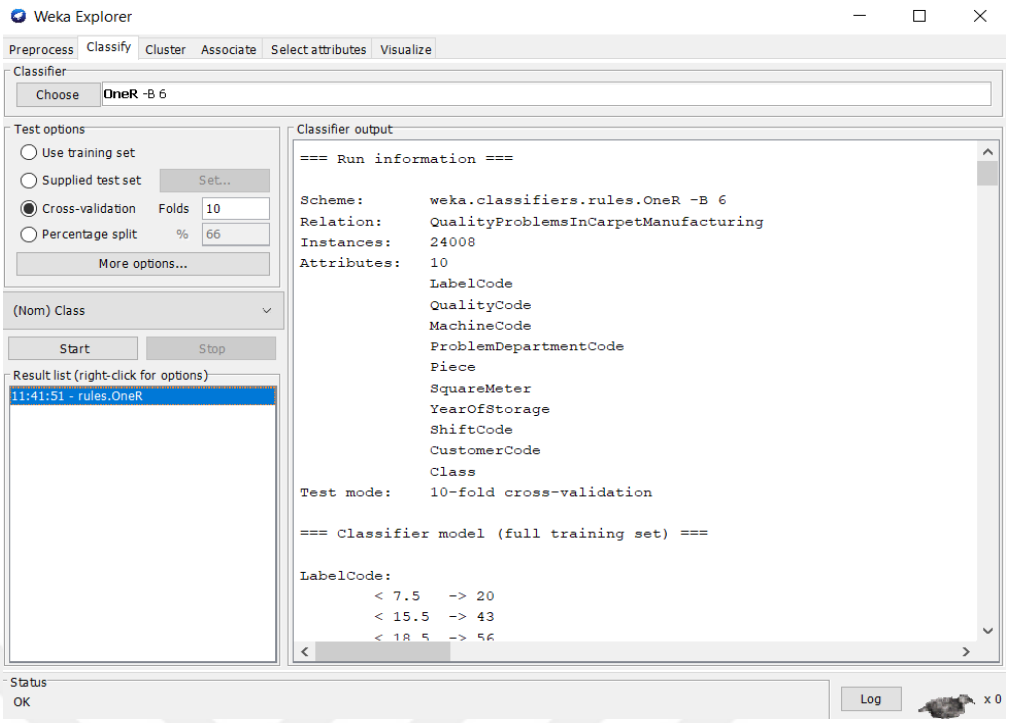

Figure 3.12 A screenshot of Weka Test Options

3.7.3 WEKA Experimenter

In data mining applications, the algorithms show performance on more than one data set. Also, more than one algorithm compares performances on the same data set. WEKA software can promote this kind of work. This provides the designed tasks to be conducted more efficiently, faster, and easier. Some studies can be made with the "Experimenter" option of WEKA software (Figure 3.13). The studies can be recorded to disk for re-use and configured or recorded experiments can be run from the command line. On this screen, there is "Setup" where the algorithms to be worked can be chosen and "Run" where the experiments can be conducted and end of the conducted work. Experiments can only be carried out on classification methods in this panel, but clustering studies are not appropriate.

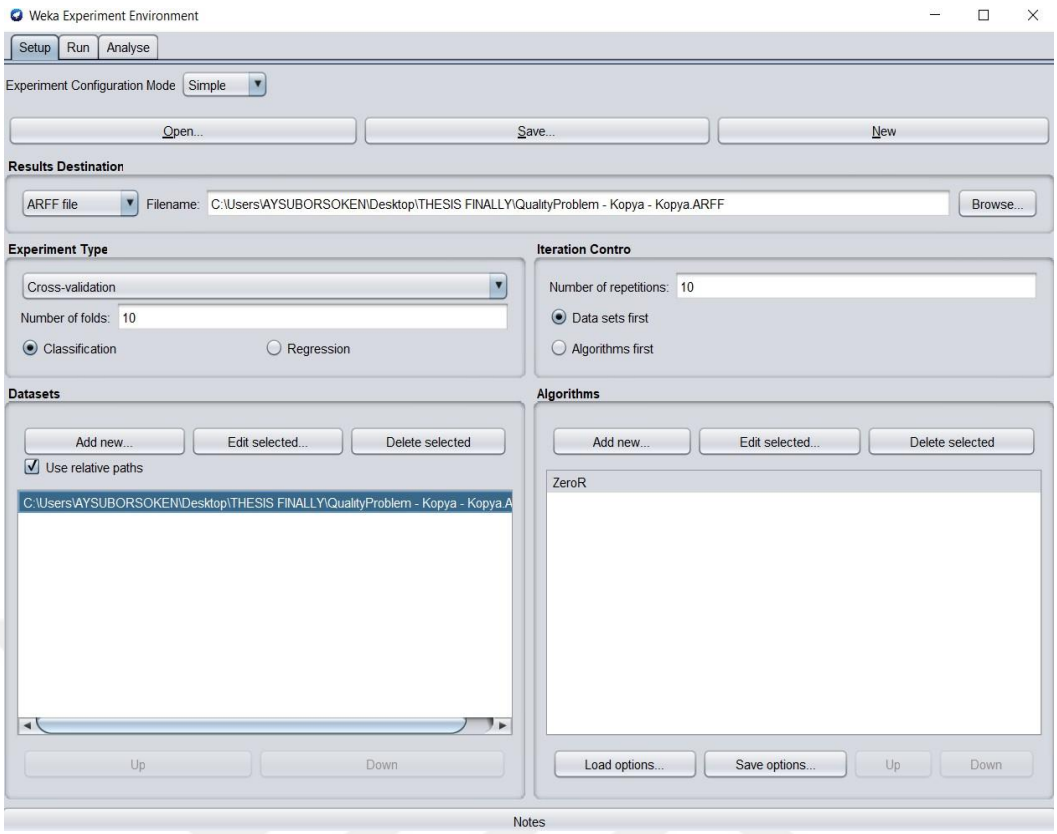

Figure 3.13 A screenshot of Weka Experimenter

3.7.4 WEKA Knowledge Flow

"Knowledge Flow" is a working environment that is improved to design the works that can be performed in the "Explorer" with visual drag-and-drop logic (Figure 3.14). The "Knowledge Flow" interface is an alternative to the Explorer. It specifies data stream and runs algorithms that stream data from one component to another.

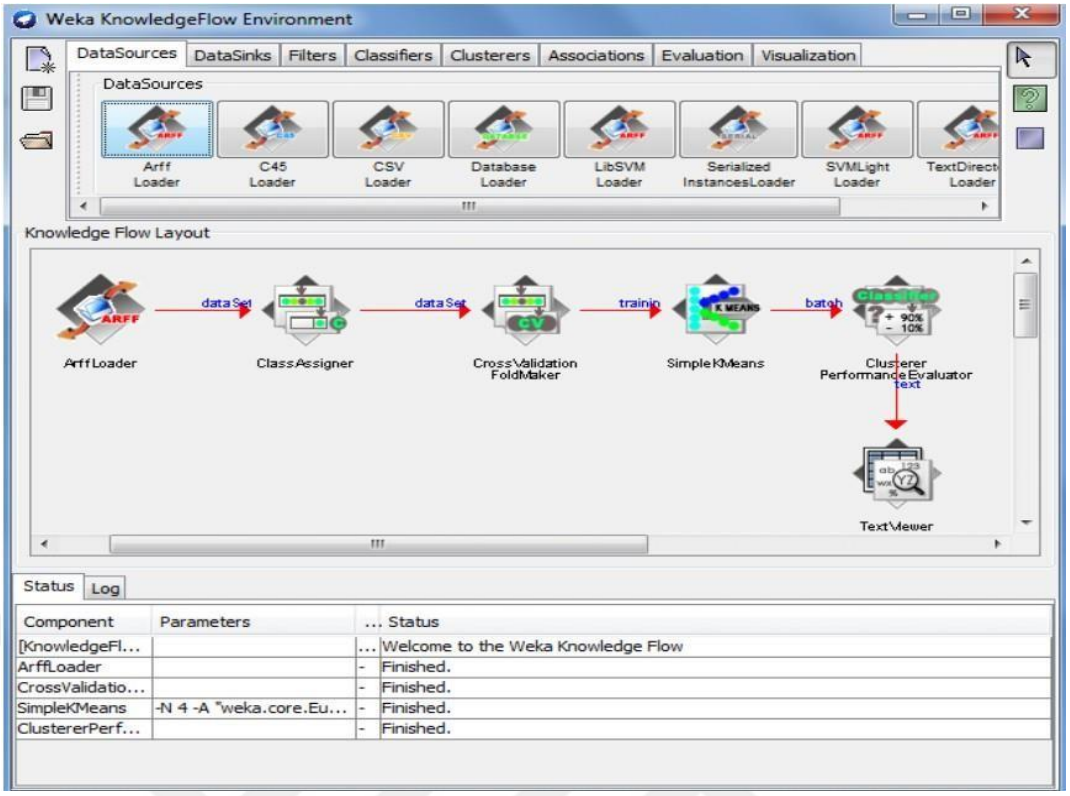

Figure 3.14 A screenshot of Weka Knowledge Flow

CHAPTER IV

EXPERIMENTAL STUDY

In this stage, a study is realized with clustering algorithms in order to group the defects according to certain characteristics by taking the "Defective Products" dataset from the database of the company that produces the carpet. The purpose of the clustering process is maximizing the similarity within the cluster and minimizing the similarity between the clusters. Clustering can be the first step in case of defective product segmentation. " Which types of products increase customer satisfaction or which features ensure quality products? " Instead of dealing with the question, first the defective products are divided into subsets according to their similar characteristic. And then, the question "Which type of product provides the best results for each cluster?" is asked. The characteristics of quality problems in the database in the carpet are examined in detail. After the algorithms are applied to the dataset, we made the comparison, check the algorithms, and found the algorithm that gave the best conclusions.

4.1 Dataset Description

During this part, the characteristics of quality problems in carpet production evaluated by managers were used as the data. In the company, the quality control team finds about 24.008 quality defects, described by 5 attributes. The quality control team discovers quality defects, when the products are in the quality control line.

There are 5 different attributes (4 numeric values and classes) belonging to quality defects in the dataset. These attributes are shown in Table 4.1.

Table 4.1 Dataset Attributes

Attributes	Name
Data of Quality Defects	Quality, Problem department, Customer, Shift, Class

Attributes including verbal data such as customer name, quality name, defect type from the data transferred to the database are digitized in the "1-2-3-..." type. In the study, it is thought that the best clustering results could be given by looking at the area of quality errors and the relationships between these attributes. There are shown digitized data and their coding in Table 4.2, Table 4.3, Table 4.4, Table 4.5.

Table 4.2 Data encoding of the class

Class Code	Class Name	Class Code	Class Name	Class Code	Class Name
1	abrage (weaving)	**23**	curve carpet	**45**	overlock defect
2	abrage (supplier)	**24**	lack of latex	**46**	faulty carving
3	barcode defect	**25**	Cutting defect from width	**47**	pallet defect (design)
4	weight giving finish (berk)	**26**	label defect	**48**	pallet defect (planning)
5	weaving machine blade defect	**27**	extra latex	**49**	pallet defect (confection)
6	camber	**28**	jacquard defect	**50**	eaves defect
7	fluff	**29**	making bouclé from pile	**51**	yellowing
8	lack of cloth	**30**	stain pile	**52**	tear of carpet (confection)
9	cutting defect from carpet's length	**31**	imperfect weaving	**53**	tear of carpet (planning)
10	dyeing	**32**	yarn twist defect	**54**	overcast defect
11	stain vapor	**33**	stain yarn	**55**	wire location
12	bouclé defect	**34**	stain machine	**56**	tolerance size defect
13	Not work bouclé	**35**	faulty cage	**57**	log
14	making pile from bouclé	**36**	thick weft	**58**	weld defect

Table 4.2. (cont'ed) Data encoding of the class

Class Code	Class Name	Class Code	Class Name	Class Code	Class Name
15	chemical finishing agent	**37**	mixed carpet	**59**	Pile length defect
16	cross weaving	**38**	faulty inflexion	**60**	burned carpet
17	cross design	**39**	mildew	**61**	carpet edge sewing
18	cross yarn	**40**	stain sourced from weaving	**62**	tear carpet
19	decadence of yarn	**41**	stain sourced from confection	**63**	tear carpet sourced from weaving
20	warp yarn defect	**42**	stain sourced from supplier	**64**	putting the carpet on the wrong shelf in the warehouse
21	design defect	**43**	difference lot	**65**	razor defect
22	defect from bulk	**44**	razor defect		

Table 4.3 Data encoding of the problem department

Problem Department Code	Problem Department
1	Confection
2	Design
3	Quality Control
4	Import and Export
5	Planning
6	Yarn
7	Weaving

Table 4.4 Data encoding of the quality

Quality Code	Quality Name	Quality Code	Quality Name	Quality Code	Quality Name
1	32x48 Casablanca	81	32x20 1 Dolu, 1 Bos, Loft	161	40x56
2	100x100 Iran	82	32x20 1 Dolu, 1 Bos, Loft	162	40x60
3	100x100 La Meridian	83	32x20 BCF	163	40x60 Fruzan
4	100x100 Les Ottomans	84	32x20 Bilbao	164	40x60 Nevada
5	100x100 Peru	85	32x22	165	40x60 Polip Simli
6	100x100 Valentino	86	32x22 1 Dolu, 1 Bos, 2 cm, Loft	166	40x60 Provence Sonil
7	100x100 Vartian Silk	87	32x22 1 Dolu, 1 Bos, 2 cm	167	40x65
8	100x75 Soho	88	32x22 1 Dolu, 1 Bos, Morocco	168	40x65.
9	100x85 Soho	89	32x22 2 Dolu, 2 Bos	169	40x70 Oymali
10	16x17 1 Dolu, 1 Boş, Queen	90	32x22 Bilbao	170	40x76
11	16x18 1 Dolu, 1 bos, Queen	91	32x23	171	40x80
12	16x18 1 Dolu, 4 Bos, Ultra Shggy	92	32x24	172	40x88
13	16x19 1 Dolu, 1 Bos, Queen	93	32X24 Bilbao	173	40x88 Galata
14	16x19 2 Dolu, 1 Bos, Queen	94	32x24 Modena Sisal	174	48x105
15	16x19 6 cm, Quattro	95	32x25	175	48x105 Bolara
16	16x20 2 Dolu, 2 Bos	96	32x25 4 Dolu, 1 Bos, Loft Shaggy	176	48x105 I. Brilant
17	16x20 6 cm, Quattro	97	32x27	177	48x105 M. Lavinia
18	16x21 1 Kalin, 1 Ince, 95 mm, low	98	32x28	178	48x105 Motion
19	16x22 1 Kalin, 1 Ince, 95 mm, Cornellia	99	32x29	179	48x105 Nessa
20	16x22 1 Kalin 1 Ince	100	32x30	180	48x105 Paintage
21	16x22 100% Dolu	101	32x30 Maya	181	48x105 Savoy
22	16x22 2 Dolu, 1 Bos	102	32x32	182	48x105 Taboo
23	16x22 2 Dolu, 2 Bos, Harmony	103	32X32 1 Dolu, 1 Bos	183	48x105 Taboo Runner
24	16x22 3 Dolu, 1 Bos	104	32x32 Casablanca	184	48x125
25	16x22 5 Bos, 4 Kalin, 1 Ince, 3 cm	105	32x32 Loren	185	48x125 Nessa
26	16x23 1 Kalin, 1 Ince, 95 mm, Cornelia	106	32x32 Oregon	186	48x125 Taboo
27	16x23 1 Kalin 1 Ince, 95 mm, Cornellia	107	32x35	187	48x140 Sehrazat
28	16x24 1 Dolu, 1 Bos, 3 cm	108	32x37	188	48x140 Shiraz
29	16x24 1 Mega, 1 Frisee	109	32x37,5	189	48x150
30	16x24 1 Mega, 2 Frisee	110	32x37,5 Low Bolivia	190	48x150 Otto
31	16x24 1 Kalin, 1 Ince Cornellia	111	32x37.5 Casablanca	191	48x150 Regnum

Table 4.4 (cont'ed) Data encoding of the quality

Quality Code	Quality Name	Quality Code	Quality Name	Quality Code	Quality Name
32	16x24 2 Dolu, 2 Bos	112	32x38	192	48x150 Wyndham
33	16x24 3 cm, Loft Shaggy	113	32x40	193	48x48 Aqua Silk
34	16x24 3 Dolu, 1 Bos	114	32x40 2 cm Tricolour	194	48x48 Barcelona
35	16x24 5 cm, 1/4 bosluklu	115	32x40 Himalaya	195	48x50
36	16x25 1 Mega, 1 Bos, 3 cm	116	32x40 Oregon	196	48x52,5
37	16x25 1 Mega, 1 Frisee	117	32x42 Lorenza	197	48x52,5 Invista
38	16x25 1 Dolu, 2 Bos	118	32x42 Oregon	198	48x57 Invista Polip
39	16x25 2 Dolu, 2 Bos, Dallas	119	32x43	199	48x60
40	16x25 2 Dolu, 2 Bos, Dallas	120	32x44	200	48x60 Versay
41	16x25 3 cm, Shaggy	121	32x45 Tricolour	201	48x62,5
42	16x25 5 Bos, 4 Kalin, 1 Ince, 3 cm	122	32x48	202	48x65
43	16x25 Astoria	123	32x48 Alhambra	203	48x65 Aqua Silk
44	16x27 2 Dolu, 1 Bos, 3 cm	124	32x48 Aura	204	48x65 Aquaslk
45	16x27 1 Dolu, 1 Bos, Loft Shaggy	125	32x48 Casablanca	205	48x65 Bahama
46	16x27 1 Mega, 2 Frisee Lotus	126	32x48 Marakesh	206	48x65 Valencia
47	16x27 2 Dolu, 2 Bos	127	32x48 Starlight	207	48x70
48	16x27 3 Dolu, 1 Bos, 3 cm, Loft	128	32x48 Starlight Saçakli	208	48x75
49	16x27 5 cm, 1 Mega, 1 Frisee	129	32x48 Valeron	209	48x75 Low Artisan
50	16x27 6 Mega, 1 Bos	130	32x48 Ziegler	210	48x75 Rapsody
51	16x28 Shaggy	131	32x48 Ziegler 2.	211	48x75 Rapsody
52	16x29 2 Hav, 1 Loop	132	32x50	212	48x75 Rapsody.
53	16x30 4 cm	133	32x52,5	213	48x80 Aqua Silk
54	16x30 Loft Shagyy	134	32x55	214	48x80 Aqua Silk
55	16x30 Shaggy Loft	135	32x60 Himalaya	215	48x80 Aqua Silk
56	16x32 1 Mega, 1 Bos	136	32x62,5	216	48x80 Artisan
57	16x32 2 Mega, 1 Bos	137	32x68	217	48x80 Invista
58	16x32 3 cm, loft shaggy	138	32x75	218	48x80 Kenitra
59	16x32 5 cm, Shaggy	139	32x75 Loop	219	48x80 Marina
60	16x32 6 cm, New Ethos	140	32x88	220	48x80 Rustyart
61	16x35 2 dolu, 1 bos	141	40x100	221	48x85 Dorian
62	16x35 2 hav, 1 Loop	142	40x100 Kunduz	222	48x88 Dorian
63	16x35 2 hav, 1 Loop, 6 cm	143	40x100 Timeless	223	48x90 Artisan
64	16x36 3 cm, Loft Shaggy	144	40x120	224	48x90 Artisan
65	16x36 5 cm, New Ethos	145	40x35	225	48x90 Artisan
66	16x38 Shaggy	146	40x40	226	48x90 Dejavu

Table 4.4 (cont'ed) Data encoding of the quality

Quality Code	Quality Name	Quality Code	Quality Name	Quality Code	Quality Name
67	16x40 3 Dolu, 1 Bos, Ethos	147	40x42	227	48x90 Dorian
68	16x40 Loft Shaggy	148	40x44	228	48x90 Melange
69	16x42 6 cm, Blanca	149	40x45	229	48x90 Ventus
70	16x44 Odessa	150	40x48	230	48x96 Bolara
71	16x45 Oymali Shaggy Loft	151	40x50	231	48x96 Esperanza
72	16x50 Shaggy	152	40x50 Sofia	232	48x96 Quasar
73	24x35 Low Lumina	153	40x50 Sofia Hb	233	50x100
74	24x48 Lumina	154	40x50 Sofia	234	50x110 Belmond
75	24x50 Qushak	155	40x52,5	235	50x130
76	24x55	156	40x53	236	50x130 Melrose
77	24x65 Bijar	157	40x55	237	48x80 Aqua Silk
78	24x65 Lumina	158	40x55 Escape	238	Motion
79	24x65 Lumina	159	40x55 Escape	239	Starlight
80	32x18	160	40x55 Escape	240	Zermatt Shaggy

Table 4.5 Data encoding of the customer

Cus. Code	Customer Name	Cus. Code	Customer Name	Cus. Code	Customer name
1	3Tex & Mehmet Sırlı	66	Ecarpet Gallery	131	NDS Kaukas
2	ABC Italya	67	Efendioglu	132	New Plan
3	Abdullah Saleh Al Avani	68	Eko Halı	133	NK Sales
4	Acem Tekstil	69	El Ketani	134	Nourison
5	Al Degheishem	70	El Rawan	135	Nuka
6	Al Gobbi	71	Empera	136	Numune
7	Al Kaffary	72	Empire Rugs	137	Obsessions
8	Al Nafea	73	Esaimar	138	Örnek Hali
9	Al Salem	74	Essam Ali Elgadi	139	Persian Dreams
10	Al Sara for Import	75	Eurofil	140	Pezzoli
11	Al- Lamushi Carpet	76	Euro-Tapis	141	P-Floor
12	Al Morky	77	Fabrics For Less	142	Polytex
13	Alpha Rug	78	Family Carpet	143	Prospero
14	Amer Rugs	79	Farzin S.A	144	PT Forum
15	American Cover Design	80	Fatex	145	Royal Carpet
16	Ansar Gallery Bahrain	81	Festival Halı	146	Rug Expo Inc.
17	Ansarmall	82	Floko	147	Rugs America
18	Apex Halı	83	Flora Home	148	Rugs Usa
19	Arctex	84	Geldof	149	Rupesh Kumar
20	Area Rug Shop	85	Gertmenian	150	Ruslan
21	Ariad Carpet	86	Giambra	151	Safavieh
22	Art De Vivre	87	Gjini Carpet	152	Safir Halı
23	Artur Magomedov	88	Habitat	153	Said
24	Asiatic	89	Heritage	154	Salih Mahmood
25	Asos Carpet	90	Heriz Gallery	155	S. Naggy Trading
26	Atlantik Halı	91	Hıdır	156	Sams Inter.
27	Aydın Carpet	92	Home Dynamix	157	Saray Rugs
28	Ayyıldız	93	Home Rus	158	Saydam Halı
29	Azeri Aloset	94	Ilyas	159	S. Vasini Rugs
30	Azeri Sefer	95	Incarpet	160	Silverstand

Table 4.5 (cont'ed) Data encoding of the customer

Cus. Code	Customer Name	Cus. Code	Customer Name	Cus. Code	Customer name
31	B. Home Value	96	Indomex	161	Sofa House
32	Babak	97	Issa Manasreh	162	Sprintzios S.A
33	Baran Hali	98	Istanwool	163	SSM
34	Baruch	99	Jagdish Store	164	Stevens Omni
35	Başaran Halı	100	Jaipur	165	Stok
36	Batyrzhan	101	Jolitex	166	Sunshine
37	Becquet	102	JY Safavieh	167	Surya Carpet
38	Beijing Yingtai	103	Karen	168	Tamir
39	Benuta	104	Kas	169	Tapis Dbouk
40	Biokarpet	105	Kenane	170	Tapis Kabalan
41	Brand Ventures	106	Khurasan	171	Tayse Rugs
42	Brandao	107	KRC Carpet	172	Tegap Inc.
43	Bursa Furniture	108	Libyalı Osama	173	Test Rite
44	Capitola Rugs	109	Loloi	174	Tianjin Huasaier
45	Carillo	110	Maddah Home	175	TMK Rugs
46	Carpet Centre	111	Madi Carpet	176	Toda Carpet
47	Carpet Prima	112	Mafour Co.	177	Toshi
48	Carpet Vista	113	Magomed	178	Trans Jordan
49	CB Rugs	114	Makro Carpet	179	Tufan
50	Cengel Sanayi	115	Mastercruft	180	Turhan- Olimpik
51	Cihan Eralpler	116	Mattex	181	Turkmen Halı
52	Classis Bvba	117	Mazan Carpet	182	Ugurser
53	Covor	118	Melrose Textile	183	Ukrayna Klast
54	Covtex	119	Meshan	184	Ukrayna Viktor
55	Cyrus	120	Mir Halı	185	Unique USA
56	Dar Asheraz	121	Miraç Turkmen	186	United Agencies
57	Decorit	122	Miraç Suudi	187	Unitex
58	Dekor Invest	123	M. Ali Baidoun	188	Ustunsoy Halı
59	Dib Alfombras	124	M. Al Bari	189	Vestis Kft
60	Diego	125	Momeni	190	Vietnam Delta
61	Dmitriy Babenko	126	Morelli	191	V. De Alfomnras
62	Doha City	127	Muhammed Dus	192	War. Carpets ltd.
63	Dolce Vita	128	Murtuz	193	Yahya Carpet
64	Domain Carpets& More	129	Museum	194	Yasin Kaplan Hali
65	Domotex Fuar	130	Natco	195	Yura

4.2 Source of Defects by the Departments

The problem department is the department to causes a quality defect. The departments are Confection Dept., Design Dept., Quality control Dept., Import and Export Dept., Planning Dept., Yarn Dept., Weaving Dept. Some examples of quality defects originating from departments is shown in Table 4.6.

Table 4.6 Examples of quality defects originating from departments

Problem Department	Examples of Quality Defect
Confection Department	Overlock defect, burned carpet, carpet edge sewing defect
Design Department.	Design defect, cross design, and pallet defect
Quality Control Department	Stain vapor, extra latex, and chemical finishing agent
Import and Export Department	Tolerance size defect, label defect, and barcode defect
Planning Department	Cross weaving, making pile from bouclé, thick weft
Yarn Department	Abrage, warp yarn defect, difference lot, yarn twist defect
Weaving department	Weaving machine blade defect, curve carpet, jacquard defect, imperfect weaving

4.3 Preparing the Data for Application

The dataset taken from the Excel file is preprocessed and transferred to the WEKA program. In the pre-processing stage, the quality defects data are made suitable for the algorithms. The file has been prepared the arff extension required by the program. The arff file is available in Appendix B.

Once the attributes are described, data is integrated sequentially after the @data command.

4.4. Implementation of Methods

In this section of the study, some algorithms are implemented to the data set. The features consisting of 5 attributes obtained in the dataset and the information of whether a sample carpet is defective or not are recorded on the disk. This data file is stored with the "Open file" command from the "Explorer" screen using WEKA software. The screenshot that occurs after the data is loaded is shown in Figure 4.1.

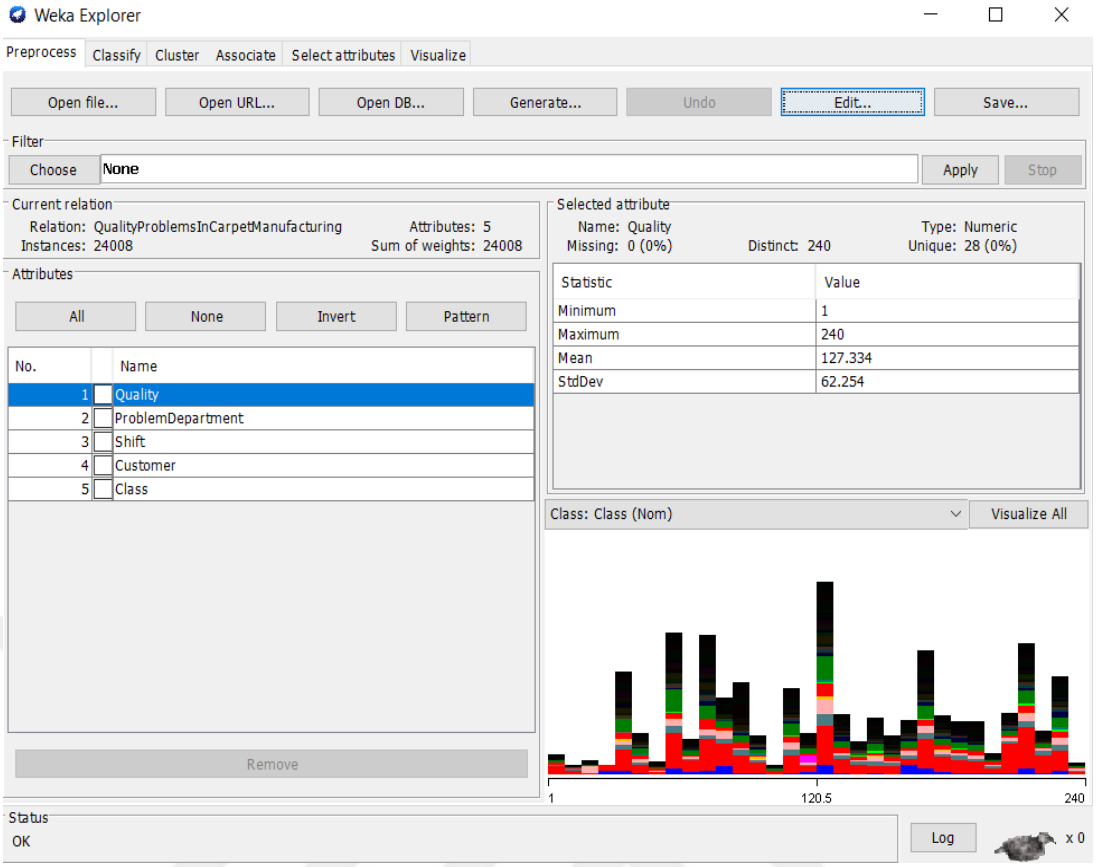

Figure 4.1 A screenshoot of interface created as a result of loading data

Under the “Current relation” heading on this screen the name of the loaded data, how many instances and attributes appears to have.

4.4.1 COBWEB Algorithm

COBWEB is an incremental system for hierarchical conceptual clustering. It incrementally organizes observations into a classification tree. Each node in a classification tree represents a class (concept) and is labeled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. This classification tree can be used to predict missing attributes or the class of a new object [76].

The Quality Problem Data set contains five attributes. All the attributes are numerical. All the attributes are considered as input attributes. The Quality Problem Data set provides the details about the features of carpets and which could be useful to cluster the carpets according to its features. Here the clustering should be done based on the category utility function which is a probabilistic measure of each data value.

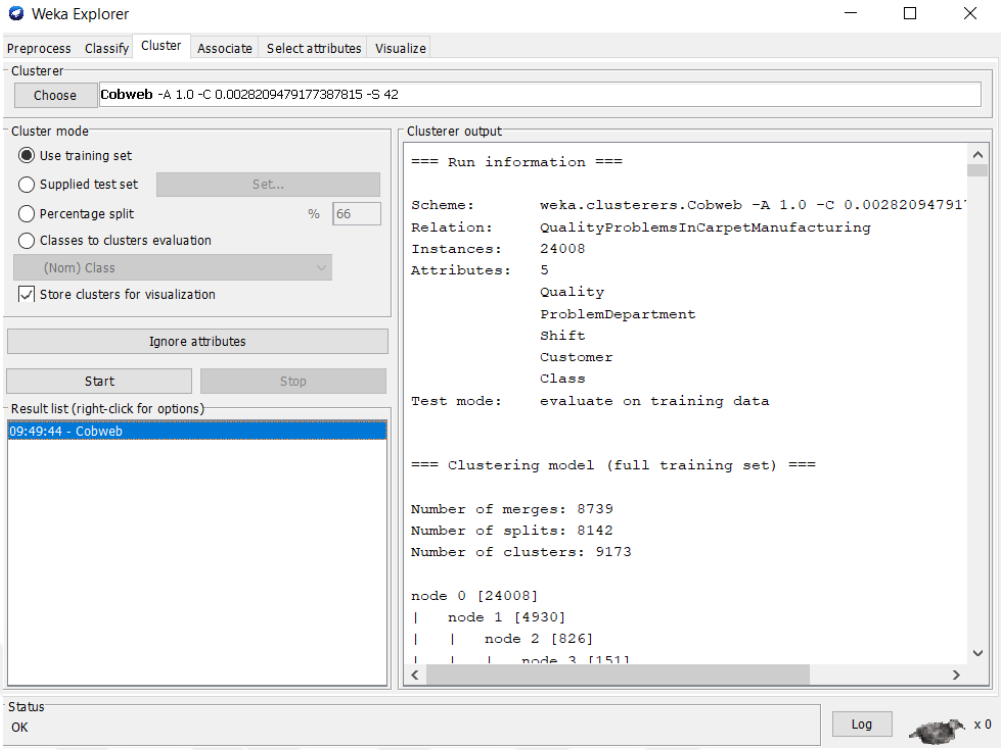

Figure 4.2 Visualize COBWEB Algorithm

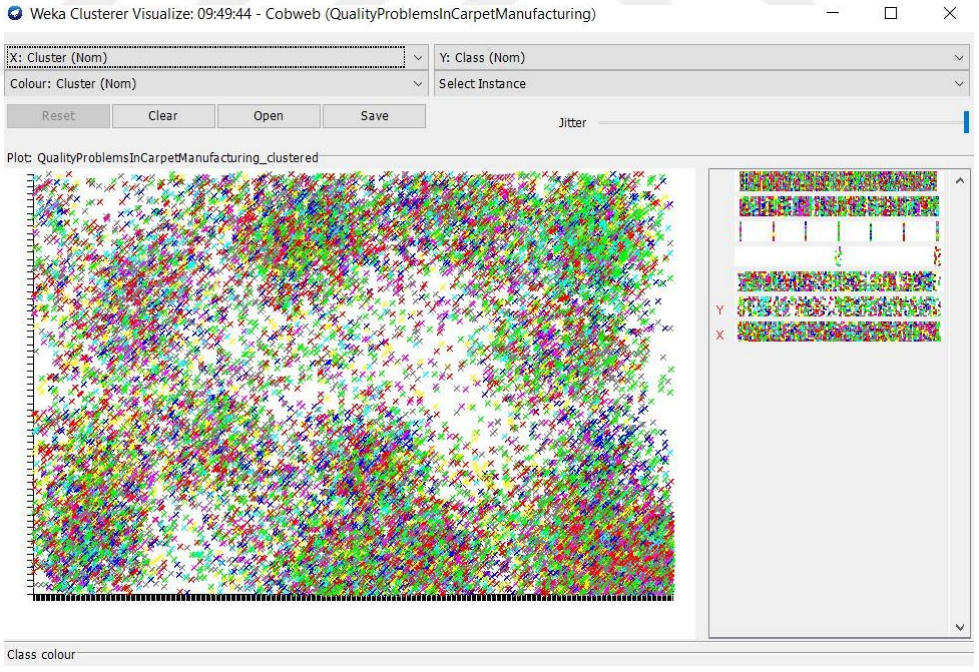

Figure 4.3 Result of COBWEB in form of graph

The figure 4.3 shows the visualization of cobweb algorithm through WEKA tool. And generating results in term of time, no of merges, no of splits and number of clusters.

Final clusters are produced for the datasets. Quality problem dataset produce 9173 clusters using five attributes. For instances, according to this result, Cluster 7823 for the 48th quality defect (Pallet defect, planning) class and this quality defect; It is seen that the second shift, customer number 185 (Unique), department number 5 (Planning) and number 107 (32X35) are included in the cluster as quality.

4.4.2 EM Algorithm

EM algorithm [79] is also an important algorithm of data mining. We used this algorithm when we are satisfied the result of k-means methods. an expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum some posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables.

The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc.

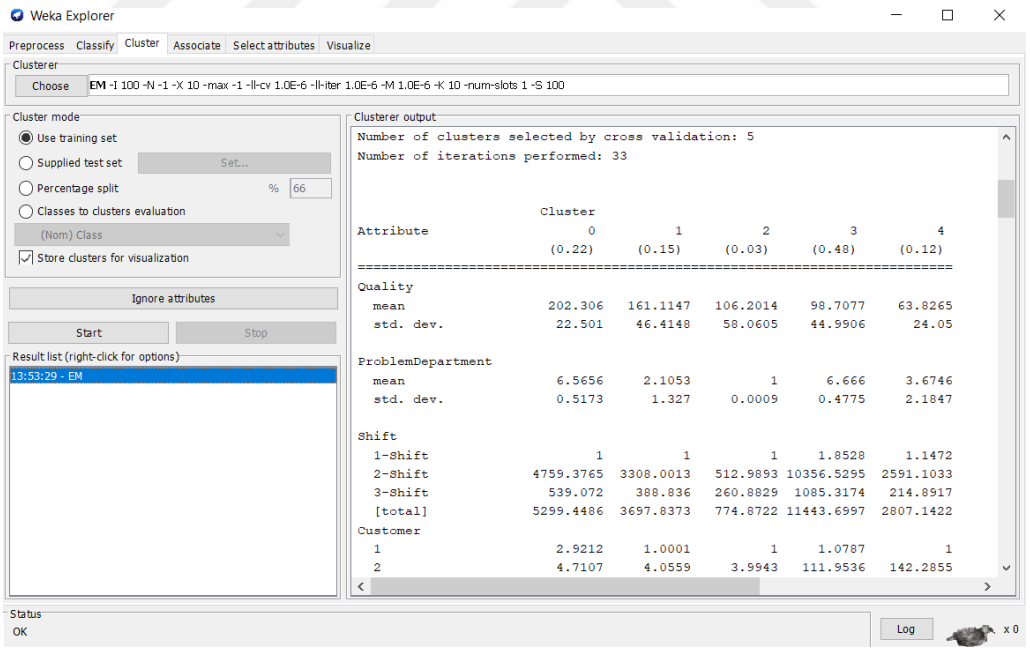

Figure 4.4 EM algorithm

This figure 4.4 shows that the result of EM algorithm. Next figure shows the result of EM algorithm in form of graph. We have seen the various clusters in the different-different colors.

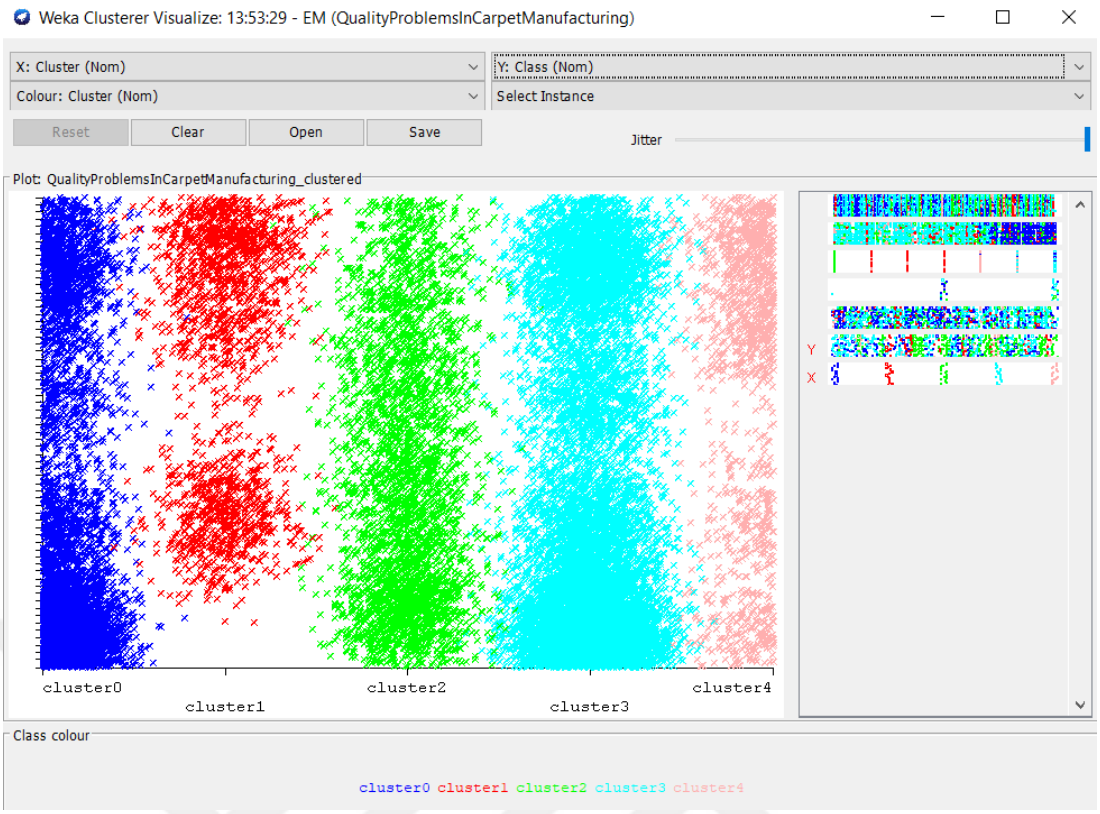

Figure 4.5 Result of EM Algorithm

Quality problem dataset produce five clusters using five attributes. For instances, according to this result, Cluster 3 for the 36th quality defect (thick weft) class and this quality defect; It is seen that the second shift, customer number 165 (Stok), department number 6 (Yarn) and number 58 (16x32 3cm Loft Shaggy) are included in the cluster as quality.

4.4.3 Farthest First Algorithm

Farthest first are a Variant of K means that places each cluster center in turn at the point furthest from the existing cluster centers. This point must lie within the data area. This greatly sped up the clustering in most cases since less reassignment and adjustment is needed.

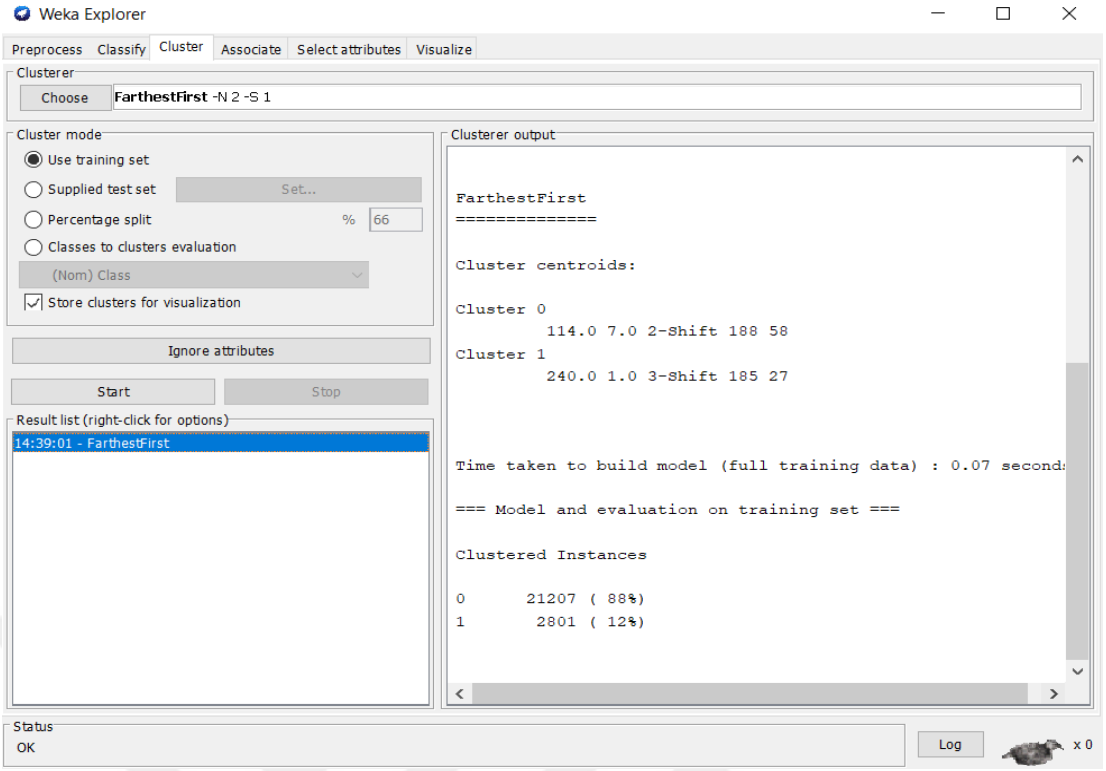

Figure 4.6 Farthest First Algorithm

Analysis of data with farthest first algorithms is shown in the figure. It is dividing the whole data set in two clusters. Each cluster had shown the lowest and higher value of the data sets.

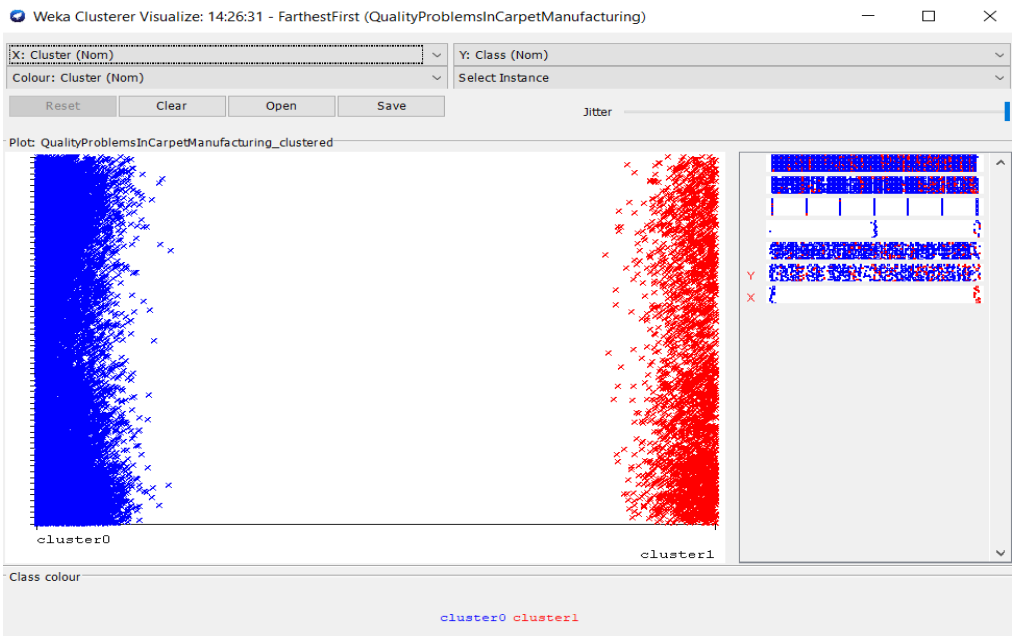

Figure 4.7 Result of Farther First Algorithm

Quality problem dataset produce two clusters using five attributes. According to this result, Cluster 0 for the 58th quality defect (weld defect) class and this quality defect; It is seen that the second shift, customer number 188 (Ustunsoy), department number 7 (Yarn) and number 114 (32X40 2 cm Tricolour) are included in the cluster as quality. Also, Cluster 1 for the 27th quality defect (extra latex) class and this quality defect; It is seen that the third shift, customer number 185 (Unique), department number 1 (Confection) and number 240 (Zermatt Shaggy) are included in the cluster as quality.

4.4.4 K-Means Clustering Algorithm

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.

The clustering algorithm is to be applied to the data by selecting the "Cluster" option in the Weka program; “Simple K-Means” is selected for the k-means clustering algorithm. The screenshot that occurs after this selection is as in Figure 4.8. In this study, no feature selection is made and all the data given for training are also used for testing. The results can be seen under the heading “Clusterer output” in the figure below.

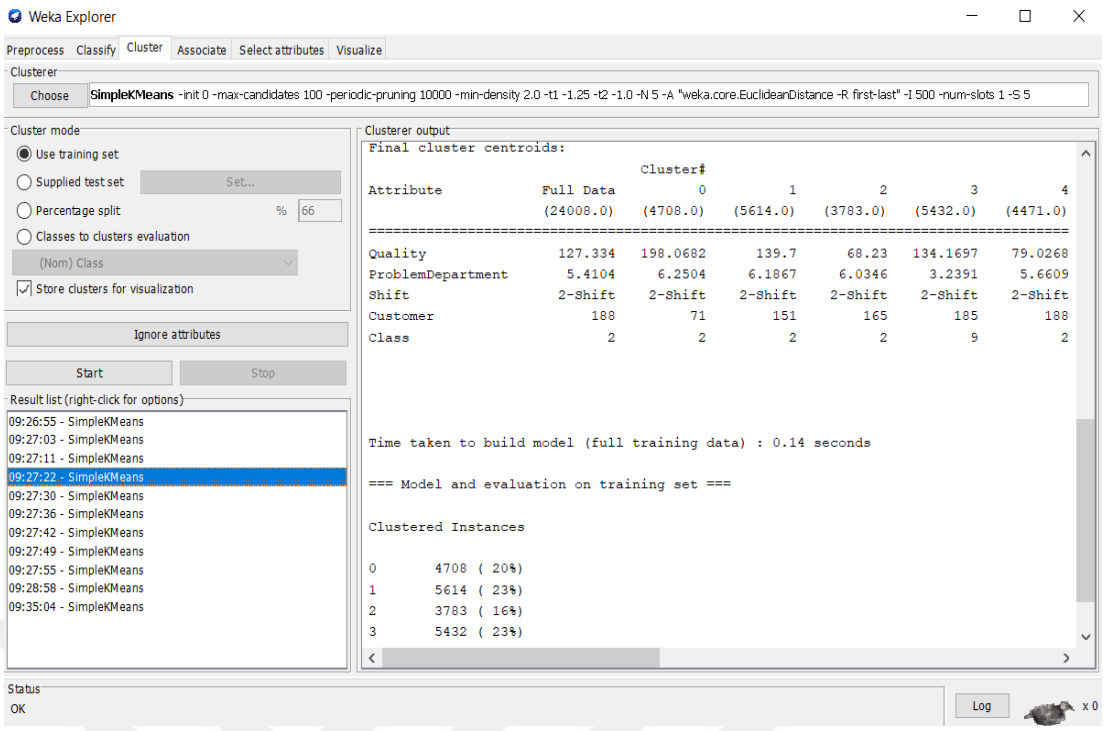

Figure 4.8 A screenshoot of Clustering with the K-means clustering algorithm

In the study, the k value, which determines the number of clusters, is taken from 2 to 10 and the above processes were repeated for different initial values (seed). It is found that as the number of clusters increased, the sum of the squares of the error decreased. In order to observe the effect of the initial value, seed values from 1 to 50 are tried for each k value. Thus, the study is repeated 450 times in total.

The results for the k=10 value are shown in Figure 4.9. According to this result, Cluster 0 for the 2nd quality defect (Abrage Supplier) class and this quality defect; It is seen that the second shift, customer number 151 (Safavieh), department number 6 (yarn) and number 130 (32X48 Ziegler) are included in the cluster as quality. According to another cluster result (for example, Cluster 4), third shift for quality defect class number 57 (log), customer number 185 (Unique), department number 4 (import and export) and quality number 121 (32X45 Tricolour) It is seen that they are included in the fourth cluster.

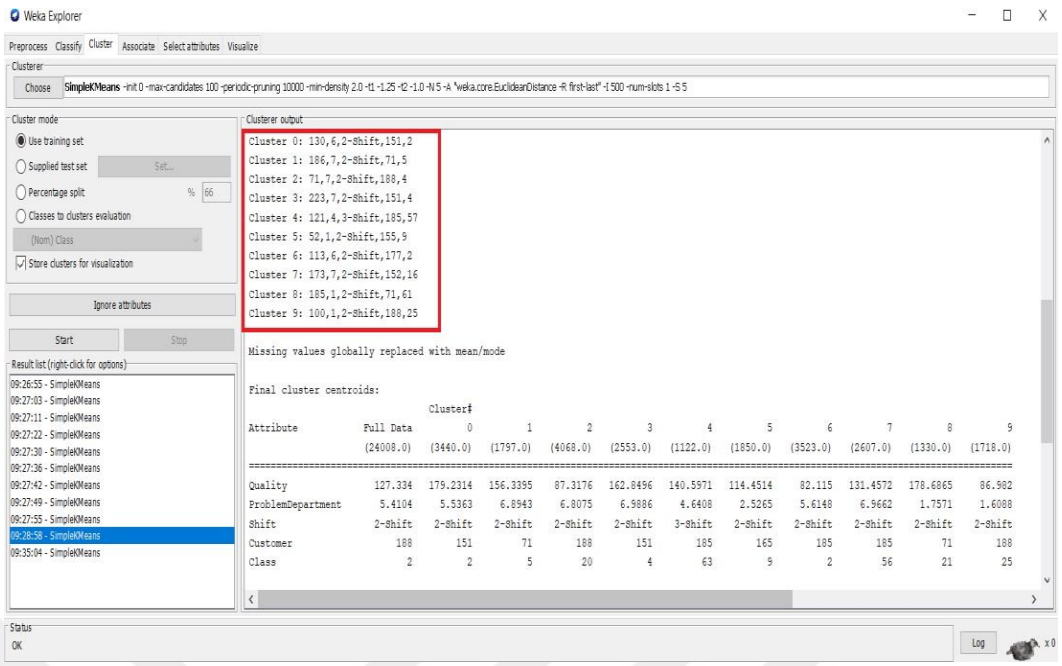

Figure 4.9 A screenshoot of result of 10 clusters

The results are obtained visually for the k=10 value, where the highest performance is obtained, with the WEKA software "Visualise" section in Figure 4.10. According to this figure, Cluster9 and Cluster2 -Cluster3 clusters are assigned with number 4 (weight giving finish) defect for quality defect class number 25 (cutting defect from width). In addition, it is seen that the product with a quality defect that occurred only in the third shift is included in the Cluster4 cluster.

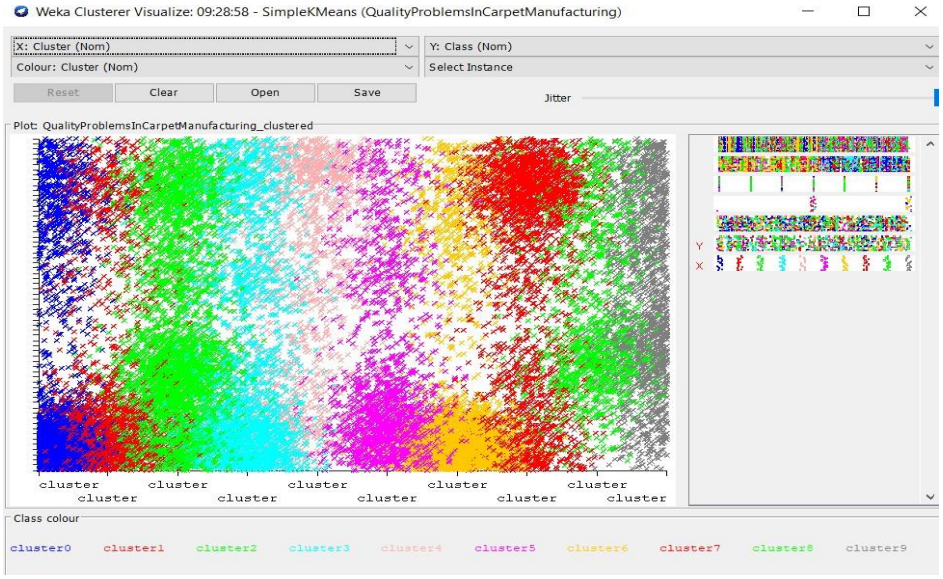

Figure 4.10 A screenshoot of visualization screen of 10 clusters

When classes are evaluated into clusters, the algorithm assigns the class label to the cluster. As see in figure 4.11, Cluster0 represents defect 9 (cutting defect from carpet's lentgh), and Cluster 3 represents defect 20 (warp yarn defect).

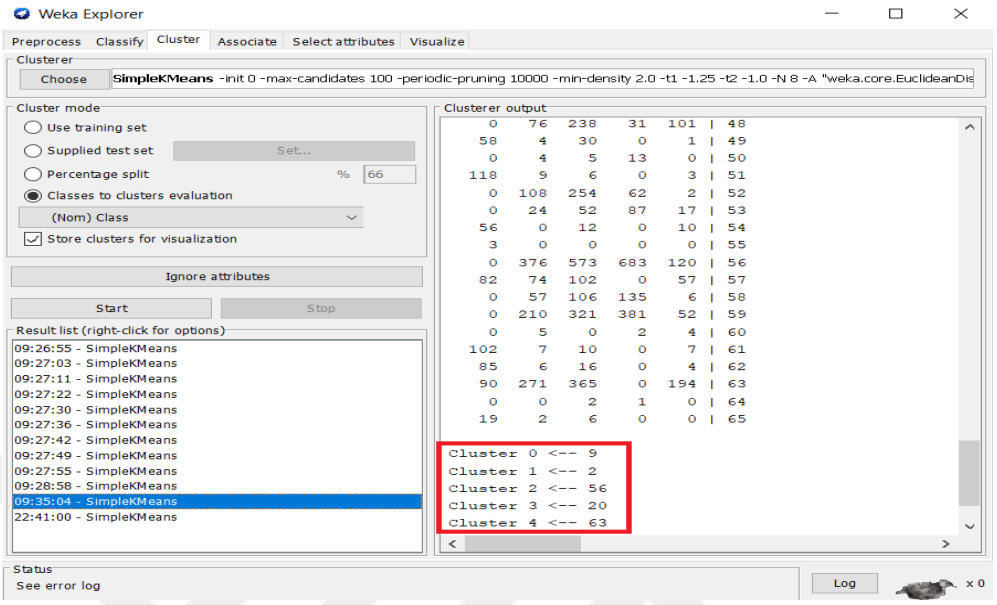

Figure 4.11 A screenshoot of evaluate Classes into Clusters

Table 4.7 Value of error terms for different number of clusters and starting values

Cluster Number (k)	Value of error terms (%)	Seed
2	43.3	3
3	38.8	25
4	36.4	25
5	35.5	5
6	33.1	5
7	32.2	3
8	30.30	5
9	30.35	15
10	30.2	15

It is summarized in Table 4.7. According to these results, the minimum success rate for k=10 in the diagnosis of quality problem in carpet manufacturing was 30.2%. Thus, 7.202 instances out of 24008 quality defects could be defined correctly.

4.3.4 Result Analysis

The following table represents the analysis process of all four algorithms:

Table 4.8 Comparison results of different clustering algorithms

Algorithm	No. of Clusters	No. of Iteration	Time Taken to build
COBWEB	9173	-	102.69s
EM	4	3	65.35s
Farthest First	2	-	0.04s
K-means	10	19	0.27s

We have performed analysis with four clustering algorithms COBWEB, EM, Farthest First, and k-mean. In all four-algorithm result is generated on the basis of similar objects to create that clusters. Best algorithm found is Farthest First clustering. It is taking less time than other clustering algorithm to find similar clusters through WEKA tool for quality problem in carpet manufacturing dataset. The datasets are applied to the WEKA (3.9.4). As the size of datasets increases time taken to form clusters increases. In partitioning based clustering algorithms Farthest First clustering algorithm took least time in forming clusters whereas Cobweb took maximum time. In non-partitioning based Cobweb clustering algorithm formed maximum clusters.

CHAPTER V

DISCUSSION AND CONCLUSION

In today's competitive market conditions, the rising volume of data, process of performing specific operations on a set of data, and ways of systematically collecting data is getting more challenging day by day. Data mining has been an important research area to find the hidden information/knowledge inside huge amounts of data. Although the data mining studies in manufacturing have been very limited, several important benefits are expected with its implementation. Gaziantep has been an industrialized city with a special focus on carpet manufacturing. While hundreds of carpet manufacturing factories have been established to produce carpets, there are several opportunities for production research. The leading problem of the industry has been the quality of carpets. In carpet manufacturing, several thousands of quality problems are recorded manually; however, they are not usually analyzed in depth. In this thesis, quality problems in carpet manufacturing are gathered in a database and the data is analyzed and clustered by using data mining algorithms. The main purpose is to categorize the similar classes in the same clusters.

Data mining and applications of it have increasing popularity in today's world. Clustering, which is one of the analysis of data mining, is used in many areas ranging from banking sector to telecommunications, health field applications to the manufacturing sector.

In this study, the quality problems in a carpet production factory are clustered. For this purpose, 24.008 data between 2017 and 2019 are examined, described by 5 attributes. To obtain the results, the problems information entered into the database by the operators are used. Software is used for easier understanding data of quality defect. One of these software, WEKA software, is introduced and used in the clustering of quality defect data occurring in carpet manufacturing.

A comparative study of clustering algorithms across quality problem datasets has been performed. The performance of various clustering algorithms is compared based on size of dataset, time taken to form clusters and the number of clusters formed. Farthest First algorithm took least time to form clusters and Cobweb algorithm formed maximum number of clusters.

According to the results, when using the k-means algorithm, since it is known that the initial center values affect the algorithm performance, the algorithm is repeated for different initial values. Other studies that are carried out using WEKA software, different starting values should be tried.

The attribute selection is not made in the study. It is predicted that more reliable results can be obtained by adding other algorithms with attribute selection and applying other algorithms that allow the selection of the test set from outside the training set.

As a result, the research can be used to describe the further steps for quality related problems in carpet manufacturing. Also, Recommendations will be presented to the operation managers concerning about the sufficiency of quality defect data, and the improvement of the collected data.

This study is prepared using only the data of the carpet company in Gaziantep province. This data is a limitation of this thesis because all companies may not work on the same qualities. For this reason, generalizing the results of this study to all companies may not be right but this study can be an instance for them.

REFERENCES

[1] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31(8)**, 651- 666.

[2] Lee, M. H. (1993). The Knowledge based factory, *Artificial Intelligence in Engineering*, 109- 125.

[3] Irani, K. B., Cheng, J., Fayyad, U. M. (1993). Applying machine learning to semiconductor manufacturing, Proc. SPIE 1293, *Applications of Artificial Intelligence VIII*, 41–47.

[4] Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning.* 1, **1**, 81–106.

[5] Gardner, M., Bieker, J. (2000), Data Mining Solves Tough Semi-Conductor Problems, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 376- 383.

[6] Chien, C. F., Wang, W. C., Chang, J. C. (2007). *Data Mining for yield enhancement in semiconductor manufacturing and an empirical study*, Expert System with Applications, **33**, 192- 198.

[7] Sebzalli, Y. M., Wang, X. Z. (2001). *Knowledge Discovery from Process Operational Data using PCA and Fuzzy Clustering*, Engineering Applications of Artificial Intelligence, **14**, 607-616.

[8] Liao, T. W., Li, D. M., Li, Y. M. (1999), *Detection of Welding Flaws from Radiographic Images with Fuzzy Clustering Methods*, Fuzy Sets and Systems, **108**, 145-158.

[9] Huang, H. Wu, D. (2005). Product quality improvement analysis using data mining: a case study in ultra-precision manufacturing industry, *in: L. Wang, Y. Jin (Eds.)*, FSKD 2005, LNAI, 3614, Springer-Verlag, Berlin, 577–580.

[10] Harding, J. A., Shahbaz, M., Srinivas, S., Kusiak, A. (2006). Data mining in manufacturing: a review, *Journal of Manufacturing Science and Engineering*, **128**, 969–976.

[11] Wang, K. (2006). Data mining in manufacturing: the nature and implications, in: K. Wang, G. Kovacs, M. Wozny, (Eds.), *International Federation for Information Processing (IFIP) Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management,* **207**, Springer-Verlag, Boston, 1–10.

[12] Rokach, L., Maimon O. (2006). Data mining for improving the quality of manufacturing: a feature set decomposition approach, *Journal of Intelligent Manufacturing*, **17(3)**, 285–299.

[13] Kusiak, A. (2000). *Computational Intelligence in Design and Manufacturing*, John Wiley, New York.

[14] Lee, J. H., Yu, S. J., Park, S. C. (2001). Design of Intelligent Sampling Methodology based on Data Mining, *IEEE Transactions on Robotics and Automation*, **17**, 637-648.

[15] Crespo, F., Webere, R. (2005), *A methodology for dynamic data mining based on fuzzy clustering*, Fuzzy Sets and Systems, **150**, 267-284.

[16] Hui, S. C., Jha, G. (2000). Data Mining for Customer Service Support, *Information and Management*, **38**, 1-13.

[17] Qian, Z., Jiang, W., Tsui, K. L. (2006). Churn detection via customer profile modelling, *International Journal of Production Research*, **44(4)**, 2913- 2933.

[18] Dudas, C., Ng, A. H. C., Pehrsson, L., Boström, H. (2013). Integration of data mining and multi-objective optimisation for decision support in production systems development. *International Journal of Computer Integrated Manufacturing*, **27(9)**, 824- 839.

[19] Neis, J. (2015). *Analyse der Produktportfoliokomplexität unter Anwendung von Verfahren des Data Mining*. Herzogenrath: Shaker.

[20] Koonce, D. A., Tsai, S. C. (2000). Using data mining to find patterns in genetic algorithm solutions to a job ship schedule. *Computers and Industrial Engineering*, **38**, 361- 374.

[21] Turker, A. K., Golec, A., Aktepe, A., Ersoz, S., Ipek, M., Cagil, G. (2020). A real-time system design using data mining for estimation of delayed orders an application. *Journal of The Faculty of Engineering and Architecture Of Gazi University*, **35**, 709- 724.

[22] Sariyer, G., Mangla, S. K., Kazancoglu, Y., Ocal Tasar, C., Luthra, S. (2021). *Data analytics for quality management in Industry 4.0 from a MSME perspective, Annals of Operations Research*, Springer, Netherlands.

[23] Saad, H. (2020). The Application of Data Mining in the Production Processes, *Industrial Engineering*, **2**, 26-33.

[24] Demir, E., Dincer, S. E. (2020). Place and Solution Proposals of Data Mining in Production Planning and Control Processes: A Business Application, *Global Business Research Congress*, **11**, 189 – 193.

[25] Zeng, S. X., Liu, H. C., Tam, C. M., Shao, Y. K. (2008). Cluster analysis for studying industrial sustainability: an empirical study in Shanghai. *Journal of Cleaner Production*, **16(10)**,1090-1097.

[26] Lovell, M. C. (1983). Data Mining, *The Review of Economics and Statist*, **65**, 1-12.

[27] Han, J., Kamber, M., Pei, J. (2011). *Data Mining: Concepts and techniques* (3rd ed.). Elsevier.

[28] Luan, J., Willet, T. (2000). *Data mining and knowledge management: A system analysis for establishing a tiered knowledge management model*. Aptos, CA: Office of Institutional Research, Cabrillo College.

[29] Wang, X. Z. (1999). *Data Mining and Knowledge Discovery for Process Monitoring and Control*, Springer London. 13-28.

[30] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17(3)**, 37.

[31] Sumathi, S., Sivanandam, S. N. (2006). *Introduction to Data Mining and its applications.* Springer.

[32] Ian H. W. (1999). *Data Mining: Practical machine learning tools and techniques.* 3rd edition. Waltham, USA: Elsevier.

[33] Girija, N., Srivatsa, S. K. (2006). A research study: Using Data Mining in knowledge base business strategies. *Information Technology Journal*, **5(3)**, 590–600.

[34] Hand, D. J. (1998). Data Mining: Statistics and more? *The American Statistician*, **52(2)**, 112–118.

[35] Kononenko, I., Kukar, M. (2007). *Machine Learning and Data Mining*. Elsevier.

[36] Data mining (2021). Intl.: www.foursmediagroup.com/DataMining.php. 12.05.2021.

[37] Data mining vs Machine Learning (2021). Intl.: www.ubuntupit.com/data-mining-vs-machine-learning-top-20-things- you-must-know. 12.05.2021.

[38] Carlin B. P. (2008). *Bayesian methods for data analysis*. CRC Press.

[39] Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*. 465-474.

[40] Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*. **2**, 230- 265.

[41] McCulloch, W. S. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*. **5**, 115-133.

[42] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17(3)**, 37.

[43] Empirical inference (2013). Festschrift in honor of Vladimir N. Vapnik. Schölkopf, Bernhard, Zhiyuan Luo, and Vladimir Vovk, Springer Science & Business Media.

[44] Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton.

[45] The White House Names Dr. DJ Patil as the First U.S. Chief Data Scientist (2015). Intl.: www.obamawhitehouse.archives.gov/blog/2015/02/18/white-house-names-dr- dj-patil-first-us-chief-data-scientist. 12.05.2021.

[46] Azevedo, A., Santos M.F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *International Journal of Intelligence Science*. 182-185.

[47] Olson, D. L., Delen, D. (2008). *Advanced data mining techniques*. Heidelberg: Springer Science & Business Media.

[48] Hammori, M., Herbst, J., Kleiner, N. (2004). Interactive Workflow Mining. *Business Process Management: Second International Conference*, BPM 2004. 211-226.

[49] Jayasree, V., Balan, S. (2013). A review on data mining in banking sector. *American Journal of Applied Sciences*, **10(10)**, 1160-1165.

[50] What is data mining? (2021). Intl.: www.talend.com/resources/what-is-data-mining/. 12.05.2021.

[51] Han, J. J. (2011). *Data mining: concepts and techniques*. 3rd edition. Waltham, USA: Elsevier.

[52] Berkhin, P. (2006). A survey of clustering data mining techniques, in *Grouping multidimensional data*. Springer. 25-71.

[53] Quinlan J. R. (2014). C4. 5: programs for machine learning. San Mateo, California: Morgan Kaufman Publishers.

[54] Buddhinath, G., Derry, D. (2003). *A Simple Enhancement to One Rule Classification*. Melbourne, Australia.

[55] Rajput, A., Ramesh, P. A., Meghna, D., Saxena, S. P., Manmohan, R. (2011). J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security (IJCSS)*- 201 - 207.

[56] Zheng, J. G., Jiao, L. C. (2001). *Using deviation detection for data mining*. Springer.

[57] Zhang, C., Zhang, Z. (2002). *Association rule mining: models and algorithms.* Springer- Verlag: Berlin Heidelberg.

[58] Sing, T. (2005). ROCR: visualizing classifier performance in *R. Bioinformatics*. **21**, 3940-3941.

[59] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*. **27**, 861-874.

[60] Tan, P. N., Steinbcah, M., Kumar, V., (2006). *Introduction to Data mining.* MA: Pearson Education.

[61] Karimollah Hajian-Tilaki, (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med.* Spring; **4(2)**, 627–635.

[62] Paidi, A. N. (2012). Data mining: Future trends and applications. *International Journal of Modern Engineering Research (IJMER).* **2**, 4657-4663.

[63] Data Mining Tutorial. Intl.: www.javatpoint.com/data-mining., 12.05.2021.

[64] Hall, M. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. **11**, 10-18.

[65] Kirkby R. E. (2006). *Weka explorer user guide for version 3.5.8.* University of Waikato.

[66] K. A. Abdul Nazeer, M. P. Sebastian (2009). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. *Proceedings of the World Congress on Engineering*. **1**, London, U.K.

[67] Dunham, M. H. (2006). *Data Mining- Introductory and Advanced Concepts*, Springer- Pearson.

[68] Eisen, M.B., Brown, P. O. (1999). DNA Arrays for Analysis Gene Expression, *Methods in Enzymology*, **303**, 179–205.

[69] Xu, R., Wunsch, D. (2005). Survey of Clustering Algorithms, *IEEE Transactions on Neural Network,* **16(3)**, 645-678.

[70] Iba, W., Langley, P. (2011). *Cobweb models of categorization and probabilistic concept formation*, *Formal approaches in categorization*. Cambridge: Cambridge University Press. 253–273.

[71] Dasgupta, S. (2003). Performance guarantees for hierarchical clustering, *International Conference on Computational Learning Theory*. 351-363.

[72] Filkov, V., Kiena, S. (2004). Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools*, **13(4)**. 863–880.

[73] Dempster, A. P., Laird, N. M., Rubin, D. B. (1997). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.* Series B (Methodological), **39(1)**, 1-38.

[74] Gupta M., Chen, Y. (2010). Theory and Use of the EM Algorithm. *Foundations and Trends in Signal Processing*, **4(3)**, 223–296.

[75] Sharma, N., Bajpai, A., Litoria, R. (2012). Comparison the various Clustering Algorithms of weka tools, *International Journal of Emerging Technology and Advanced Engineering*, **2(5)**, (78-80).

[76] Celeux, G., Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 781- 793.

APPENDIX

APPENDIX A: Pseudocodes of Online Learning Algorithms

A. The J48 Classifier: Algorithm 1. A pseudo code of the J48 classifier

Algorithm of J48 (D)

Input: a dataset D

begin

Tree = {}

If (D is "pure") || (other stopping criteria met) then terminate;

For all attribute a α € D D do

Compute criteria of impurity function if we split on α;

a best= Best attribute according to above computed criteria

Tree = Create a decision node that tests $abest$ in the root

Dv= Induced sub-datasets from D based on $abest$

For all Dv, do

begin

$Treev$= J48(Dv)

Attach Tree v to the corresponding branch of Tree end

return Tree end

B. Naive Bayes: Algorithm 2. Pseudocode of the Naive Bayes

Input:

Training dataset T,

F= ($f1, f2, f3,.., fn$) // value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;

3. Repeat

Calculate the probability of fi using the gauss density equation in each class;

Until the probability of all predictor variables ($f1$, $f2$, $f4$,.., fn) has been calculated.

4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

C. K-Nearest Neighbor: Algorithm 3. Pseudocode of the KNN classification.

k-Nearest Neighbor

Classify (X, Y, x) // X: training data, Y: class labels of X, , x: unknown sample for i =1 to m do

Compute distance d(Xi, x)

end for

Compute set / containing indices for the k smallest distances d(Xi, x). return majority label for {Yi, where i € I}

D. RIPPER: Algorithm 4. Pseudocode of the RIPPER Algorithm.

procedure BUILORULESET(P,N) P = positive examples

N = negative examples

RuleSet = {}

DL = DeseriptionLength(RuleSet, PN)

while P ≠ {}

// Grow and prune a new rule

split (P, N) into (Grow Pos, GrowNeg) and (PrunePos, PruneNeg)

Rule := GrowRule(GrowPos, GrowNeg)

Rule := PruneRule(Rule, PrunePos, PruneNeg)

add Rule to RuleSet

if DescriptionLangth(RuleSet, P,N) > DL + 64 then

// Prune the whole nule set and exit

for each rule 2 in RuleSet (considered in reverse order)

if DescriptionLength(RuleSet —(Dft}. PN) < DL then

delete R from RuleSet

DL = DeseriptionLength(RuleSet, P, N)

end if

end for

return (RuleSet)

end if

DL := DeseriptionLength{ RuleSet, P, N)

delete from P and N all examples covered by Rule end while

end BuiLDRuLeser

procedure OpTimizeRuLeSer(RuleSet, P, N)

for each rule # in HuleSet

delete # from RuleSet

U Pos := examples in P not covered by RuleSet

UNeg := examples in N not covered by RuleSet

split (UPos,UNeg) into (Grow Pos, Grow Neg) and (Prune Pos, PruneNeg)

RepRule := GrowRule(Grow Pos, Grow Neg)

RepRule := PruneRule(RepRule, Prune Pos, PruneNeg)

RevRule := GrowRule(GrowPos, GrowN eg, R)

RevRule := PruneRule(Rev Rule, Prune Pos, PruneNeg)

choose better of RepRule and RevRule and add to RuleSet end for

end OpTimizeRuLeset

procedure Rippen(P, N, k)

RuleSet := BuioRuceSer(P, N)

repeat k times RuleSet := OpTiizeRuLeSet(RuleSet, P, N)

return (RuleSet)

E. K-MEANS CLUSTERING: Algorithm 5. Pseudocode of the K-Means Algorithm.

Input:

D = {d1, d2,. ,dn} //set of n data items.

k // Number of desired clusters

Output:

A set of k clusters.

Steps:

1. Arbitrarily choose k data-items from D as initial centroids;
2. Repeat

   Assign each item di to the cluster which has the closest centroid;

   Calculate new mean for each cluster;

   Until convergence criteria is met.

F. HIERARCHICAL CLUSTERING: Algorithm 6. Pseudocode of the Hierarchical Clustering Algorithm.

1. Turn each input element into a singleton, i.e, into a cluster of a single element.
2. For each pair of clusters c1, c2 calculate their distance d (c1, c2).
3. Merge the pair of clusters that take the smallest distance.
4. Continue the step 2, until the termination criterion is satisfied.
5. The termination criterion most commonly used, a threshold of the distance value.

G. COBWEB ALGORITHM: Algorithm 7. Pseudocode of the COBWEB Algorithm.

```
COBWEB(root, record):
Input: A COBWEB node root, an instance to insert record
if root has no children then
        children: = {copy(root)}
        newcategory(record) \\ adds child with record's feature values.
        insert(record, root) \\ update root's statistics
else
        insert (record, root)
        for child in root's children do
                calculate Category Utility for insert (record, child),
                set best1, best2 children w. best CU.
        end for
        if newcategory(record) yields best CU then
                newcategory(record)
        else if merge (best1, best2) yields best CU then
                merge(best1, best2)
                COBWEB (root, record)
        else if split(best1) yields best CU then
                split(best1)
                COBWEB (root, record)
        else
                COBWEB(best1, record)
        end if
end
```

APPENDIX B: Arff File

@relation QualityProblemsInCarpetManufacturing

@attribute Quality NUMERIC
@attribute ProblemDepartment NUMERIC @attribute Shift {1-Shift,2-Shift,3-Shift}
@attribute Customer NUMERIC @attribute Class
{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,
37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65}

@Data
16,1,2-Shift,92,15
71,1,2-Shift,165,9
71,1,2-Shift,165,25
86,1,2-Shift,165,61
86,1,2-Shift,188,54
33,3,2-Shift,165,39
45,3,2-Shift,165,39
93,4,2-Shift,48,63
143,4,2-Shift,56,63
86,4,2-Shift,188,57
177,6,2-Shift,35,2.

CURRICULUM VITAE

PERSONAL INFORMATION

Name Surname: Aysu BORSÖKEN

EDUCATION

Degree	Field	School/University	Graduation
M.Sc	Industrial Engineering	Gaziantep University	2021
B.Sc	Industrial Engineering	Gaziantep University	2017
High School		Özel İdare Anatolian High School	2012

PUBLICATIONS / PRESENTATIONS

Borsöken A. Unutmaz Durmuşoğlu Z. (2021), Classification of Quality Problems in Carpet Manufacturing by Using Data Mining Algorithms, IEOM 2021: First Central American and Caribbean Industrial Engineering and Operations Management.