DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEVELOPMENT OF DECISION SUPPORT ALGORITHMS ON RFID SYSTEMS OF STORES

by

Boran Taylan BALCI

November, 2016

İZMİR

DEVELOPMENT OF DECISION SUPPORT ALGORITHMS ON RFID SYSTEMS OF STORES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Sciences
in Computer Engineering

by
Boran Taylan BALCI

November, 2016
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DEVELOPMENT OF DECISION SUPPORT ALGORITHMS ON RFID SYSTEMS OF STORES"** completed by **BORAN TAYLAN BALCI** under supervision of **PROF. DR. ALP KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Alp KUT

Supervisor

Y. Doç. Dr. Derya BİRANT

(Jury Member)

Y.Doç.Dr. Özgür CAN

(Jury Member)

Prof.Dr. Emine İlknur CÖCEN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Dr. Alp Kut who allowed me to work in this project and I appreciate for his valuable ideas, support and guidance throughout this project.

Also I would like to thank Giltaş Inc. employees and their CEO Feyzullah Oktay who allowed me to participate in 7441350 numbered TÜBİTAK-approved project. Special thanks to Gamze Özçelik and Barış Hamil for their great support on various parts of the project.

Boran Taylan BALCI

DEVELOPMENT OF DECISION SUPPORT ALGORITHMS ON RFID SYSTEMS OF STORES

ABSTRACT

In today's world, RFID technology is playing effective role in many parts of our lives, such as supply chain, logistics, item-inventory tracking and so on. This technology is seen by some as inevitable replacement for barcodes especially in clothing industry. The increase of this method in order to identify products and customers, motivated Giltaş Inc. to develope a project called "Smart Fitting Room" that involves auto identification and recommendation system based on both customers and products in retail industry. We are assigned to this project for RFID implementation and recommendation system. There is also Nebim Integration part which is excluded on this study.

This project involves 2 sub projects; RFID implementation and data mining. RFID implementation has been done concerning various parts of the physical store. These parts consist of fitting rooms, express lanes, spesific locations which the owner of a store can decide to trace customers in the store and exit doors for security purposes. In second project, a data warehouse has been established by filtering the existing data from Nebim ERP data tables. Based on the purchased item transactions, first the products sold together, which in other words frequent itemsets are extracted by using Apriori algorithm. Then the frequent itemsets are used to make recommendations based on the products which are identified in the fitting room. Classification and clustering methods have been used together to provide customer-based recommendation. Data attributes have been determined as catogerical data which includes gender, age (different age intervals have been grouped to describe spesific audience), city and top hierarchies of items (for instance; blue 2015 v-neck t-shirt expressed as only t-shirt). First customers have been clustered by using k-means algorithm based on these predetermined data attributes. Then cluster results have been used as target classes for classification with the same attribute set. J48 algorithm has been executed to construct a decision tree. The decision tree has been used to classify a customer who is identified in the fitting room. Once the customer has been classified, best-selling products related to that cluster have been recommended to customer.

The scope of the research has been considered as the companies that use ERP and offer service in retail industry. The project has been implemented mainly as web project and it also contains Android application for push notifications.

Keywords: Data warehouse, data mining, web services, ERP, clustering, classification, RFID, decision tree, association rule mining, android

MAĞAZA RFID SİSTEMLERDE KARAR DESTEK ALGORİTMALARININ GELİŞTİRİLMESİ

ÖZ

Günümüz dünyasında RFID teknolojisi, tedarik zinciri, lojistik, ürün envanter takibi gibi hayatımızın benzeri birçok alanında önemli rol oynamaktadır. Bu teknoloji bazı şirketler tarafından özellikle giyim sektöründe barkodun yerine kaçınılmaz bir değişim olarak görülmektedir. Müşterileri ve ürünleri tanımlamakta kullanılmaya başlanan bu teknolojinin artması, Giltaş AŞ’yi “Akıllı Kabin” olarak adlandırılan bir proje geliştirmeye sevketmiştir. Proje, içerisinde müşteri ve ürün bazlı otomatik tanımlama ve öneri sistemlerini içermektedir. RFID uygulaması ve veri madenciliği bölümleri tarafımızca gerçekleştirilmiştir. Ayrıca Nebim Entegrasyon bölümü de bulunmakta, fakat bu çalışmada yer almamıştır.

Bu proje, RFID uygulaması ve veri madenciliği olarak 2 alt projeden oluşmaktadır. RFID uygulaması, fiziksel mağazanın çeşitli bölümlerine ilişkin olarak gerçekleştirilmiştir. Bu bölümler; soyunma kabinleri, hızlı kasalar, mağaza sahibinin müşteri takibi için karar vereceği belirli yerler ve güvenlik amaçlı yapılmış çıkış kapılarıdır. Projenin ikinci kısmında, Nebim ERP veri tablolarında var olan veriler filtrelenerek bir veri ambarı oluşturulmuştur. Satış verileri temel alınarak, birlikte satılan ürünler diğer bir deyişle sık tekrar eden ürün kümeleri Apriori algoritması yardımıyla belirlenmiştir. Daha sonra bu ürün kümeleri, kabinde tanımlanan ürüne ilişkin öneri oluşturmakta kullanılmıştır. Sınıflandırma ve kümeleme metotları, müşteri bazlı öneri yapmak için birlikte kullanılmıştır. Veri özellikleri kategorik olarak seçilmiş ve içinde cinsiyet, yaş (belirli yaş aralıkları belirli kitleleri ifade edecek şekilde gruplandı), en üst hiyerarşi bilgilerini (örneğin; mavi 2015 v-yaka tişört, sadece tişört olarak ifade edildi.) içermektedir. İlk olarak müşteriler, önceden karar verilen bu özellikler kullanılarak gruplanmıştır. Daha sonra kümeleme sonuçları sınıflandırma için hedef sınıf olarak, aynı özellik kümesi ile beraber kullanılmıştır. Bir karar ağacı oluşturmak için J48 algoritması kullanılmıştır. Karar ağacı, soyunma kabininde algılanan müşteriyi sınıflandırmak için kullanılmıştır.

Müşteri sınıflandırıldığı anda, ait olduğu kümeye ait en çok satılan ürünler müşteriye önerilmiştir.

Bu projenin odağı olarak gıda dışı perakende sektöründe ERP kullanan firmalar ele alınmıştır. Proje temel olarak web projesi olup, bildirim alımları için ayrıca Android uygulaması barındırmaktadır.

Anahtar kelimeler: Veri ambarı, veri madenciliği, web servisleri, ERP, kümeleme, sınıflandırma, RFID, karar ağacı, birliktelik kuralı analizi, android

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Background

Radio Frequency IDentification (RFID) defines the designated transmission channel to accomplish object identification, from stock control and object tracing to cardreading. Comparing to barcode identification technology, RFID is much quicker and more profound for identifying and gathering information about objects (Di Marco, Santucci & Fischione, 2014). This is one of the reasons that we used RFID technologies for recognitions of products and customers. Also there are a lot of tools available you can easily programme the RFID Reader.

Billions of bytes of data is being travelled across the network and stored in devices regarding many aspects of life such as business, science, medicine and so on. This massive data is being generated by businesses which have millions of transactions, scientific corporations which gather continuous observations via remote sensors, health industries which use medical records and so on. To transform these big chunks of data into knowledge can be named as data mining in other words knowledge discovery (Han, Pei & Kamber, 2011). Many companies including Giltaş Inc. understood that knowledge discovery in these enormous data is essential for providing marketing insights (Shaw, Subramaniam, Tan & Welge, 2001). In order to keep their customers' attention, companies use CRM applications which shape their assistances and supplies based on customer choices (Peppers, Rogers & Dorf, 1999). Since these companies store transactions related to their sales in detail, that gives opportunity to get better understanding about diverse customer profiles (Rygielski, Wang & Yen, 2002). The aim of better analysis of the customers can be achieved by extracting hidden customers' distinctive properties and creating a model from existing data (Giraud-Carrier & Povel, 2003). Classification, regression, association rule mining, clustering, visualization etc. methods have been described in many papers depending on needs of organisations (Ngai,Xiu & Chau, 2009). In our study we focused on classification, clustering and association rule mining methods which will be explained in next chapters.

The main reason that Giltaş Inc. advanced this groundbreaking project is there was no AI/DC (Automatic Identification and Data Capture) supported smart systems in the country. Smart keyword implies here segmentation of customers based on purchasing habits, discovering customer tendency so called *trends*, producing recommendations based on association between products and customers.

1.2 Purpose

In this thesis we aim to use data mining methods mentioned above shortly to provide functionality of analysis of data and recommendations. Also those technological novelties including identification and recommendation will be integrated in retail industry for the first time in our country. The satisfaction of customers will improve with the recommendations provided. Also with mobile part implementation, interaction between employees and valuable customers will be more satisfactory, since, before these VIP customers call for aid, a salesperson will be contacting with them by knowing their identity.

We are aiming that innovations made in both technological and operational parts will improve the companies' sales, make them lead the industry and take company one step further in the industry by using these tools. This project will be an infrastructure for further projects. And it will be widely used across all the country.

1.3 Organization of the Thesis

This thesis includes 7 chapters and the rest of the thesis is organised as follows:

In Chapter 2, case studies have been discussed regarding automatic identification and data mining methods.

In Chapter 3, different identification systems have been described shortly and RFID technology has been explained with detail.

In Chapter 4, RFID implementation of the project has been shown with the screenshots which represent the main aspects for the project.

In Chapter 5, data mining concepts related to association rule mining, clustering and classification have been described.

In Chapter 6, data mining implementation of the project has been illustrated with the screenshots regarding data preparation, execution and results.

In Chapter 7, experimental results of proposed system has been shown.

In Chapter 8, the conclusion and future works have been discussed.

CHAPTER TWO
LITERATURE REVIEW

2.1 Automatic Identification and Data Capturing in RFID Clothing Industry

There are plenty of ways to identify an object, in clothing sector it usually refers barcodes. With the recent innovations in technology it is possible to apply a complete supply chain including end user experience.

Usually the clothing industry had been run late to utilize and improve their technology that other industries like electronic and automative have. But, with the existence and widely usage of RFID technology, a number of clothing enterprises took the oppurtinity for the development of this technology rapidly and improved it to go one step further with their competitives. RFID technology has now been widely used among the clothing industry, varying from clothing retailers to stores and logistics (Wong & Guo, 2014).

The technology is not used in fashion widely by saying that RFID technology had the power to make a remarkable impact on retail supply chain (SC) operations. By remotely reading the item code and other information on a tag that is coded the bits it has in it and making data available to information systems like SC, RFID technology provides better inventory control and reducing stock-out by having the control of inventory, savings in costs due to keeping track with RFID is much easier, and to yield fewer transaction errors. But, altough several pilot implementations in the retail fashion sector, there is still a low understanding of the value of RFID technology and, especially, RFID item tagging (Wong & Guo, 2014).

2.1.1 Case Studies

2.1.1.1 Zara (Retailer Company)

According to the article published in The Wall Street Journal, by 2016 the transformation of their products to RFID in Zara company has been nearly completed. The companies such as WalMart wanted to use this technology, but it hasn't fitted very well because of metal interaction between a reader and a tag. They learn from competitors' experience and now Inditex SA is rolling out RFID technology for their operations (Bjork, 2014).

Before the RFID tags were introduced, employees had to scan barcodes for each product and these inventory checking had been performed once every six months. By saving time with RFID tags, the company keeps track of their stock items every six weeks, getting a more spesific information about what the trend is and what they are selling well and any styles that are outworn (Bjork, 2014).

Also the company has a mobile implementation. When the customer cannot find the desired size or color of an item, the salesperson is able to take a picture with his mobile device and checks the similar items and availability in that store or other stores of Zara nearby (Bjork, 2014).

The companies like Wal-Mart had to downsize the project after suppliers complained about the cost of the technology and the company didn't face with the problem because they had their own manufacturers (Bjork, 2014).

2.1.1.2 S.Culture International Holdings Limited

S.Culture Internation Holdings Limited one of the retail companies which mainly sells shoes from different brands like Clarks, Josef Seibel etc. What makes this company special is one of the stores in Hong Kong made a fully integrated RFID integration in all operational processes. A small part of what they have done covers what we did in our project.

You can find the video on youtube under the title “How RFID Benefits Retail Fashion: Host Louis Sirico”. Louis Sirico is the CEO of Executive Lifestyle Furnished Homes, Inc. & ELF Express Hotels. His responsibilities include company P&L, operations, corporate expansion, and new business development. He was formerly the creator and host of The RFID Network, an educational TV program broadcast on 23 US cable channels to an audience of 1.5 million viewers, and is a well-known industry expert in the field of Radio Frequency Identification. In 2002, he was nominated for Entrepreneur of the Year by the Entrepreneur Council of Maryland and then in 2004, sold the business he founded, RFID Wizards, for 5 times revenue.

2.1.1.3 Microsoft’s Smart Fitting Room Is Like A Robo-Shop Clerk

Mark Wilson, who is a writer in *Fastcodesign*, describes the futuristic fitting room made by Accenture and Microsoft corporation. Every product in the store is labeled with an RFID tag. RFID readers are located inside the fitting room. Whenever customer enters the fitting room, the screen located inside displays different size or color of the item. When the user clicked the button on the screen, the notification of requested color or size is received by the clerk without asking for help physically in the room (Wilson, 2014).

The article also emphasizes that the recommendation part can be applied like in Amazon for a customer. When the customer brings a T-shirt into the fitting room, it can be deduced that he will probably buy that product. What it can be done next is recommending other items like Amazon’s “who bought a T-shirt also purchased this” (Wilson, 2014).

About privacy issues that might be an obstacle in the future that customer might not want you to know he or she wants a XL size. That doesn’t seem a big threat to the application but still one of the nominees that might be a problem (Wilson, 2014).

2.2 Data Mining in Implicit Data Sets

As online shopping plays an essential role and nowadays become very popular, an important task emerged, retrieving the right item over wide range of products to satisfy the customers most. One of the well known approaches is called recommender systems (Hu, Koren & Volinsky, 2008).

2.2.1 Recommendation Systems

Recommender systems are based on different types of input. Most suitable is the high quality explicit data, which includes explicit feedback by users who rates the items based on their judgement. For instance, Netflix gathers feedbacks for its movies. Clients make the feedback for their movie and TV series selection by hitting like or dislike buttons. But, explicit data cannot be reached all the time and even might not exist. So, recommendations could be deduced for client selections using bigger implicit dataset, which mediately the preferences of user behavior (Hu, Koren & Volinsky, 2008). Types of implicit input could cover the topics like purchasing, browsing in websites, the keyword you wrote on search boxes, or even mouse click and moves.

The article “*A TV Program Recommender Framework”* aims to filter the channels with information gathered by satellites or digital video archives and make a recommendation system based on these meta-data. The recommendation part is implemented to achieve faster browsing through the channels by filtering or featuring them based on his profile. This profile is fed by implicit dataset which is gathered from TV habits (Chang, Irvan & Terano, 2013).

Amazon is also doing recommendations, which is an implicit dataset example, based on the selected item. System understands, for instance, that users who looked at the title *Jupiter's Travels: Four years*, also glanced the media CD of *Long Way Round* which tells about experiences of a young biker and his bosom friend. Like in the example of image below in Figure 2.1, this notion is managed to present products for the part "Customers who viewed this also viewed." (Zacharski, 2015).

Figure 2.1 Customers who viewed this also viewed from Amazon

Another implicit rating can be which product a client really purchases. The firm additionally monitors this data and with using this data produces recommendations "Frequently Bought Together" and "Customers Who Bought This Item Also Bought" like in Figure 2.2 (Zacharski, 2015).

Figure 2.2 Frequently bought together and customers who bought this item also bought

Even though recommendation systems are used by various companies, there are some drawbacks such as cold start and data sparsity. Usually retail companies have big chunk of itemsets in their database, a problem occurs related to customers since they purchase or rate very small part of that big data. Inferring preferences of these cold users cause recommendations to be inaccurate (Guo, Zhang & Thalmann, 2014). Because of these reasons, we wanted to proceed with different approach by combining clustering with classification.

2.2.2 Clustering via Classification

According to the paper published by M. Lopez, it is suggested to create a classifier that process a cluster with a classification algorithm dependently the idea that every cluster matches to a label (Figure 2.3). Initially, the data from student portal activity had to be gathered and processed. After, the arbitrary attribute deduction algorithm like SVM (Support Vector Machine) or PCA (Principle Component Analysis) can be applied in order to select spesific attributes that matters more than the others. Next step, they executed a clustering approach by obtaining the training data, then filtering target label, matching targets between clusters and classes. That matching was used to guess target classes on behalf of unknown objects in the training set. Putting it differently, target label didn't get selected for clustering, but it was selected for assess the clusters as classifiers. They used this approach to indicate if there is correlation between participation in forums and passing or failing the course (Lopez, Luna, Romero & Ventura, 2012).

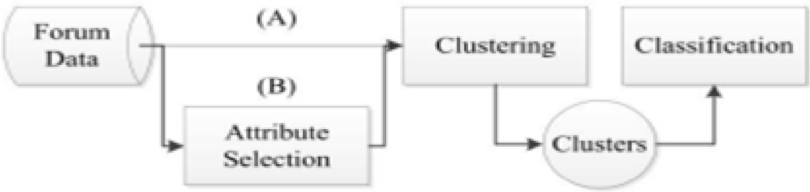


Figure 2.3 Proposed classification via clustering approach

R. S. Kamath mentions similar approach for educational data analysis in his book. In his research, academicians' achievement details are gathered and after filtering, a clustering algorithm is implemented. Then extracted clusters are used as meta-classifier for classification process. The aim of the research is to infer the performance of students in further examination and to find out the ones who need private attention and advices (Kamath & Kamat, 2016).

CHAPTER THREE
CUSTOMER AND PRODUCT IDENTIFICATION

3.1 Automatic Identification and Data Capture Technologies

This section examines some of the technologies available for AI/DC. There are 4 main approaches that might be used to identify the object; barcode systems, biometric systems, smart card systems, rfid systems.

3.1.1 Barcode Sytems

Barcode technology is widely used in the retail shopping sector. It is a low-cost and simple technology. It is binary cipher including fields like bars with spaces regulated for a collateral arrangement. These bars with spaces are regulated in preset model and points out a symbol (Liwan, 2015).

Despite appearing identical, there are considerable differences between each barcode. This is the result of the different coding techniques used in the design. The European Article Number (EAN) is the most common type of coding used for designing barcodes. A barcode has a data density of 100 bytes. Barcodes are usually read with optical scanners. The different reflections of the laser gleam to dark bars following with white spaces assist in interpreting the bars and graphs on a barcode numerically and alphanumerically. The optical scanner has to be placed very close (10-50 cm) and in the line of sight of the barcode for data to be read from it (Finkenzeller, 2013).

A barcode system could have been an ideal approach for the current manual system because barcode systems are cheap and easy to operate. However, these advantages are negated by the fact that they are affected by dirt, highly susceptible to wear and tear, and fail completely if the barcode is blocked from the direct view of the optical scanner. Also it requires a lot affort to be identified by optical scanner, alignment of barcode and keeping close to the scanner.

3.1.2 Biometric Sytems

Biometrics is the study of methods of uniquely recognising people based upon one or more intrinsic physical characteristic. There are various biometric techniques, but in keeping with the context of this project, dactyloscopy or fingerprint scanning will be examined in some detail (Cole, 2009).

Fingerprint scanning was first used, and is still being used, by criminologists. Criminal offenders are fingerprinted when they are charged with a crime. If there is a match between a fingerprint found at a crime scene and the one stored in the criminal database, this is regarded as conclusive evidence against the criminal, as fingerprints differ in every person (Cole, 2009).

Fingerprint readers are used in dactyloscopy. Users must first register their fingerprints in the central database. This is done by placing the fingertip on the reader. The reader framework computes the information over the fingerprint figure and stores the info in a memory (Cole, 2009).

Once the fingerprints of the users have been registered in the database, the fingerprint reader can be used to identify the users. Every time a user enters the fitting room, their fingers are scanned. A match between the scanned images and those already stored in the database will confirm the user.

The advantage of the biometric scanning systems is being very accurate, compact and resistant to data tampering. However, the high cost and complexity of the system make it less attractive compared to the other technologies available for automating identification.

3.1.3 Smart Card Sytems

Smart card systems are mainly used for electronic data storage. Their applications range from prepaid telephone cards to the SIM cards used in GSM mobile phones. Smart cards are equipped with galvanic contacts. The smart card is provided with the necessary voltage and pulse from the smart card reader when the two come into contact with each other (Rankl, 2014).

Two different classes of smart cards exist, namely memory cards and microprocessor cards. Memory cards have an Electrically Erasable Programmable Read-Only Memory (EEPROM). The end application that needs to be run using the memory card is stored in the EEPROM. The security algorithms used in the card are also stored in the EEPROM (Rankl, 2014). The advantage of the memory card is that it is very cheap to manufacture. However, low data storage capacity and susceptibility to wear and tear have resulted in memory cards slowly being phased out of the market.

Microprocessor cards, on the other hand, have different sectors, namely a Read-Only Memory (ROM), a Random Access Memory (RAM) and an EEPROM. As a result microprocessor cards can store many more applications. This advantage of the microprocessor card is negated by its cost.

3.1.4 RFID Sytems

RFID systems are similar to smart cards except that they do not have to be physically in contact with the RFID reader. Data stored in an RFID card are transferred via radio waves to the RFID reader.

RFID systems comprise of an RFID transponder, an RFID reader and RFID middleware. RFID transponders have a very high data density. RFID transponders are small microchips that can store data. RFID systems are not influenced by dirt or by obscuring the tags.

RFID readers have a range of up to 5 m without the transponder being in the line of sight of the RFID reader. The advantages mentioned in this section have prompted the use of RFID in automating the identification of both customers and products.

3.2 Comparison between Various Automatic Identification Systems

In previos section the different technologies available for automating the customer and product identification were discussed. This section will examine the pros and cons of each of these technologies, bearing in mind the objective of this project. The aim of this section is to narrow down the technology that can be used for identification of customers and products.

A comparison is made of the technologies with respect to some of the vital parameters concerned with identifying products and customers (Table 3.1).

Table 3.1 Comparison between different automatic identification systems

Parameters	Barcode system	Biometric system (Dactyloscopy)	Smart card system	RFID systems
Data density (bytes)	Low data density 100 bytes	High data density	Very high data density -16-64 kb	Very high data density
Influence of dirt	Very high	No influence	High if contacts come in contact with dirt	No influence
Influence of covering the data carrier	Total failure of system	Total failure as system works on contact	Total failure as system works on contact with the smart card	No influence
Influence of direction between reader and data carrier	Failure - if no line-of-sight communication	Not applicable as direct contact is needed	Not applicable as direct contact is needed	No influence as data are transferred via radio waves
Wear and tear of data carrier	Limited - if not tampered with intentionally	Not applicable	Possible with extended use	No influence
Purchasing cost	Low	High	Low	Low
Operating cost	Low	High	Low	None
Reading speed in seconds	Low - up to 4 seconds	Very low - 5-10 seconds	Low – 4 seconds	Very fast - 0.5 to 1 second
Distance between reader and data carrier in centimetres	0-50 cm	Direct contact	Direct contact	0-6 m depending on the frequencies used

3.3 Introduction to Radio Frequency Identification (RFID) Technology

The history of RFID date back to middle of 20th century. In the time of World War II, British forces developed a system called IFF (Identity Friend or Foe) in order to differentiate their aircrafts from the enemies'. Since then, a lot of companies use this technology all over the world for different purposes (Domdouzis, Kumar & Anumba, 2007). In 2004, Wal-Mart advanced a pilot RFID application as an example of retailer. After that, various retail firms adopted RFID implementations and that gave birth to academic studies in clothing sector (Moon & Ngai, 2008).

In our study, the RFID technology has been considered an ideal solution for identifying customers and products in retail clothing industry. Section 3.1.4 briefly introduced the RFID technology. This section aims to elaborate on that discussion.

An RFID framework originates in three main sub parts, namely the RFID tag, the RFID reader and RFID middleware.

3.3.1 RFID Tags

RFID tags (here after referred to as tags) are also called transponders. The tag is placed on the object that needs to be identified. It contains an internal antenna and a microchip. The microchip stores the data which define and distinguish each tag. There are three types of tags in use; active tags, passive tags and semi-passive tags.
Active tags incorporate a battery along with the antenna and the microchip. The battery affects the cost and size of active tags. As a result active tags are not very commonly used (Peris-Lopez, Hernandez-Castro, Estevez-Tapiador & Ribagorda, 2015).

Passive tags do not have a built-in battery. The power requirements of a passive tag are generated from the electric or magnetic fields generated by the RFID reader. Passive tags are very cheap and smaller than active tags (Peris-Lopez, Hernandez-Castro, Estevez-Tapiador & Ribagorda, 2015). As a result they have been used in our study for identifying products and customers

Semi-passive tags have an onboard power source and may have onboard sensors. The onboard power source provides a continuous power source for the sensors. This enables the semi-passive tags to transfer data even in the absence of an RFID reader. The semi-passive also has an increased read range. The cost of semi-passive tags lies between the costs of active and passive tags (Peris-Lopez, Hernandez-Castro, Estevez-Tapiador & Ribagorda, 2015).

3.3.2 RFID Reader

All RFID tags contain a microchip which stores data that distinguish each tag. The data contained in each tag must be transmitted. The transmission midpoint of an RFID system is referred to as the RFID reader (referred to as a reader from now on). The reader reads the data in the tag and sends the data to the RFID middleware (Müller, 2013).

This section examines the physical components of the reader and the different types of readers available. There are three components to the reader: the antenna, the controller and the network interface (Müller, 2013). It is shown in Figure 3.1:

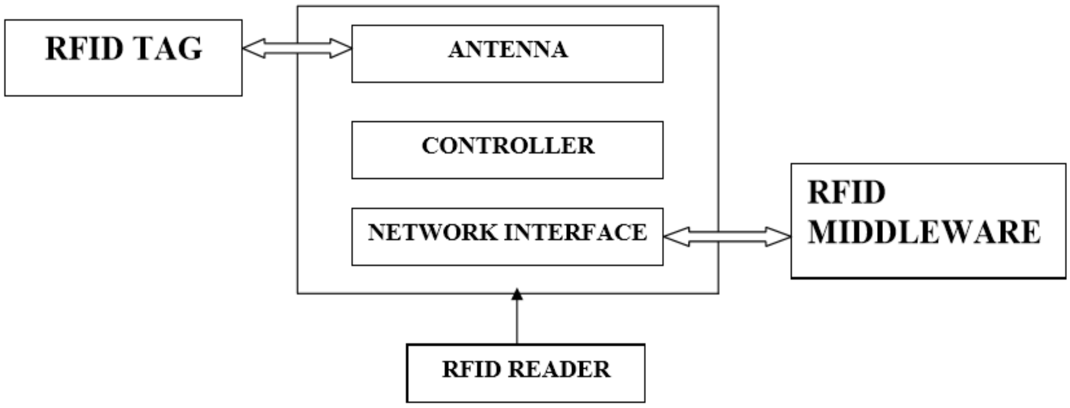


Figure 3.1 Phsyical components of an RFID reader

All RFID readers need an antenna as the tag communicates with the reader using radio frequency (RF). The antenna acts as a receptor of the RF waves. This makes the antenna the most important component of an RFID reader (Müller, 2013).

The antenna is designed such that the radio frequency waves it receives are optimised for the centre frequency ranges. This is high-precision work which requires considerable attention during the antenna design stage and fine tuning of the design properties (Müller, 2013).

All readers need a controller to run the different processes involved in reading RFID tags. The complexity of the reader varies. A reader can be equipped with only a single embedded chip which can function as a simple-state machine, or it can run an entire operating system with substantial hard disk space and RAM.

The data read from the tags by the reader must be transferred to a device which recognises and manipulates the data. This is where a network interface is needed.

3.3.3 RFID Middleware

RFID middleware serves three main purposes: to capture data from the network interface of the reader and input it to an end-user application; to process the data from the reader so as to allow the end-user application to see only the necessary data, and to provide an application level interface for managing the reader.

Based on this, basic RFID middleware should consist of three principal components; the reader adapter, the event management unit and the application level interface. A block diagram illustrating these three components is given in Figure 3.2.

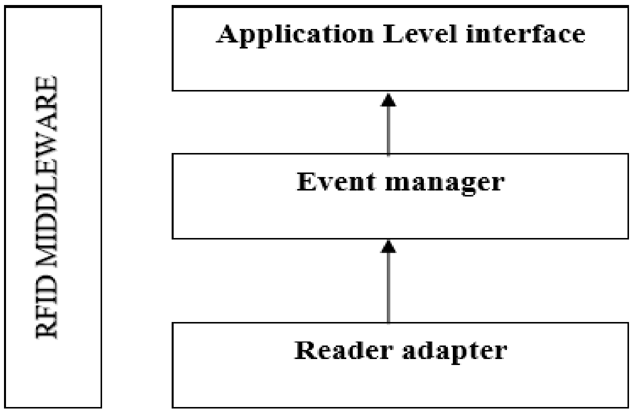


Figure 3.2 Components of middleware

CHAPTER FOUR

RFID IMPLEMENTATION IN OUR PROJECT

In our study, Octane SDK has been used as a middleware which is provided by Impinj company for programming in .NET environment. Before using RFID tags in our project, in order to initialize the RFID tags' EPC values with different values, we used the software shown in Figure 4.1 and random EPCs has been written to the selected RFID tags. The reason of necessity of this process was that RFID tags usually come with the same EPC at the first place. There has to be an initiation process for each tags.

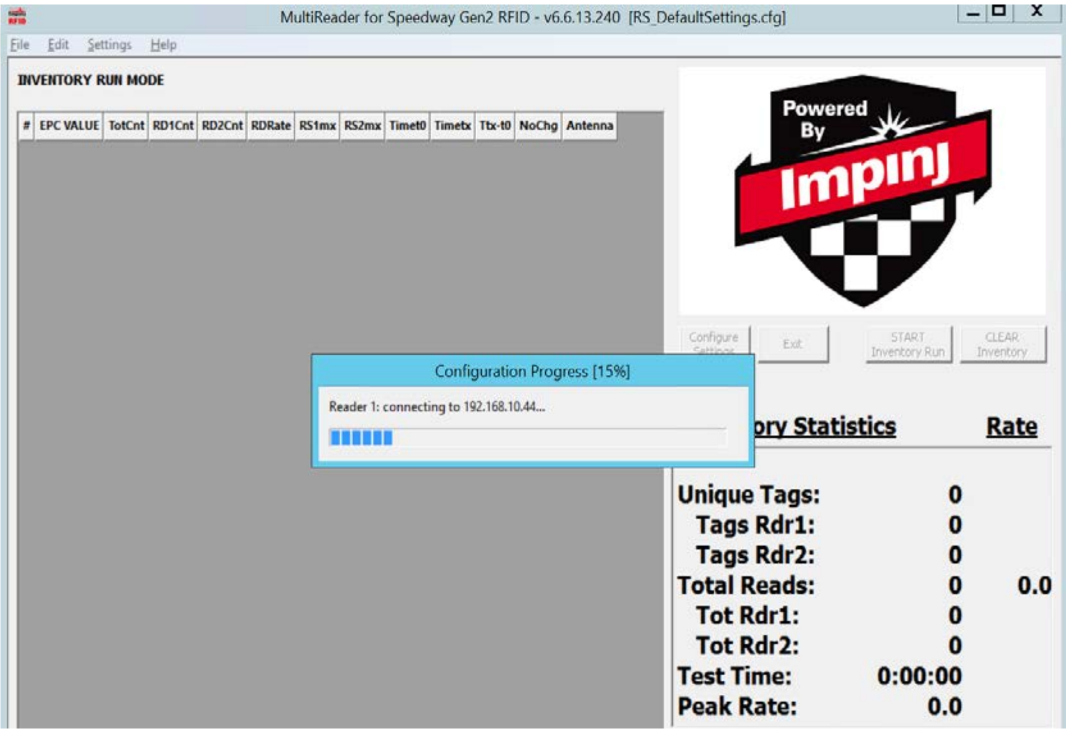
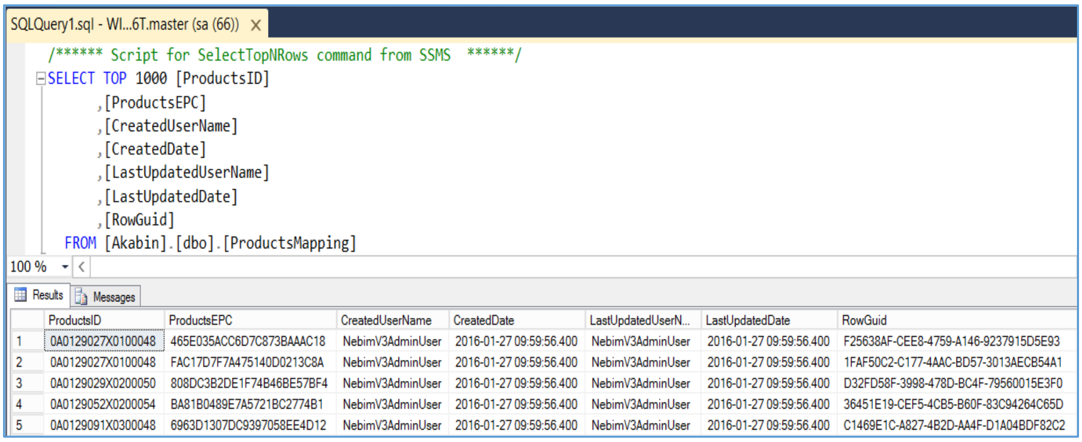


Figure 4.1 MultiReader for Speedway powered by Impinj

After the initiation part, the tags' EPCs are mapped to desired product IDs which are in this case barcode numbers. The mapping phase hasn't been automated yet. This process has been made manually along with Giltaş Inc.. You can see on Figure 4.2 and Figure 4.3 the data structure and the records of mapping for both customer and product.



SQLQuery1.sql - WI...6T.master (sa (66))

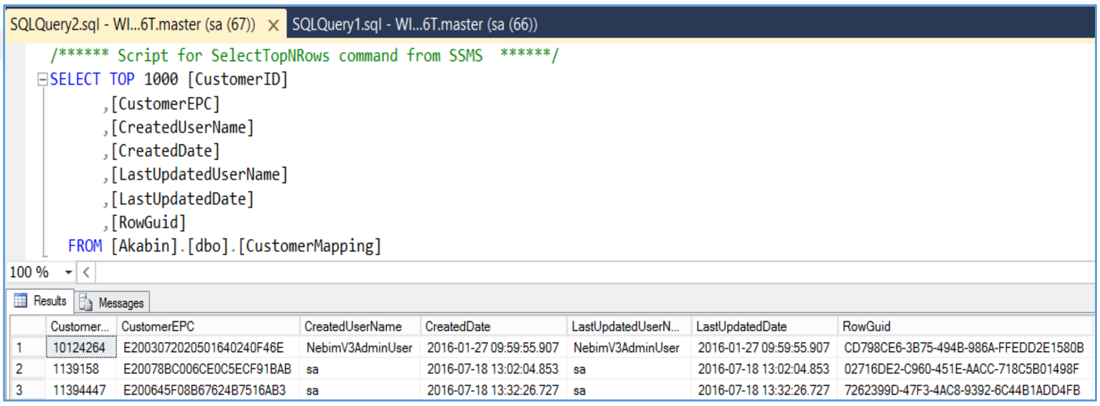
```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [ProductsID]
      ,[ProductsEPC]
      ,[CreatedUserName]
      ,[CreatedDate]
      ,[LastUpdatedUserName]
      ,[LastUpdatedDate]
      ,[RowGuid]
  FROM [Akabin].[dbo].[ProductsMapping]
```

100 %

Results | Messages

	ProductsID	ProductsEPC	CreatedUserName	CreatedDate	LastUpdatedUserN...	LastUpdatedDate	RowGuid
1	0A0129027X0100048	465E035ACC6D7C873BAAAC18	NebimV3AdminUser	2016-01-27 09:59:56.400	NebimV3AdminUser	2016-01-27 09:59:56.400	F25638AF-CEE8-4759-A146-9237915D5E93
2	0A0129027X0100048	FAC17D7F7A475140D0213C8A	NebimV3AdminUser	2016-01-27 09:59:56.400	NebimV3AdminUser	2016-01-27 09:59:56.400	1FAF50C2-C177-4AAC-BD57-3013AECB54A1
3	0A0129029X0200050	808DC3B2DE1F74B46BE57BF4	NebimV3AdminUser	2016-01-27 09:59:56.400	NebimV3AdminUser	2016-01-27 09:59:56.400	D32FD58F-3998-478D-BC4F-79560015E3F0
4	0A0129052X0200054	BA81B0489E7A5721BC2774B1	NebimV3AdminUser	2016-01-27 09:59:56.400	NebimV3AdminUser	2016-01-27 09:59:56.400	36451E19-CEF5-4CB5-B60F-83C94264C65D
5	0A0129091X0300048	6963D1307DC9397058EE4D12	NebimV3AdminUser	2016-01-27 09:59:56.400	NebimV3AdminUser	2016-01-27 09:59:56.400	C1469E1C-A827-4B2D-AA4F-D1A04BDF82C2

Figure 4.2 An example of products mapping



SQLQuery2.sql - WI...6T.master (sa (67)) | SQLQuery1.sql - WI...6T.master (sa (66))

```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [CustomerID]
      ,[CustomerEPC]
      ,[CreatedUserName]
      ,[CreatedDate]
      ,[LastUpdatedUserName]
      ,[LastUpdatedDate]
      ,[RowGuid]
  FROM [Akabin].[dbo].[CustomerMapping]
```

100 %

Results | Messages

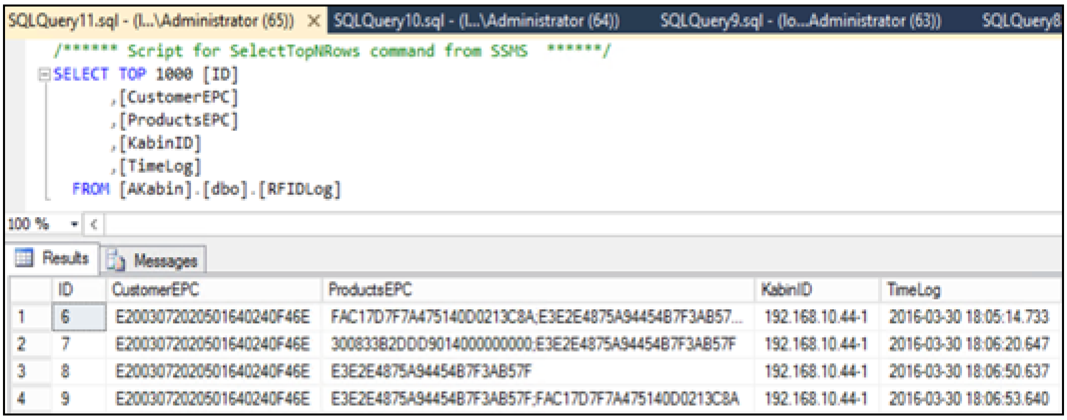
	Customer...	CustomerEPC	CreatedUserName	CreatedDate	LastUpdatedUserN...	LastUpdatedDate	RowGuid
1	10124264	E2003072020501640240F46E	NebimV3AdminUser	2016-01-27 09:59:55.907	NebimV3AdminUser	2016-01-27 09:59:55.907	CD798CE6-3B75-494B-986A-FFEDD2E1580B
2	1139158	E20078BC006CE0C5ECF91BAB	sa	2016-07-18 13:02:04.853	sa	2016-07-18 13:02:04.853	02716DE2-C960-451E-AACC-718C5B01498F
3	11394447	E200645F08B67624B7516AB3	sa	2016-07-18 13:32:26.727	sa	2016-07-18 13:32:26.727	7262399D-47F3-4AC8-9392-6C44B1ADD4FB

Figure 4.3 An example of customer mapping

There are 4 RFID projects which all have spesific aims and processes need to be accomplished. These are Console Applications. There is no UI interface to observe. They just store the data in spesific data tables. Those 4 projects cover all the steps from fitting room to sale and sale to security. Below sections will cover the projects with detail.

4.1 RFID Main Recorder

This is the backbone of the entire RFID project. The main recorder solution gathers the EPC tags and cards which are representing product and customer IDs, in a fitting room and store them in a log table called “RFIDLog”. In order not to inflate the log table with reduntant data, there is another process checking the products with previous state and comparing them to find out any changes occured in the fitting room.



SQLQuery11.sql - (l...\Administrator (65)) | SQLQuery10.sql - (l...\Administrator (64)) | SQLQuery9.sql - (lo...Administrator (63)) | SQLQuery8

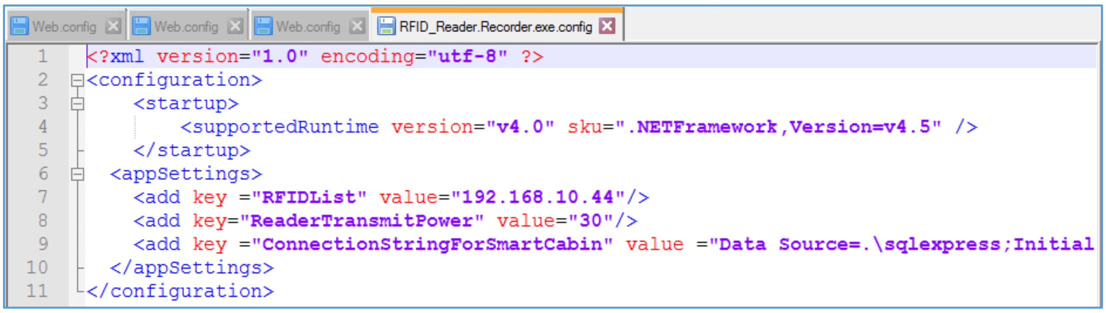
```sql
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [ID]
      ,[CustomerEPC]
      ,[ProductsEPC]
      ,[KabinID]
      ,[TimeLog]
  FROM [AKabin].[dbo].[RFIDLog]
```

Results | Messages

	ID	CustomerEPC	ProductsEPC	KabinID	TimeLog
1	6	E2003072020501640240F46E	FAC17D7F7A475140D0213C8A;E3E2E4875A94454B7F3AB57...	192.168.10.44-1	2016-03-30 18:05:14.733
2	7	E2003072020501640240F46E	300833B2DDD9014000000000;E3E2E4875A94454B7F3AB57F	192.168.10.44-1	2016-03-30 18:06:20.647
3	8	E2003072020501640240F46E	E3E2E4875A94454B7F3AB57F	192.168.10.44-1	2016-03-30 18:06:50.637
4	9	E2003072020501640240F46E	E3E2E4875A94454B7F3AB57F;FAC17D7F7A475140D0213C8A	192.168.10.44-1	2016-03-30 18:06:53.640

Figure 4.4 An example of RFID Log

As you can see in Figure 4.4, multiple tags are seperated by semicolon. There is a KabinID column and that is seperated with a dash sign and a number follows. That is actually the number of the antenna. For trial purposes we worked on one antenna mainly, due to detection range of second one was not proper for the fitting room. There can be more than one antenna and each antenna represents different fitting rooms. That information will be used for setting front-end UI and notification of salesperson about product request and where it came from. There is also config file worth to mention in Figure 4.5 that you can set RFID IP list, transmit power and connection string to database which data will be stored. The chunk of data diagram is shown in Figure 4.6 that shows the columns of Mapping tables and RFID Log table.



Web.config | Web.config | Web.config | RFID_Reader.Recorder.exe.config

```xml
<?xml version="1.0" encoding="utf-8" ?>
<configuration>
    <startup>
        <supportedRuntime version="v4.0" sku=".NETFramework,Version=v4.5" />
    </startup>
  <appSettings>
    <add key ="RFIDList" value="192.168.10.44"/>
    <add key="ReaderTransmitPower" value="30"/>
    <add key ="ConnectionStringForSmartCabin" value ="Data Source=.\sqlexpress;Initial
  </appSettings>
</configuration>
```

Figure 4.5 Rfid main reader config file

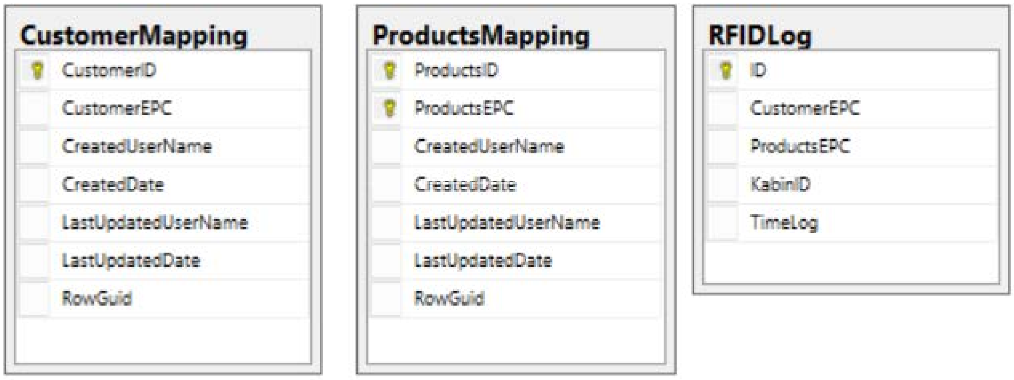
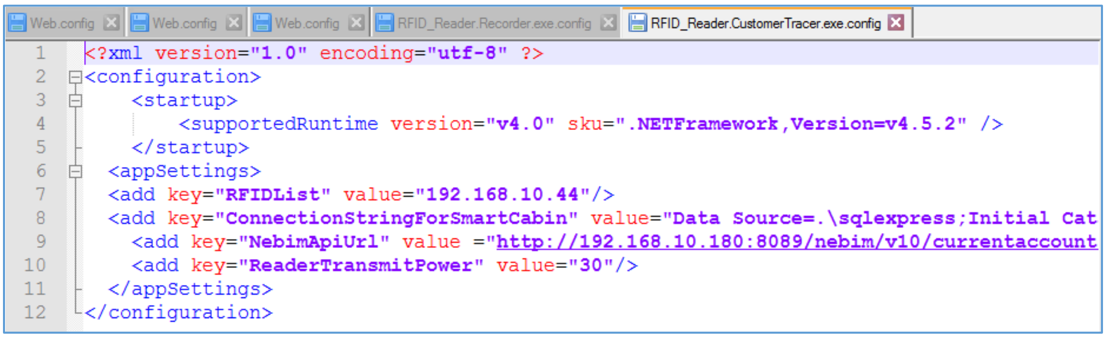


Figure 4.6 Customer, products mapping and RFIDLog tables

4.2 RFID Customer Tracer

Every solution after RFID Main Reader project has been implemented as sub projects of it. There are always similar approaches with little differences. In Customer Tracer project, the aim is to keep track of special customers' RFID Cards. After identifying the VIP customer, by using one of free push notification server called "parse.com", the registered mobile devices will be notified with the existence of the VIP customer. In Figure 4.7, there is the same key value pair with Main RFID project but one additional line is added. That line is the way to communicate with the WEB API which is written for integration with Nebim ERP and through this WEB API to send request to parse.com in order to push notification.




```
<?xml version="1.0" encoding="utf-8" ?>
<configuration>
    <startup>
        <supportedRuntime version="v4.0" sku=".NETFramework,Version=v4.5.2" />
    </startup>
  <appSettings>
  <add key="RFIDList" value="192.168.10.44"/>
  <add key="ConnectionStringForSmartCabin" value="Data Source=.\sqlexpress;Initial Cat
    <add key="NebimApiUrl" value ="http://192.168.10.180:8089/nebim/v10/currentaccount
    <add key="ReaderTransmitPower" value="30"/>
  </appSettings>
</configuration>
```

Figure 4.7 Customer tracer config file

The database diagram is showed in Figure 4.8:

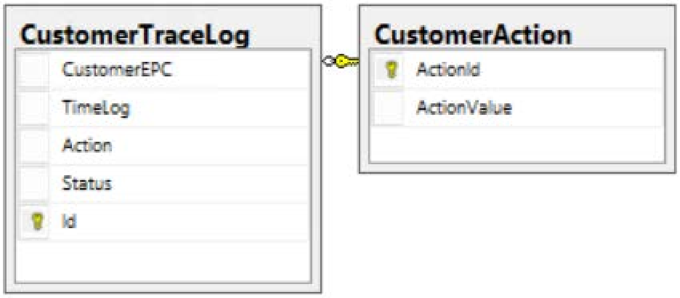


Figure 4.8 Data diagram of customer tracer

The examples of data records are illustrated in Figure 4.9 and 4.10. Whenever the reader starts to log, it means a VIP customer is passing through the antennas. Those antennas might be located in different parts of the store. Table CustomerAction describes where exactly the customer past through; entrance, exit, fitting room and so on. The reader keeps logging from the first moment it detects, till the notification process ends or the customer walks away to out of range of antenna.

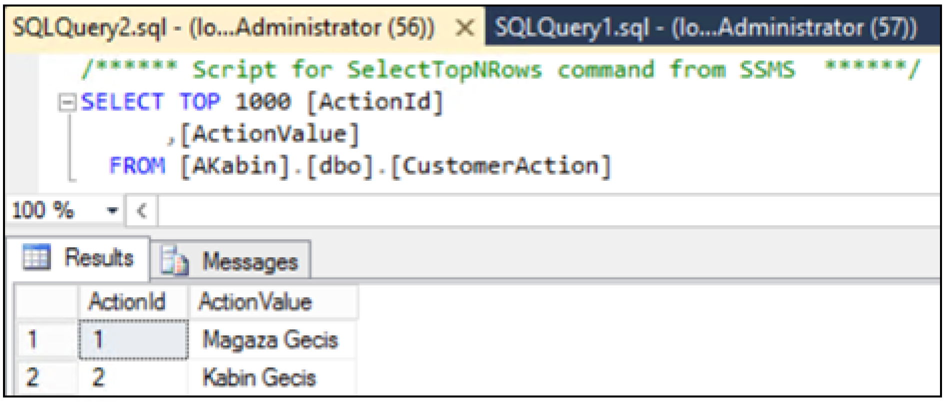


Figure 4.9 Customer action table

SQLQuery1.sql - (lo...Administrator (57)) WIN-F501UA8TT6T\...Kabin - Diagram_0*

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [CustomerEPC]
      ,[TimeLog]
      ,[Action]
      ,[Status]
      ,[Id]
  FROM [AKabin].[dbo].[CustomerTraceLog]
```

100 %

Results | Messages

	CustomerEPC	TimeLog	Action	Status	Id
1	E2003072020501640240F46E	2016-04-08 18:11:56.023	1	Open	1
2	E2003072020501640240F46E	2016-04-08 18:13:10.900	1	Open	2
3	E2003072020501640240F46E	2016-04-08 18:15:06.443	1	Open	3
4	E2003072020501640240F46E	2016-04-08 18:20:46.013	1	Open	4
5	E2003072020501640240F46E	2016-04-08 18:20:47.390	1	Notified	5

Figure 4.10 Customer tracelog table

4.3 RFID Express Lane Recorder (In Progress)

The aim of this part is to log products and customers for the last phase of shopping. The same principle has been applied as Main Reader project. Customers and products have been stored in different columns, yet multiple products are seperated with semicolon. There are additional tables related to sale process based on logs. The small chunk of diagram has been showed in Figure 4.11.

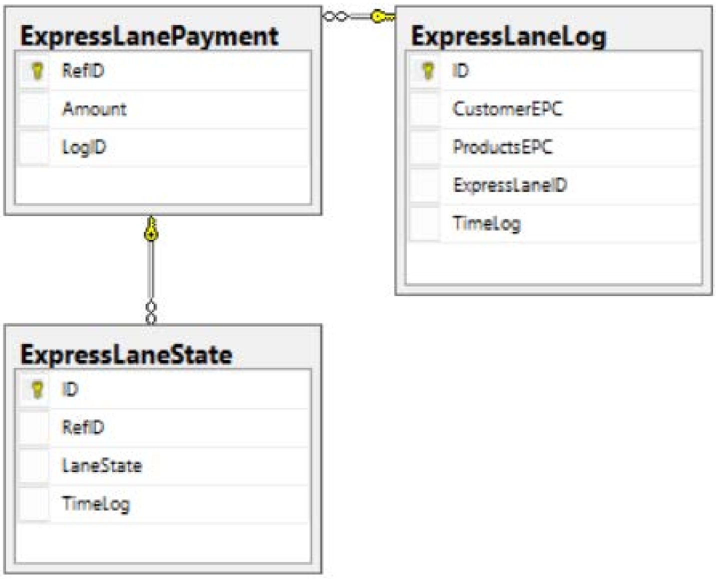
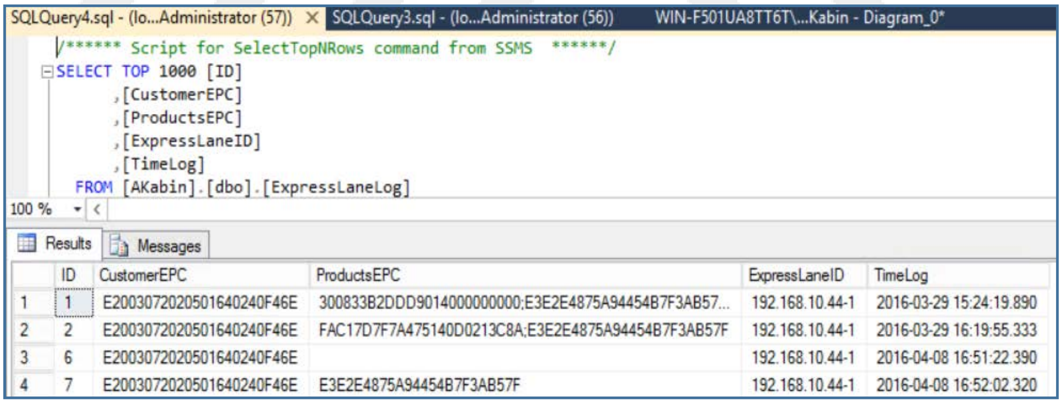


Figure 4.11 Data diagram of express lane

Express Lane Log example is illustrated in Figure 4.12:



	ID	CustomerEPC	ProductsEPC	ExpressLaneID	TimeLog
1	1	E2003072020501640240F46E	300833B2DDD9014000000000;E3E2E4875A94454B7F3AB57...	192.168.10.44-1	2016-03-29 15:24:19.890
2	2	E2003072020501640240F46E	FAC17D7F7A475140D0213C8A;E3E2E4875A94454B7F3AB57F	192.168.10.44-1	2016-03-29 16:19:55.333
3	6	E2003072020501640240F46E		192.168.10.44-1	2016-04-08 16:51:22.390
4	7	E2003072020501640240F46E	E3E2E4875A94454B7F3AB57F	192.168.10.44-1	2016-04-08 16:52:02.320

Figure 4.12 An example of express lane log

Express Lane Payment example is shown in Figure 4.13:

SQLQuery5.sql - (lo...Administrator (58)) SQLQuery4.sql - (lo...Administrator (57))

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [RefID]
      ,[Amount]
      ,[LogID]
  FROM [AKabin].[dbo].[ExpressLanePayment]
```

100 %

Results | Messages

	RefID	Amount	LogID
1	14032273-B027-41A2-9AA5-04BBBB624021	14,50	15
2	4828A7EF-85BB-42C3-A252-0513E3ACBD14	132,01	15
3	1FA926B6-B19D-4F00-9557-2A9C16D9015B	132,01	2
4	D98848DA-A8DA-4C63-A723-44DD475E6046	132,01	15
5	1D634762-337A-47B6-9B37-4D11976070A0	143,00	15

Figure 4.13 An example of express lane payment

Express Lane State records are shown in Figure 4.14:

SQLQuery6.sql - (lo...Administrator (60)) SQLQuery5.sql - (lo...Administrator (58)) SQLQ

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [ID]
      ,[RefID]
      ,[LaneState]
      ,[TimeLog]
  FROM [AKabin].[dbo].[ExpressLaneState]
```

100 %

Results | Messages

	ID	RefID	LaneState	TimeLog
1	24	1FA926B6-B19D-4F00-9557-2A9C16D9015B	pending	2016-04-07 11:21:04.213
2	25	83FC0ACF-9EC3-4677-846A-5E2E3C17EBF0	pending	2016-04-08 16:56:14.153
3	26	83FC0ACF-9EC3-4677-846A-5E2E3C17EBF0	fail	2016-04-08 16:56:53.200
4	27	83FC0ACF-9EC3-4677-846A-5E2E3C17EBF0	fail	2016-04-08 16:59:00.557
5	28	B11EC9B9-F68A-4424-BAA4-5B2671D7BDBB	pending	2016-04-11 09:39:17.717
6	29	B11EC9B9-F68A-4424-BAA4-5B2671D7BDBB	fail	2016-04-11 09:41:06.067
7	30	D98848DA-A8DA-4C63-A723-44DD475E6046	pending	2016-04-11 09:53:45.640
8	31	D98848DA-A8DA-4C63-A723-44DD475E6046	fail	2016-04-11 09:54:07.543
9	32	0C7DCE55-C0DE-47C4-A085-8CF1060F65B6	pending	2016-04-12 15:10:59.113
10	33	0C7DCE55-C0DE-47C4-A085-8CF1060F65B6	success	2016-04-12 15:12:17.457
11	34	4828A7EF-85BB-42C3-A252-0513E3ACBD14	pending	2016-04-12 15:23:54.137
12	35	4828A7EF-85BB-42C3-A252-0513E3ACBD14	success	2016-04-12 15:24:12.393

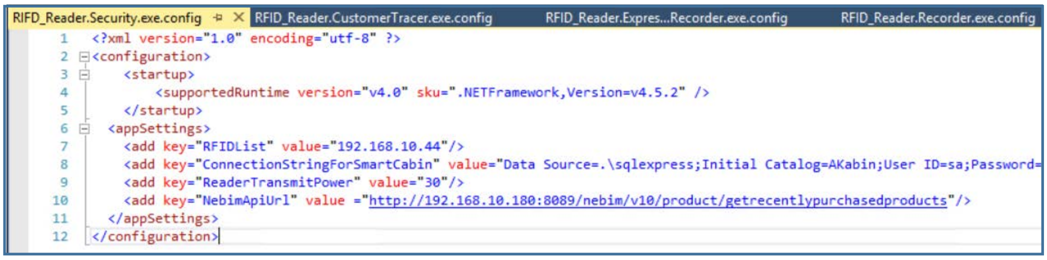
Figure 4.14 An example of express lane state

Those 3 tables work in a consecutive way. After consistent logging part, when a customer clicks the button and confirms the products that are shown in the Kiosk screen which are retrieved from Express Lane Logs, a record will be inserted to Express Lane Payment. This record will be with a unique GUID and Log ID with the price calculated through Nebim WEB API. After the customer clicks to pay button, a record will be inserted to Express Lane State with the same GUID that is created on previous step and state information which will describe the current state of payment. Automatically pending value will be inserted as state information at the first step. After the confirmation of payment from banks or some paying devices has been detected, another record will be inserted with the state information whether success or fail. Depending on the second state, front end will notify the user by selecting the last row's state information of current GUID.

As you can notice this process has been temporarily suspended. The latest law regulation says that in order to produce a software that has payment in it, you need to get special permission and be controlled by the experts to get it. Since it is a research and development project we didn't have neither time nor budget to face with this problem.

4.4 RFID Security

The best part of RFID systems is providing security without that big detectors attached to the clothes. Since RFID tags will be used in clothes there is no necessity for detector. There hasn't been need to create log table for security issues. The security process is implemented directly into the reader.



RIFD_Reader.Security.exe.config | RFID_Reader.CustomerTracer.exe.config | RFID_Reader.Expres...Recorder.exe.config | RFID_Reader.Recorder.exe.config

```
<?xml version="1.0" encoding="utf-8" ?>
<configuration>
    <startup>
        <supportedRuntime version="v4.0" sku=".NETFramework,Version=v4.5.2" />
    </startup>
  <appSettings>
    <add key="RFIDList" value="192.168.10.44"/>
    <add key="ConnectionStringForSmartCabin" value="Data Source=.\sqlexpress;Initial Catalog=AKabin;User ID=sa;Password=
    <add key="ReaderTransmitPower" value="30"/>
    <add key="NebimApiUrl" value ="http://192.168.10.180:8089/nebim/v10/product/getrecentlypurchasedproducts"/>
  </appSettings>
</configuration>
```

Figure 4.15 RFID security config file

As you can see in Figure 3.17 above, there is another url defined in the config file. By using that link, RFID Security whenever reads a tag, it retrieves all the sold products for that day. The working mechanism is related to Express Lane Recorder project. In express lane, after payment is succesfully done, the related invoice information is being inserted to ERP Tables using Nebim V3 Entegrator. In that invoice information there is a column called line description for each product sold that you can write additional info about the product. That part has been filled by EPC values. Thanks to this property, we were able to differentiate the products that have same barcode. So after calling Nebim WEB API on return, there is an array full of EPCs. If product IDs would have been brought instead of EPCs, we wouldn't be able to distinguish the products that have same barcode value.

4.5 Mobil Part and Push Notifications

The mobile part has been implemented in accordance with Customer Tracer project. An Android project has been implemented in order to broadcast notifications. A spesific URL has been mentioned for pushing notifications in chapter 4.2. This URL actually is an indirect request to "parse.com" API to deliver notifications the users who have the applications. When you registered the site, you are able to post REST requests to https://api.parse.com/ with the given application Id and REST API key. The post requests have been done to https://api.parse.com/1/push URL. Then mobile devices which have the app have been notified by the text written in the body of request. The notification text and app's UI is shown in Figure 4.16:

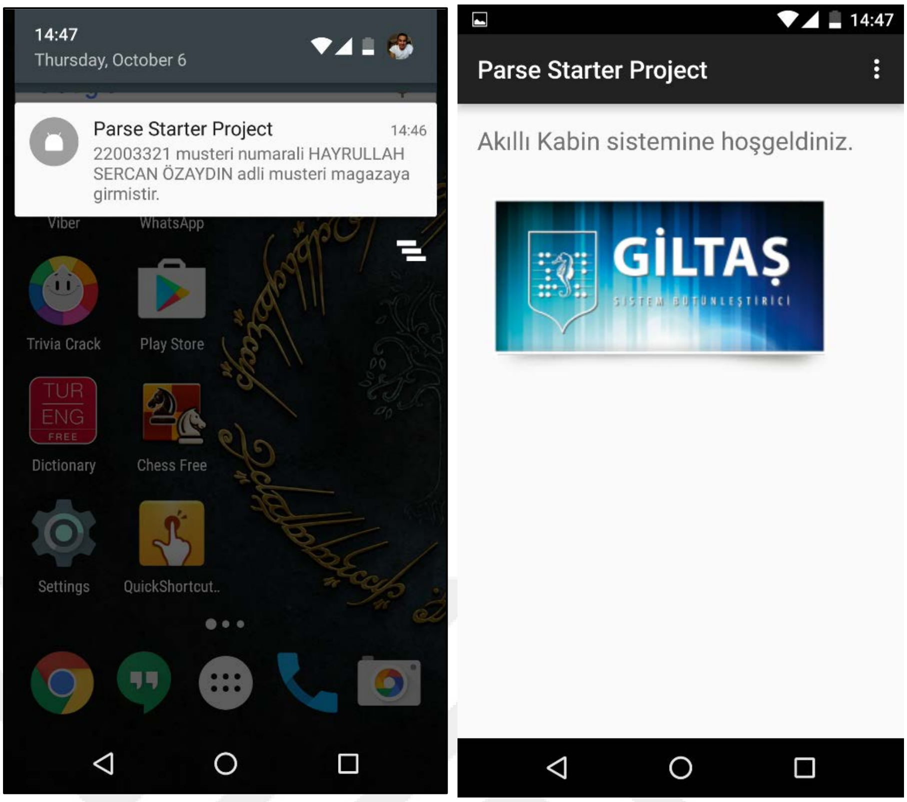


Figure 4.16 Notification Example

CHAPTER FIVE
DATA MINING METHODS

5.1 Clustering

Clustering can be defined as exploring groups without knowing characteristics of objects in a dataset. Clustering methods have been used in many applications of computer science such as text mining, machine learning, computer vision and so on (Kogan, 2007).

Since a cluster can be defined as a set of objects that are “similar” to each other, it wouldnt be wrong if we think the objects are “dissimilar” to the other ones.

As you can see on Figure 5.1 simple clustering example:

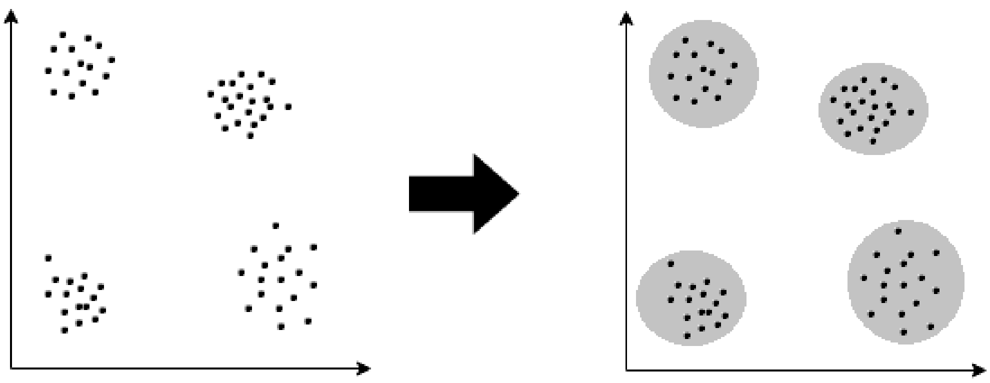


Figure 5.1 Clustering example

Since clustering is bundling the similar data points, some sort of measure should be defined in order to detect whether two objects are similar or dissimilar. Famous distance measures can be summarized as: Eucledian Distance, Minkowski Distance, Manhattan Distance etc.

After describing the similarity function, there are couple of methods which are grouped regarding the results. Mainly it can be categorized as partition-based, hierarchical and density-based clustering.

Partitioning-based clustering: It creates one level segregation of objects which are assigned to k numbers of clusters depend on their similarities. The examples of methods can be explained as (Kriegel, 2005):

- K-means
- K-medoids
- CLARA and CLARANS

Hierarchical clustering: In this approach, unlike partitioning based methods, it segments data into several groups (Gan, Ma & Wu, 2007).

- Divisive approach is a top down approach.
- Agglomerative approach is a bottom up approach.

Density-Based Methods: is also widely used and can be used for exploring closely packed points. A cluster will be occured if density of data reaches a predetermined number. Common approaches are:

- DBSCAN
- OPTICS

Although there are plenty of methods available, k-means still has most popularity among others and simple to implement. In our study, we have used k-means for clustering customers.

5.1.1 K-means Algorithm

K-means is one of the widely used approaches between clustering algorithms. The algorithm pursues a easy approach in order to classify the object by definite quantity (what k describes) of clusters. The essential part is defining k centroids which will be the centers of each cluster. These centroids shoud locate carefully because of locating them in different position may cause different result. Usually main approach is to put centroid to the furthest points from each other. The further phase is to get each represented object and calculate the distance between centroids whether to associate one of the clusters. When all objects are assigned, the first grouping is done. Afterwards new k centroids should be calculated by getting the mean of represented objects. After we have these k new centroids, the same approach shoul be applied in a loop until that k centroids stay still (Ebrahimzadeh, 2012).

Finally, below equation is an objective function which determines to reduce the value, like in below equation which is a squared error function. Objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{5.1}$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is selected for dissimilarity criteria between a represented object $x_i^{(j)}$ and the cluster mid-point c_j , is an pointer of distance measure for n objects and corresponding cluster centroids.

The algorithm flows like mentioned below:

- Get K objects into the dimensional platform which involves data points that will be clustered. These objects will identify first cluster centroids.
- Place every data point to the cluster that has the nearest centre.
- After every represented data points are placed, rearrange the locations of the K centres.
- Do Steps 2 and 3 repeatedly till the centre points stay still.

Despite k-means approach could be validated that the algortihm invariably ends, the k-means method does not certainly pull out the best regulation, based on the global objective function minimum. This approach additionally can be counted as vulnerable to where the initialization of centroids is done. Executing k-means clustering method several instances lower this impact (Ebrahimzadeh, 2012).

5.2 Classification

The business of object classification has numerous solutions in a wide diversity of mining concerns. That is by reason of the algortihm tries to elicit the dependance among a group of attributes and an objective inconstants. As a lot of applicative matters could be described as associations among attribute and objective inconstants, that yields wide set of practicality of the approach. The classification approach might be explained like in below (Kumar, 2010):

Dedicated a group of training objects in addition with matched target classes, establish the object label for an unlabeled sample data (Kumar, 2010).

Classification approaches generally consists of two parts:

• Training Part: In training part, a pattern is built by using set of training objects.
• Testing Part: In testing part, the pattern obtained previously assigns target classes to an unlabeled sample data.

For several cases, for instance lazy learning, the training process could be skipped completely, while the classification could be done straightly by using the dependance between training objects and sample data. Approaches like the nearest neighbor classifiers (K-NN) is one of the main approaches following this way. Yet in some situations, an initiative part like nearest neighbor index processing might handle to guarantee effectiveness for testing part (Kumar, 2010).

The result of a classification method might represent for a test data point in two ways:

- Discrete Label: For this approach, the target class returns to the test sample.
- Numerical Score: For this approach, a numeral score returns to every target class. Keep in mind that the numerical score can be represented as a discrete target class for sample data with choosing the label which has the topmost score for that sample data. The positive thing about numeral scoring would be that it presently comes in available to contrast the notional tendance of various data samples refering to a label of significance, and order data if necessary (Kumar, 2010).

One of the well known algorithm is called J48 algorithm in Weka which generates a Decision Tree. Based on the rules extracted, the new instance is classified. In our project this approach has been used. Naive-bayes, k-nearest-neighbour algorithm, neural networks etc. also are used widely.

5.2.1 Decision Tree Method

Decision tree creates classification or regression patterns for a structure of a tree. It shrinks the sample set to minor datasets meantime concurrently a related decision tree is constructed. The last outcome occurs as desicion tree including nodes and leaf nodes (Friedl, 1997). A decision node might have multiple branches whilst leaf node defines a target whether classification or decision. The node placed to the top of the tree which represents optimal predictor named root node. Decision trees could be used for nominal and numeral data. In Figure 5.2, illustration is showed:

Outlook	Eye Color	Height	Gender	Is European
Dark	Brown	Tall	Female	No
Dark	Brown	Tall	Male	No
Brunette	Brown	Tall	Female	Yes
Blonde	Hazel	Tall	Female	Yes
Blonde	Blue	Normal	Female	Yes
Blonde	Blue	Normal	Male	No
Brunette	Blue	Normal	True	Yes
Dark	Hazel	Tall	Female	No
Dark	Blue	Normal	Female	Yes
Blonde	Hazel	Normal	Female	Yes
Dark	Hazel	Normal	Male	Yes
Brunette	Hazel	Tall	Male	Yes
Brunette	Brown	Normal	Female	Yes
Blonde	Hazel	Tall	Male	No

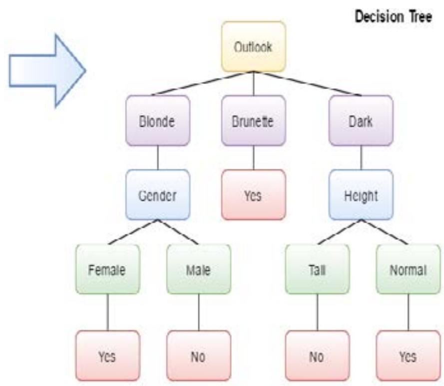


Figure 5.2 Example of decision tree

The construction of algorithm based on 2 factors that will be mentioned in following chapters; Entropy and Information Gain.

5.2.2.1 Entropy

A decision tree might be built from top which is a root node to bottom and contains segmenting the sample set into small sets that subsumes data points which have approximate values. The methodology defines entropy in order to compute the similarity of test dataset. If test data set has total uniform then the entropy is zero and likewise if test data has even division, we can say it has the value one for entropy.

In order to construct the decision tree, initially computation of entropies using frequency tables in two different ways illustrated below:

a) Entropy computation by frequency table of an attribute showed in Figure 5.3.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (5.2)$$

Is European	
Yes	No
9	5

Entropy(IsEuropean) = Entropy(5,9)

= Entropy(0.36,0.64)

= - (0.36 $\log_2$0.36)-(0.64 $\log_2$0.64)

= 0.94

Figure 5.3 Calculation of entrophy

b) Entropy computation with frequency table of two dimensions showed in Figure 5.4.

$$E(T,X) = \sum_{c\in X} P(c)E(c) \tag{5.3}$$

		Is European		
		Yes	No	
Outlook	Blonde	3	2	5
	Brunette	4	0	4
	Dark	2	3	5
				14

$E(IsEuropean, Outlook)$ = P(Blonde) x E(3,2) + P(Brunette) x E(4,0)+ P(Dark) x E(2,3) = (5/14) x 0.971 + (4/14) x 0 + (5/14) x 0.971 = 0.693

Figure 5.4 Calculation of entrophy based on probability

5.2.2.2 Information Gain

Information gain calculation plays essential role to divide training set by choosing the best attribute for seperation. These results are used to decide which attributes should be put for every iteration. For building the tree, it should be looked for the attribute which fetches the best score of information gain (Sugumaran, Muralidharan & Ramachandran, 2007).

Step 1: Calculate entropy of the target (Figure 5.5).

$$Entropy(IsEuropean) = Entropy(5{,}9)$$
$$= Entropy(0.36{,}0.64)$$
$$= -(0.36 \log_2 0.36)-(0.64 \log_2 0.64) = 0.94$$

Figure 5.5 First step of information gain calculation

Step 2: The sample data afterwards divides to varied attributes. The entropy is computed for every branch. Afterwards it is attached relatively, in order to fetch overall entropy in order to divide. The calculated entropy value is subtracted from the entropy before the division. The outcome becomes Information Gain, or reduction in entropy is illustrated in Figure 5.6.

		Is European	
		Yes	No
Outlook	Blonde	3	2
	Brunette	4	0
	Dark	2	3
Gain = 0.247			

		Is European	
		Yes	No
Eye Color	Brown	2	2
	Hazel	4	2
	Blue	3	1
Gain = 0.029			

		Is European	
		Yes	No
Height	Tall	3	4
	Normal	6	1
Gain = 0.152			

		Is European	
		Yes	No
Gender	Female	6	2
	Male	3	3
Gain = 0.048			

Figure 5.6 Second step of information gain calculation

Step 3: Select the attribute which has the greatest information gain value for the decision node, like in Figure 5.7.

		Is European	
		Yes	No
Outlook	Blonde	3	2
	Brunette	4	0
	Dark	2	3
Gain = 0.247			

Figure 5.7 Third step of information gain calculation

Step 4a: A branch which has zero entropy score represents leaf node in Figure 5.8.

Eye Color	Height	Gender	Is European
Brown	Tall	Female	Yes
Blue	Normal	True	Yes
Hazel	Tall	Male	Yes
Brown	Normal	Female	Yes

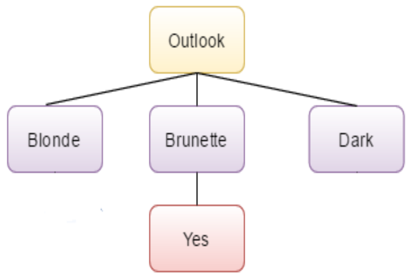


Figure 5.8 One possible result of fourth step of information gain calculation

Step 4b: A branch which has entropy score greater than 0 requires more division like in Figure 5.9.

Eye Color	Height	Gender	Is European
Hazel	Tall	Female	Yes
Blue	Normal	Female	Yes
Blue	Normal	Male	No
Hazel	Normal	Female	Yes
Hazel	Tall	Male	No

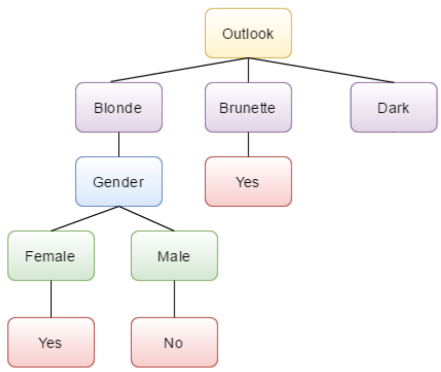


Figure 5.9 Other possible result of fourth step of information gain calculation

Step 5: The method should be executed repeatedly for non-leaf branches, till every object gets labeled.

Previously built decision tree could easefully be turned into pile of rules with mapping starting from the top node to the leaf nodes like in Figure 5.10.

R_1: IF(Outlook=Blonde) AND (Gender=Female) Then European
R_2: IF(Outlook=Blonde) AND (Gender=Male) Then Non-European
R_3: IF(Outlook=Brunette) Then European
R_4: IF(Outlook=Dark) AND (HEIGHT=Tall) Then Non-European
R_5: IF(Outlook=Dark) AND (HEIGHT=Normal) Then European

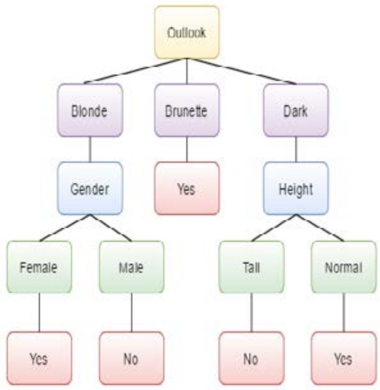


Figure 5.10 Transformation to set of rules

In our project we are using this kind of rules to assign a class to a given new instance.

5.3 Association Rule Mining

Association rule mining was preliminary announced at then end of 20th century and one of the essential and well studied methods of data mining. It achieves to deduct relations, frequent patterns and relations for group of objects sometimes in massive transaction tables (Buttar & Kaur, 2013).

One of the well-known method is called Apriori algorithm which we are using in our project. There is another alternative approach to reduce time complexity called FP-Growth but we didnt follow that approach, the reason that there is a possibility to overload the RAM with the tree data which constructed by all sales.

5.3.1 Apriori Algortihm

The Apriori Algorithm is an efficient approach in order to extract frequent itemsets.

Essential topics are Frequent Itemsets, Apriori Rule, Join Process. Frequent Itemsets, the group of objects that have lowest support (represented as L_i for i-th-Itemset). Apriori Rule is that any subset of frequent itemset also should be frequent. Join Process is for finding L_k which is a group of candidate k-itemsets will be produced with using join process for L_{k-1} with itself (Tanna & Ghodasara, 2014).

The summarized and brief version of the algorithm:

• Look for the frequent itemsets: the group of objects that have lowest support value.

– Think of Apriori Rule for instance:

• if {XY} is a frequent itemset, both {X} and {Y} must be a frequent itemset

– Iteratively look for frequent itemsets by using number from 1 to k (k-itemset)

• Define the frequent itemsets for producing association rules.

The Pseudo code as follows:

• Join Function: C_k is produced by joining process L_{k-1} with itself.

• Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.

• Pseudo-code:

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};

for(k = 1;L_k !=∅; k++) do begin

C_{k+1} = candidates produced by using Lk;

for every transaction t in transaction table do

increase the number of total candidates in C_{k+1} that are within t

L_{k+1} = candidates in C_{k+1} with min_support

End return ∪kL_k;

CHAPTER SIX

DATA MINING IMPLEMENTATIONS IN OUR PROJECT

In our project, we used Weka libraries for .NET. Programming phase has been done in Visual Studio with C# along with Microsoft MS SQL. For Apriori algorithm an executable file downloaded from Internet has been used. For k-means and J48 Weka methods have been used.

6.1 Data Preparation

Since both Weka methods and Apriori.exe file demands in its own input file in a format, there need to be a data preparation before putting them into a process. Weka formatted files include ".arff" extension. There is certain way to describe attributes, target class and data with '@' tags.

Approximately 1.5 million transactions and 100.000 customers have been taken as the scope of sales covering three-month period between March and June. We wanted to exclude season change, in order to provide consistency over data.

In this section some processes are written in run.bat file (Figure 6.1) will be examined. The first, the third and the fifth ones are related to data preparation phase.

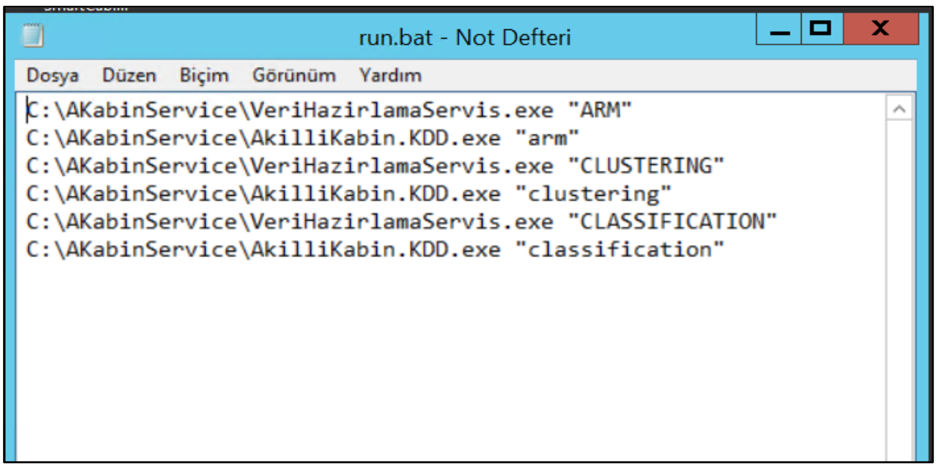


Figure 6.1 run.bat file that executes every process with a order

The whole data structure is defined as in Figure 6.2 and 6.3.

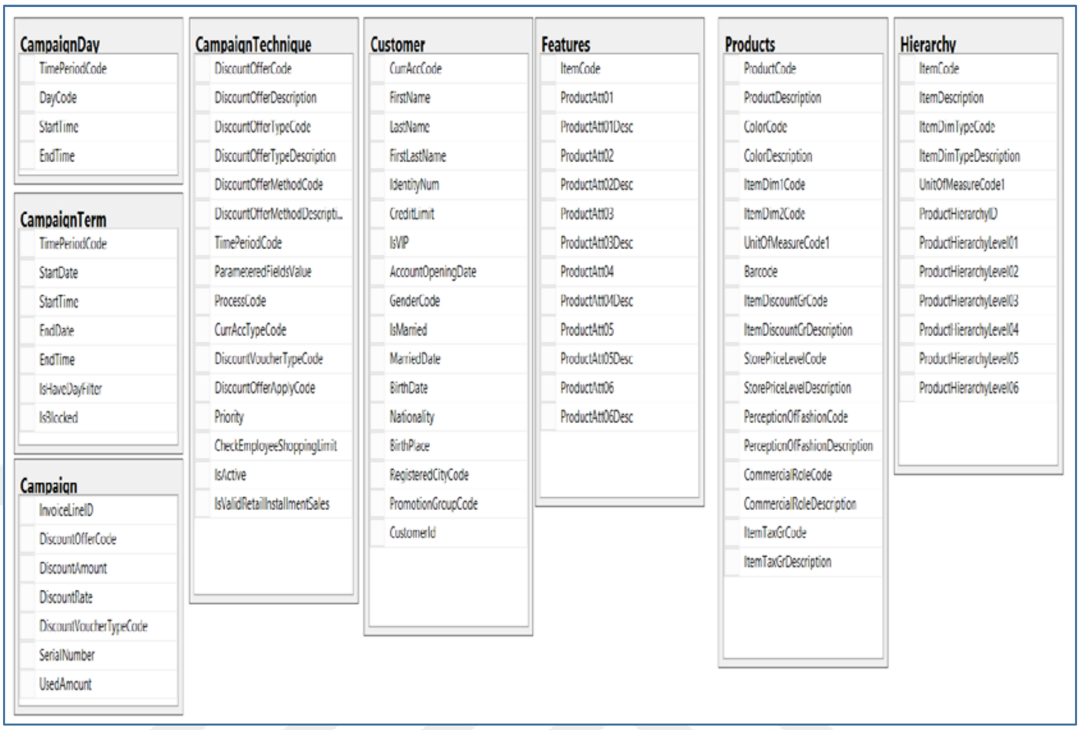


Figure 6.2 All data structure part-1

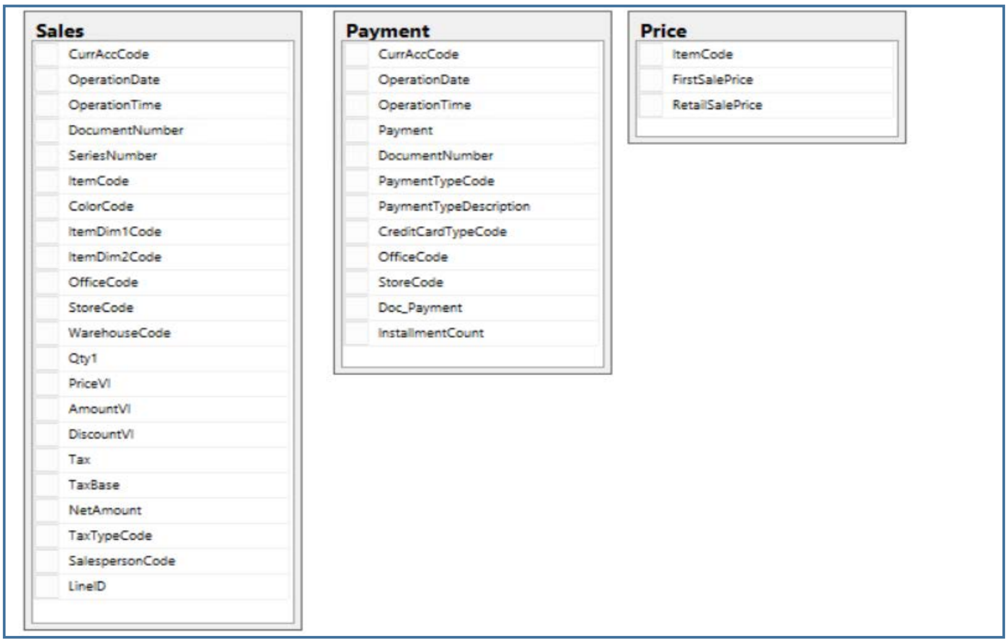


Figure 6.3 All data structure part-2

Most of the tables are the parts of ERP database, and they are fetched from multiple ERP tables into single one.

6.1.1 Data Preparation for Association Rule Mining

The first part of data preparation phase has been done for association rule mining. If the products had been represented with an itemcode which contains spesific number and string, we haven’t been able to apply a sufficient algorithm due to high diversity of items. The Features data table has been used to reduce the attributes to plausible quantity. We used 1st and 2nd parents of hierarchy and put the ColorCode between them to represent items in more generic way (For instance; 1st hierarchy-ERKEK, ColorCode-BRD, 2nd hierarchy-PANTOLON). In Figure 6.4 is the small part from the text file:

```
ERKEK-MAV-GOMLEK_CVC,ERKEK-BYZ-GOMLEK_CVC
ERKEK-STN-GOMLEK,ERKEK-YSL-GOMLEK
ERKEK-BYZ-GOMLEK_CVC,ERKEK-BYZ-GOMLEK_CVC
ERKEK-BYZ-GOMLEK_CVC,ERKEK-BYZ-GOMLEK_CVC
ERKEK-SK-PANTOLON,ERKEK-SSA-PANTOLON
ERKEK-HAK-PANTOLON_5_CEP,ERKEK-RAN-PENYE,ERKEK-RAN-PENYE
ERKEK-303-CEKET,ERKEK-303-KABAN
ERKEK-kma-GOMLEK,ERKEK-GRM-GOMLEK,ERKEK-SK-GOMLEK
ERKEK-STN-GOMLEK,ERKEK-kma-GOMLEK,ERKEK-SK-GOMLEK,ERKEK-FME-CORAP,ERKEK-FME-CORAP
ERKEK-RAN-CEKET,ERKEK-RAN-KRAVAT,ERKEK-RAN-KRAVAT,ERKEK-RAN-KRAVAT,ERKEK-RAN-KRAVAT,ERKEK-RAN-KRAVAT
ERKEK-ATR-TAKIM_ELBISE_DUZ,ERKEK-KLC-TAKIM_ELBISE
```

Figure 6.4 The input raw data for ARM

The process is observed through command line outputs like in Figure 6.5:

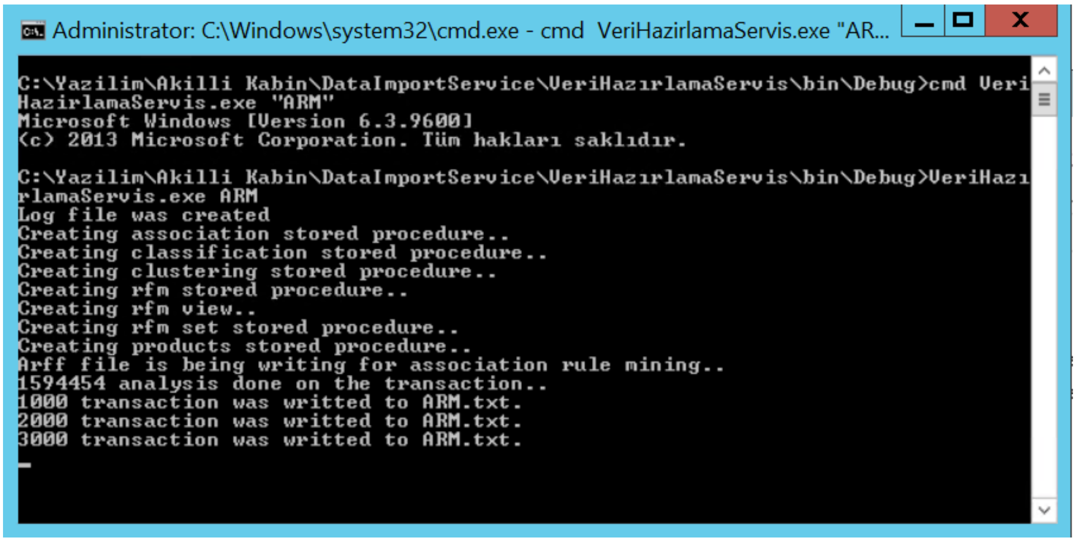


Figure 6.5 ARM data preparation

6.1.2 Data Preparation for Cluster Analysis

For clustering part, customer data has been transformed into arff files with the related purchasing history. In Figure 6.6 is the image of program during the creation of arff. Null data have been filtered. Items have been fetched from transactions in their basic hiearchy name such as “black leather belt” has taken into account as only “belt”. Then retrieved items have been selected as attributes.

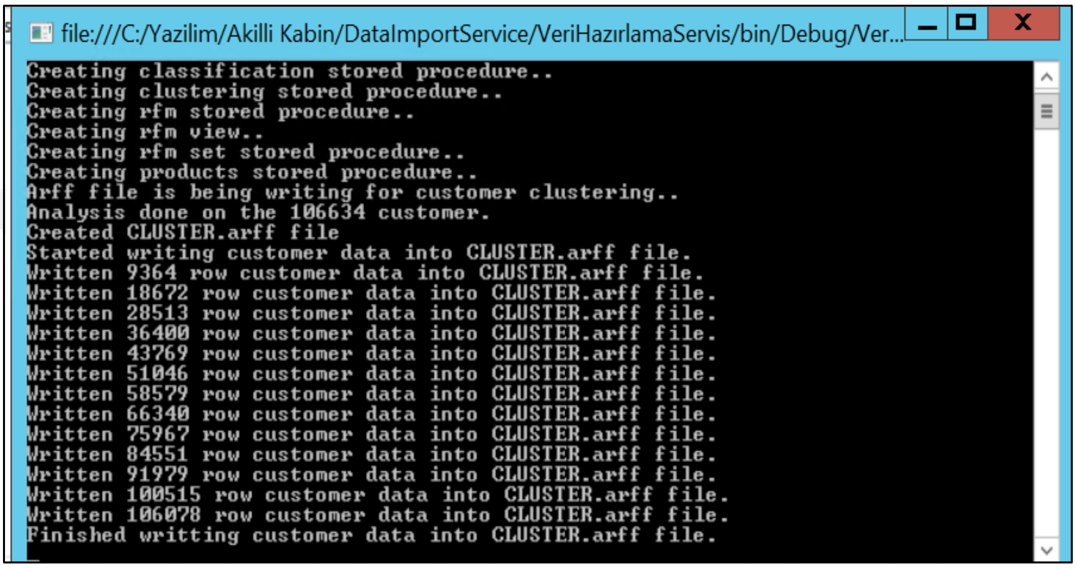


Figure 6.6 Clustering data preparation

After the process completed succesfully, the arff file for clustering has became ready. Attribute definitions and data are shown in Figure 6.7 in Figure 6.8.

```
@RELATION musteri

@ATTRIBUTE CurrAccCode real
@ATTRIBUTE GenderCode {3,1,2}
@ATTRIBUTE BirthDate {0-14,15-18,18-23,24-30,30-40,40-55,55-}
@ATTRIBUTE CityDescription {ISTANBUL,YALOVA,ARTVIN,ELAZIG,RIZE,TRABZON,ORDU,ERZINCAN,ANTALYA,BINGOL,ADANA,
@ATTRIBUTE SORT real
@ATTRIBUTE SAL real
@ATTRIBUTE KUSAK real
@ATTRIBUTE GOMLEK_ATAYAKA real
@ATTRIBUTE DERI_KEMER real
@ATTRIBUTE TRIKO&PENYE real
@ATTRIBUTE GOMLEK_SATEN real
@ATTRIBUTE CAMASIR real
@ATTRIBUTE YELEKLI_TAKIM_ELBISE_DUZ real
@ATTRIBUTE BOT real
@ATTRIBUTE BLUZ real
@ATTRIBUTE PANTOLON real
@ATTRIBUTE DERI_MONT real
@ATTRIBUTE KOL_DUGMESI real
@ATTRIBUTE YELEK-FLAR real
@ATTRIBUTE PAPYON&KUSAK real
@ATTRIBUTE ETEK real
@ATTRIBUTE AYAKKABI real
@ATTRIBUTE PENYE real
@ATTRIBUTE CIZME real
@ATTRIBUTE KARTVIZITLIK real
@ATTRIBUTE KRAVAT real
@ATTRIBUTE PIU_MONT real
@ATTRIBUTE KEMER real
@ATTRIBUTE PANTOLON_5_CEP real
@ATTRIBUTE PARFUM real
@ATTRIBUTE OZEL_DIKIM_GOMLEK real
@ATTRIBUTE OZEL_DIKIM_PANTOLON real
@ATTRIBUTE CEKET real
@ATTRIBUTE ELDIVEN real
@ATTRIBUTE PANTOLON_KLASIK real
@ATTRIBUTE OZEL_DIKIM_CEKET real
@ATTRIBUTE PAPYON real
@ATTRIBUTE SEMSIYE real
@ATTRIBUTE YELEK_KLASIK real
@ATTRIBUTE KABAN real
@ATTRIBUTE ELBISE real
@ATTRIBUTE AKSESUAR real
@ATTRIBUTE KAZAK real
@ATTRIBUTE AKS_8 real
@ATTRIBUTE MAYO real
@ATTRIBUTE DENIM_PANTOLON real
@ATTRIBUTE PALTO real
@ATTRIBUTE MANTO real
@ATTRIBUTE ATKI real
@ATTRIBUTE PANT._ASKISI real
@ATTRIBUTE TSHIRT real
@ATTRIBUTE SMOKIN_TKE real
@ATTRIBUTE GOMLEK_CVC real
@ATTRIBUTE YELEKLI_TAKIM_ELBISE real
```

Figure 6.7 Arff file definition part, ready to do clustering for Weka

```
@data

10330864,1,30-40,VAN,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10331065,1,30-40,KOCAELI,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,
10332671,2,24-30,ANTALYA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10333438,1,30-40,BURSA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10336053,1,40-55,ANKARA,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
10339484,1,55-,VAN,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10339910,1,55-,ANKARA,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0
10339963,1,30-40,ISTANBUL,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
10343791,1,24-30,ISTANBUL,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
10344261,1,18-23,BALIKESIR,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10005366,1,30-40,ISTANBUL,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
10013513,1,24-30,KAHRAMANMARAS,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10014637,1,30-40,TRABZON,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1001866,1,30-40,ISTANBUL,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,
10020709,2,30-40,ANKARA,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
10020904,1,30-40,HATAY,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,
10021475,1,30-40,GAZIANTEP,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
10023036,1,24-30,ISTANBUL,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1
10024183,2,18-23,IZMIR,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Figure 6.8 Arff file data part, ready to do clustering for Weka

6.1.3 Data Preparation for Classification

Data preparation of classification part is similar to previous one. But for handling arff file Clustering Data Mining Part should have been executed. The classification of this project is done by assigning the output of cluster info into the customer. The reason is to analyze the new customer based on demographic attributes and the products he or she brought to the fitting room and classify him/her. The details will be shown in Data Mining Part. In Figure 6.9, a small piece of command line output has been shown:

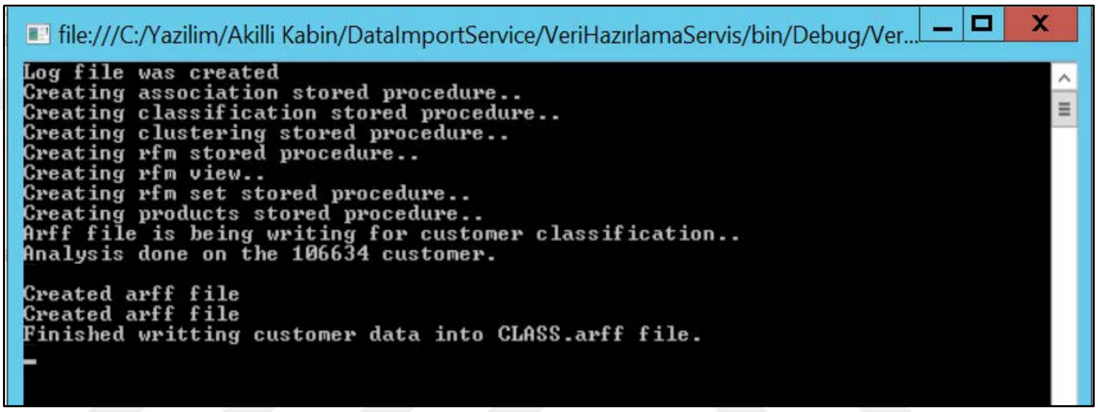


Figure 6.9 Classification data preparation

After finishing the arff file creation, CLASS.arff file looked like in Figure 6.10:

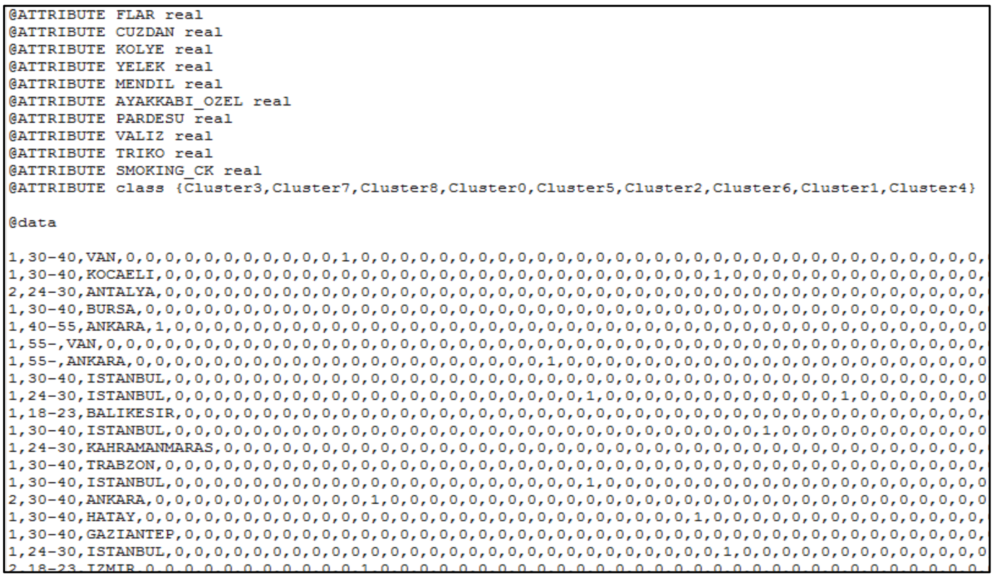


Figure 6.10 Arff file ready to classify new instance based on decision tree

6.2 Data Mining Studies in the Project

After the long preparation phase, the order of mining processes follows as ARM, clustering and classification.

6.2.1 Application of Association Rule Mining

After preparing the text file in a comma seperated file, apriori.exe is executed. The exe file, after processing all the data with the minimum support value -0.01-, writes as command line output (different minimum support values will be shown in Experimental Results chapter). In our implementation, we loop through every output and wrote it directly to the data table called KDSAssociationRules with the ratio value like in Figure 6.12. The created tables after execution has been shown in Figure 6.11.

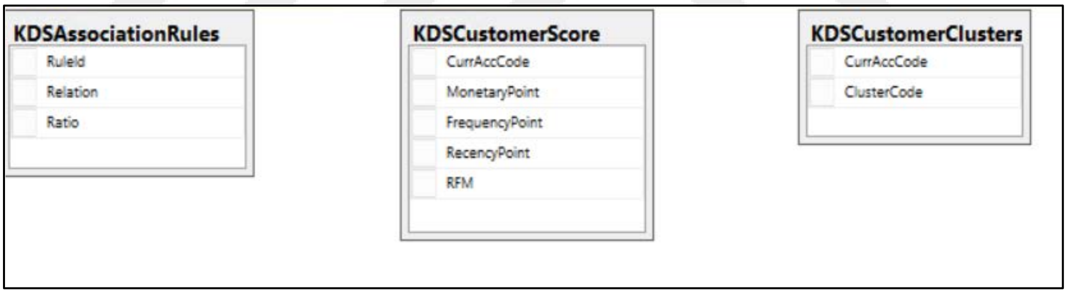


Figure 6.11 Data mining tables

```
SQLQuery7.sql - (lo...Administrator (61))   SQLQuery6.sql - (lo...Administrator (60))
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [RuleId]
      ,[Relation]
      ,[Ratio]
  FROM [AKabin].[dbo].[KDSAssociationRules]
```

100 %

Results | Messages

	RuleId	Relation	Ratio
1	1	ERKEK-RAN-GÖMLEK	102462
2	2	ERKEK-STN-PANTOLON CHINO	108476
3	3	ERKEK-STN-TRIKO	111987
4	4	ERKEK-R19-PANTOLON CHINO	109759
5	5	ERKEK-SYH-PANTOLON CHINO	113819
6	6	ERKEK-LCK-TRİKO&PENYE	105484
7	7	ERKEK-204-GÖMLEK	11504
8	8	ERKEK-KLN-TRİKO&PENYE	118277
9	9	ERKEK-422-TRIKO	115437

Figure 6.12 Example of association rules

6.2.2 Application of Clustering Algorithm

After preparing arff file, k-means clustering has been applied. There are 82 attributes in the dataset. The number *k* for k-means have been selected by human observation based on the error rate like in Figure 6.13. So after examining the outputs we decided to make 9 cluster (It will be discussed in Experimental Studies chapter).).

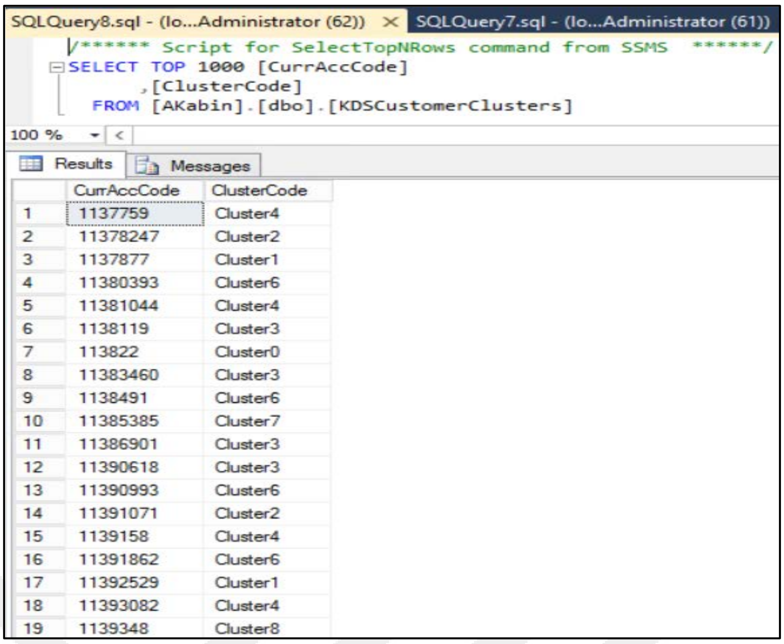
```
kMeans
======

Number of iterations: 33
Within cluster sum of squared errors: 63329.10745540852
Missing values globally replaced with mean/mode

Cluster centroids:
                                   Cluster#
Attribute               Full Data         0         1         2         3         4         5         6         7         8
                          (80703)    (8369)    (5973)    (9769)   (12533)   (14959)    (3075)    (6245)    (3812)   (15968)
==========================================================================================================================
GenderCode                      1         1         1         1         1         1         2         2         2         1
BirthDate                   30-40     30-40     30-40     40-55     24-30     40-55     24-30     30-40     40-55     30-40
CityDescription          ISTANBUL     IZMIR  ISTANBUL  ISTANBUL  ISTANBUL    ANKARA     IZMIR    ANKARA     IZMIR   ANTALYA
GOMLEK_ATAYAKA             0.0345    0.0174    0.1061    0.0229      0.08    0.0024    0.1772    0.0102    0.0121    0.0055
KUSAK                      0.0023    0.0008    0.0069    0.0003    0.0049    0.0003    0.0176    0.0014    0.0018         0
GOMLEK_SATEN               0.0425    0.0356    0.0584    0.0509    0.0503    0.0315    0.0387    0.0263    0.0283    0.0498
YELEKLI_TAKIM_ELBISE_DUZ   0.0045     0.002    0.0028    0.0036    0.0108    0.0031    0.0039    0.0026    0.0005     0.005
BLUZ                       0.0142    0.0109     0.004    0.0039    0.0032    0.0073    0.0205    0.0397    0.1306     0.002
YELEK-FLAR                 0.0023    0.0011    0.0097     0.001    0.0053    0.0001    0.0072    0.0008    0.0008    0.0009
ETEK                       0.0027    0.0022    0.0003    0.0014    0.0002    0.0008    0.0055    0.0094    0.0231    0.0005
CIZME                      0.0002         0         0         0    0.0001    0.0001         0    0.0003    0.0018    0.0001
```

Figure 6.13 The cluster output of Weka

After setting the cluster number, final results have been inserted to data table called KDSClusters like in Figure 6.14:



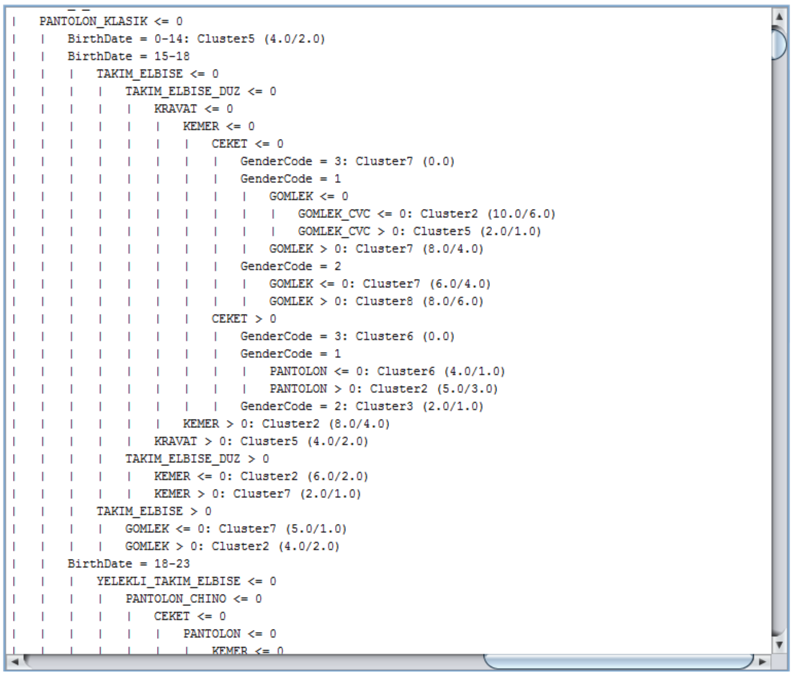
```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [CurrAccCode]
      ,[ClusterCode]
  FROM [AKabin].[dbo].[KDSCustomerClusters]
```

	CurrAccCode	ClusterCode
1	1137759	Cluster4
2	11378247	Cluster2
3	1137877	Cluster1
4	11380393	Cluster6
5	11381044	Cluster4
6	1138119	Cluster3
7	113822	Cluster0
8	11383460	Cluster3
9	1138491	Cluster6
10	11385385	Cluster7
11	11386901	Cluster3
12	11390618	Cluster3
13	11390993	Cluster6
14	11391071	Cluster2
15	1139158	Cluster4
16	11391862	Cluster6
17	11392529	Cluster1
18	11393082	Cluster4
19	1139348	Cluster8

Figure 6.14 Final output of clustering

6.2.3 Application of Classification Techniques

After preparing our arff file based on the output of clustering which is shown in Figure 6.10, we proceed with J48 algorithm using Weka library. After execution of the process, a part of the constructed decision tree is shown in Figure 6.15.



```
|   PANTOLON_KLASIK <= 0
|   |   BirthDate = 0-14: Cluster5 (4.0/2.0)
|   |   BirthDate = 15-18
|   |   |   TAKIM_ELBISE <= 0
|   |   |   |   TAKIM_ELBISE_DUZ <= 0
|   |   |   |   |   KRAVAT <= 0
|   |   |   |   |   |   KEMER <= 0
|   |   |   |   |   |   |   CEKET <= 0
|   |   |   |   |   |   |   |   GenderCode = 3: Cluster7 (0.0)
|   |   |   |   |   |   |   |   GenderCode = 1
|   |   |   |   |   |   |   |   |   GOMLEK <= 0
|   |   |   |   |   |   |   |   |   |   GOMLEK_CVC <= 0: Cluster2 (10.0/6.0)
|   |   |   |   |   |   |   |   |   |   GOMLEK_CVC > 0: Cluster5 (2.0/1.0)
|   |   |   |   |   |   |   |   |   GOMLEK > 0: Cluster7 (8.0/4.0)
|   |   |   |   |   |   |   |   GenderCode = 2
|   |   |   |   |   |   |   |   |   GOMLEK <= 0: Cluster7 (6.0/4.0)
|   |   |   |   |   |   |   |   |   GOMLEK > 0: Cluster8 (8.0/6.0)
|   |   |   |   |   |   |   CEKET > 0
|   |   |   |   |   |   |   |   GenderCode = 3: Cluster6 (0.0)
|   |   |   |   |   |   |   |   GenderCode = 1
|   |   |   |   |   |   |   |   |   PANTOLON <= 0: Cluster6 (4.0/1.0)
|   |   |   |   |   |   |   |   |   PANTOLON > 0: Cluster2 (5.0/3.0)
|   |   |   |   |   |   |   |   GenderCode = 2: Cluster3 (2.0/1.0)
|   |   |   |   |   |   KEMER > 0: Cluster2 (8.0/4.0)
|   |   |   |   |   KRAVAT > 0: Cluster5 (4.0/2.0)
|   |   |   |   TAKIM_ELBISE_DUZ > 0
|   |   |   |   |   KEMER <= 0: Cluster2 (6.0/2.0)
|   |   |   |   |   KEMER > 0: Cluster7 (2.0/1.0)
|   |   |   TAKIM_ELBISE > 0
|   |   |   |   GOMLEK <= 0: Cluster7 (5.0/1.0)
|   |   |   |   GOMLEK > 0: Cluster2 (4.0/2.0)
|   |   BirthDate = 18-23
|   |   |   YELEKLI_TAKIM_ELBISE <= 0
|   |   |   |   PANTOLON_CHINO <= 0
|   |   |   |   |   CEKET <= 0
|   |   |   |   |   |   PANTOLON <= 0
|   |   |   |   |   |   |   KEMER <= 0
```

Figure 6.15 Output J48

Since classification requests will be based on demand, this rule tree should be used again. After keeping this for our classifier, we are ready to publish our web service.

6.3 Web Service

Web Service Implementation has been done by using .NET WEB API. Since clustering is intermediary step for classification, there is no UI interaction that would trigger clustering. ARM and classification part has been proposed as Web services. In Figure 6.16, the spesific url requests with the parameters which are special representations of items will retrieve the itemcode and color data of corresponding item.

Figure 6.16 Web service for ARM

In Figure 6.17, the customer ID has been sent by query string. The workflow goes as follows; if customer doesn’t belong to any predefined clusters, classify the customer based on the items which he/she brought to cabin by using decision tree. If customer exists in clusters, then, compare with some other customer who exists in the same cluster and process the attributes of that other customer, extract top counts of items and transform to itemcode in order to represent on UI.

Figure 6.17 Web service for classification

Using Nebim API and UI, all parts have been integrated. The final look of the project is shown in Figure 6.18.

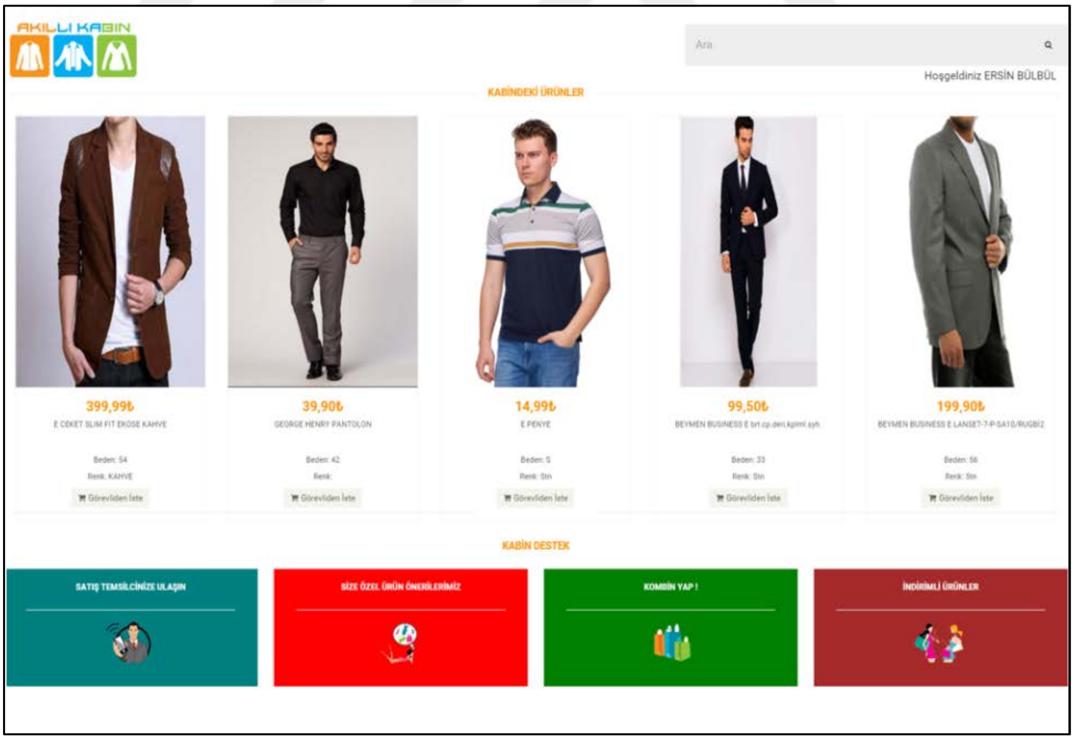


Figure 6.18 Final outlook of project

CHAPTER SEVEN
EXPERIMENTAL STUDIES

7.1 ARM Minimum Support Selection

Different minimum support values have been tested against apriori process. Since output will be proposed as item-based recommendations to customer, we had to guarantee, we have enough numbers of recommendations. Different minimum support values have been tested. Default confidence value has been chosen as 80%.

Table 7.1 Comparison of different minimum support values

Minimum Support / N frequent -itemsets	S=10%	S=5%	S=2%	S=1%	S=0.5%	S=0.1%
2	0	1	12	31	112	1001
3	0	0	0	5	17	438
4	0	0	0	0	1	88
5	0	0	0	0	0	2
6	0	0	0	0	0	0

As it is shown in Table 7.1, different numbers of rules have been extracted. In our study, minimum support value has been selected as 1%. Since these rules represent items' top hierarchy names with colors (like ERKEK-BEYAZ-GÖMLEK), even one rule can produce dozens of products due to decapsulation process from hierarchy names to real products. Thus, total 36 association rules would be enough for our work.

7.2 K-means Cluster Analysis

Different k values have been tested in order to find out the best k value for clustering. During the process, sum of squared errors has been taken into account as estimator value.

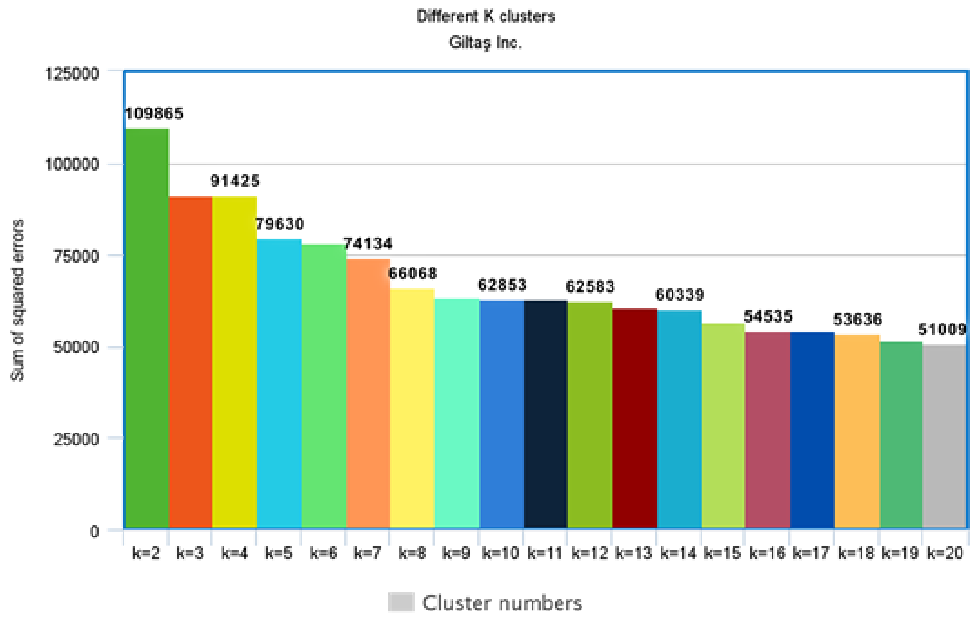


Figure 7.1 Cluster Analysis

As you can see in the histogram graph (Figure 7.1), there are couple of critical points that makes drastic drop in sum of squared errors. As we observed, after k value 9, drops become smoother for each iteration. Hence, k=9 has been chosen as centroid numbers for the clustering algorithm.

In order to get better results, attributes have been prefiltered for clustering and classification. It has downsized to 16 attributes. Demographic properties have been removed and customers have been populated to 300,000 instances since customers who have null values on removed attributes, have been taken into account again. Different k values and sum of squared errors have been showed in Figure 7.2. K=12 has been selected after the observations.

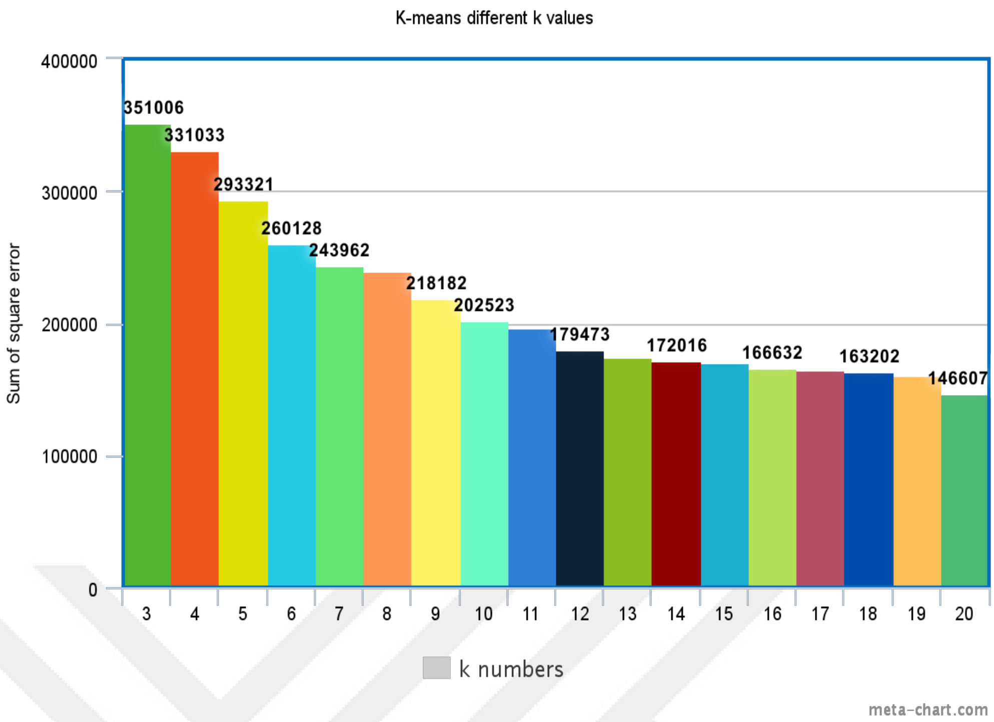


Figure 7.2 Prefiltered 16-attribute cluster analysis

7.3 Evaluation of Classification Results

Different approaches have been followed during the classification process. 10-folds cross validation and 66% split methods have been chosen as testing options to observe J48 algorithm results. 16 GB RAM, i7 5500U 2.40 GHz CPU computer has been used for mining operations. Prefiltered dataset has been used for experimental results. For 10-folds cross validation, 57 seconds spent for building model and 12 minutes spent to test 10 folds (Figure 7.3).

```
=== Summary ===

Correctly Classified Instances      107329              34.2583 %
Incorrectly Classified Instances    205964              65.7417 %
Kappa statistic                          0
Mean absolute error                      0.1363
Root mean squared error                  0.2611
Relative absolute error                 99.9995 %
Root relative squared error            100      %
Total Number of Instances           313293

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,013     Cluster0
                 1,000    1,000    0,343      1,000    0,510      0,000    0,500     0,343     Cluster1
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,071     Cluster2
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,093     Cluster3
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,196     Cluster4
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,068     Cluster5
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,035     Cluster6
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,044     Cluster7
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,043     Cluster8
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,033     Cluster9
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,034     Cluster10
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,029     Cluster11
Weighted Avg.    0,343    0,343    0,117      0,343    0,175      0,000    0,500     0,182
```

Figure 7.3 Ten cross-validation J48 algorithm sesults

For 66% split method, approximately 2 minutes spent to test the data. Model construction is the same (Figure 7.4).

```
=== Summary ===

Correctly Classified Instances       36431              34.2011 %
Incorrectly Classified Instances     70089              65.7989 %
Kappa statistic                          0
Mean absolute error                      0.1363
Root mean squared error                  0.2611
Relative absolute error                 99.9993 %
Root relative squared error            100      %
Total Number of Instances           106520

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,013     Cluster0
                 1,000    1,000    0,342      1,000    0,510      0,000    0,500     0,342     Cluster1
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,069     Cluster2
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,094     Cluster3
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,197     Cluster4
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,067     Cluster5
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,035     Cluster6
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,045     Cluster7
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,043     Cluster8
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,033     Cluster9
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,033     Cluster10
                 0,000    0,000    0,000      0,000    0,000      0,000    0,500     0,029     Cluster11
Weighted Avg.    0,342    0,342    0,117      0,342    0,174      0,000    0,500     0,182
```

Figure 7.4 Split 66% J48 algorithm results

As an addition, other well-known classification methods have been tried. The results are shown in the Table 7.2.

Table 7.2 Comparison of classification methods based on success ratio.

Method Name	Classification Accuracy %
BayesNet	34.26
NaiveBayes	34.20
Multilayer Perceptron	34.18
Random Forest	33.80

7.4 Experimental Results

Different approaches have been followed for clustering and classification. As a result, customer-based basket analysis has been done by using clustering and classification as an alternative to recommender systems. The success ratio has been obtained around 34.20% approximately.

Since clustering categorical data with overlap measure (if same attributes have same values, 1, otherwise 0) is susceptible to overextension of clusters, that could affect classification phase. As cluster1 has one-third of instances, classification methods didn't produce good results due to overextension (Figure 7.5).

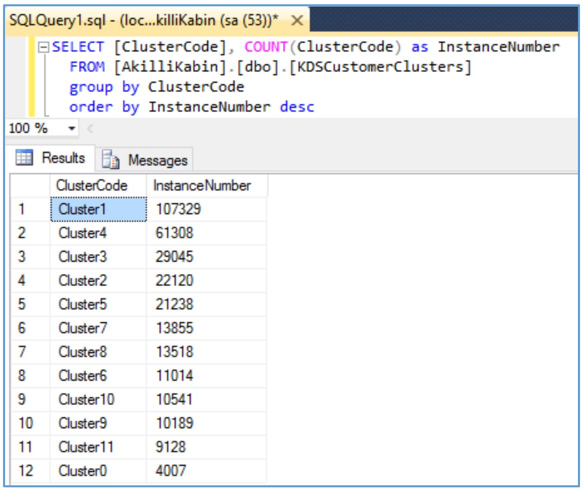


Figure 7.5 Clusters' instance numbers

CHAPTER EIGHT
CONCLUSION AND FUTURE WORK

The study has shown the functionalities and components of a smart fitting room on ERP systems and the integration with data mining methods. In this project, as a communication between real objects in a store with ERP data records, a transition database is established in order to map EPC numbers with product IDs which are barcodes. The security has been implemented in the store with the help Nebim Integrator. EPC numbers of the purchased items have been stored as a record in invoice line which represents each item sold. Thereby the deficieny due to barcodes redundancy has been removed and this made possible to make impenetrable security gates on exit based on RFID with the legacy systems.

An alternative approach has been followed to the recommender systems, due to cold start and data sparsity problems. The transition database is also used for data mining purposes. First a bulk insert operation has been done for the initial phase of clustering, classification and ARM (Association Rule Mining). After transferring the data by filtering the customers and products data that have null values in significant columns, ARM has been applied and results have been recorded in the database as a set of rules. Afterwards attribute selection has been done for clustering. Each attribute has represented the simple name (like pants, shirts and so on) of the items. The results of clustering process have been used to create a decision tree by using J48 algorithm. Corresponding cluster numbers have became the target class for J48 algorithm. After setting the decision tree, desired customer has been classified and top attributes which stands for products have been selected in the same cluster and given back to the web service as a result.

In this study, there are also few drawbacks has to be examined again. By using ARM we have discarded the insignificant sales but that might have resulted with no suggestions with the rare item that is sold in really few transactions.

Clustering and classification with 16 attributes (initially 82) gave us a bit inconsistent results. The reason lies behind the clustering phase. ROCK or LIMBO like clustering methods could have been used.

As a future work, express lane integration with devices such as Ingelico, Hugin should be done to observe and test how the security works even it means more cost to get permission for your software. Also when we were discussing with the owner of the company, he told the main idea was to accomplish everything in a cabin. Even though that may result long waiting queues could be an innovative approach.

Additionally the implementation of security provides a platform free purchasing oppurtinity. What it means that, the entire purchasing process can be moved to the mobile platform, end users can buy what they see by phone application and security doors wont be alarmed since the RFID info has been stored in invoice line.

REFERENCES

Bjork, C. (2014). Zara builds its business around RFID. *Wall Street Journal*, B1-B2.

Buttar, H., & Kaur, R. (2013). Association technique in data mining and its applications. *International Journal of Computer Trends and Technology (IJCTT)-4.* (4)April 2013.

Chang, N., Irvan, M., & Terano, T. (2013). A TV program recommender framework. *Procedia Computer Science, 22,* 561-570.

Cole, S. A. (2009). *Suspect identities: A history of fingerprinting and criminal identification*. Harvard University Press.

Di Marco, P., Santucci, F., & Fischione, C. (2014, June). *Modeling anti-collision protocols for RFID Systems with multiple access interference*. In 2014 IEEE International Conference on Communications (ICC) (pp. 5938-5944). IEEE.

Domdouzis, K., Kumar, B., & Anumba, C. (2007). Radio-Frequency Identification (RFID) applications: A brief introduction. *Advanced Engineering Informatics, 21*(4), 350-355.

Ebrahimzadeh, A., Addeh, J., & Rahmani, Z. (2012). Control chart pattern recognition using K-MICA clustering and neural networks. *ISA Transactions*,*51*(1), 111-119.

Finkenzeller, K. (2003). *RFID Handbook: Fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication*,(3rd ed.). John Wiley and Sons.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399-409.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Siam.

Giraud-Carrier, C., & Povel, O. (2003). Characterising data mining software. *Intelligent Data Analysis*, *7*(3), 181-192.

Guo, G., Zhang, J., & Thalmann, D. (2014). Merging trust in collaborative filtering to alleviate data sparsity and cold start. *Knowledge-Based Systems*,*57*, 57-68.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques.* Elsevier.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 263-272.

Kamath, R. S., & Kamat, R. K. (2016). *Educational data mining with R and Rattle*. River Publishers.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press.

Kriegel, H. P., & Pfeifle, M. (2005, August). Density-based clustering of uncertain data. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (pp. 672-677).

Kumar, V. (2010). *Chapman & Hall/CRC data mining and knowledge discovery series.* Taylor & Francis Group.

Laheurte, J. M., Ripoll, C., Paret, D., & Loussert, C. (2014). *UHF RFID technologies for identification and traceability*. John Wiley & Sons.

Liwan, S. R. (2015). *The framework of improving on-site materials tracking for inventory management process in construction projects* Doctoral dissertation, Universiti Tun Hussein Onn Malaysia.

Lopez, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. *International Educational Data Mining Society*.

Moon, K. L., & Ngai, E. W. T. (2008). The adoption of RFID in fashion retailing: a business value-added framework. *Industrial Management & Data Systems, 108*(5), 596-612.

Müller, J. (2013). *A real-time in-memory discovery service. Leveraging hierarchical packaging information in a unique identifier network to retrieve track and trace information*. Springer.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592-2602.

Nguyen, H. H., Harbi, N., & Darmont, J. (2012). *An efficient clustering-based classification approach for intrusion detection.*

Peppers, D., Rogers, M., & Dorf, B. (1999). Is your company ready for one-to-one marketing. *Harvard Business Review, 77*(1), 151-160.

Peris-Lopez, P., Hernandez-Castro, J. C., Estevez-Tapiador, J. M., & Ribagorda, A. (2016). Lightweight cryptography for low-cost RFID tags. *Security in RFID and Sensor Networks*, 121-150.

Rankl, W., & Effing, W. (2004). *Smart card handbook*. John Wiley & Sons.

Roberts, C. M. (2006). Radio frequency identification (RFID). *Computers & Security*, *25*(1), 18-26.

Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, *24*(4), 483-502.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*(1), 127-137.

Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical Systems and Signal processing*, *21*(2), 930-942.

Tanna, P., & Ghodasara, Y. (2014). Using apriori with WEKA for frequent pattern mining. A*rXiv Preprint ArXiv:1406.7371*.

Tudorache, I. C., & Vija, R. I. (2015). Data mining and customer relationship management for clients segmentation. *International Journal of Economic Practices and Theories*, *5*(5), 571-578.

Wilson, M. (2014) *Microsoft's smart fitting room is like a robo-shop clerk*. Retrieved July 15, 2016 from https://www.fastcodesign.com/3031689/microsofts-smart-fitting-room-is-like-a-robo-shop-clerk

Wong, C., & Guo, Z. X. (Eds.). (2014). *Fashion supply chain management using radio frequency identification (RFID) technologies*. Elsevier.

Zacharski, R. (2015). *A programmer's guide to data mining: THE ancient art of the numerati.* Retrieved July 15 2016 from, http://guidetodatamining.com/assets/guideChapters/DataMining-ch1.pdf