İSTANBUL COMMERCE
UNIVERSITY

GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCES

ASSOCIATION RULES AND MARKET BASKET ANALYSIS: A CASE STUDY IN RETAIL SECTOR

Pınar YAZGAN

Assist. Prof. Dr.
Ali Osman KUSAKCI

M. SC. THESIS
DEPARTMENT OF INDUSTRIAL ENGINEERING
ISTANBUL – 2016

ACCEPTANCE AND APPROVAL PAGE

The thesis with the title "**Association Rules And Market Basket Analysis: A Case Study In Retail Sector "** which is written and presented successfully by **Pınar Yazgan** is approved as a thesis of **"M.Sc. "** by the committee in Institute of Natural Science **Industrial Engineering** on 06/06/2016.

Supervisor	**Assist. Prof. Dr. Ali Osman KUSAKCI** Istanbul Commerce University	
Member of Committee	**Assist. Prof. Dr. Berk AYVAZ** Istanbul Commerce University	
Member of Committee	**Assist. Prof. Dr. Nezir AYDIN** Yıldız Technical University	

Date of Approve : 06/06/2016

Prof. Dr. Doğan KAYA
Director of Nature Science Institute

DECLARATION OF CONFORMITY
ACADEMIC AND ETHICS

I declare in this work which I prepared in accordance with the rules of writing thesis of Institue of Natural Science that

- I obtained all information and documents in this thesis within the framework of academic rules,
- I provide all the visual, audio and written information and results according to scientific ethics,
- in the case of making use of others' works, I referred to those related artifacts in accordance with scientific standards,
- I referred to the all works which I used as a source,
- I did not make any alterations in the used data,
- And I did not present any section of this thesis as another thesis work neither at this university nor at any other university.

06.06.2016

Pınar YAZGAN

CONTENTS

ÖZET

Yüksek Lisans Tezi

ASSOCIATION RULES AND MARKET BASKET ANALYSIS: A CASE STUDY IN RETAIL SECTOR

Pınar YAZGAN

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Yrd. Doç. Dr. Ali Osman KUŞAKCI
2016, 76 sayfa

Hızla gelişen teknoloji sayesinde marketler ve işletmeler verilerini kolayca saklayabilmektedirler. Gerçekleştirilen her işlem depolanarak veri setlerini oluşturmaktadır. Gittikçe büyüyen bu veri setlerinden yararlı bilgiler elde edilmesi gerekmektedir. İşte bu aşamada veri madenciliği devreye girmektedir.

Bu çalışmada, öncelikle veri madenciliğinin temelleri, aşamaları, kullanım alanları ve temel algoritma çeşitlerinden bahsedilmiştir.

Daha sonra Veri Madenciliği modellerinden olan "Birliktelik Kuralları" algoritmaları üzerinde durulmuş, bu algoritmalar arasında bir değerlendirme yapılarak Apriori algoritması tercih edilmiştir.

Son bölümde Türkiye'deki bir perakende satış mağazasının verileri ile Apriori algoritması kullanılarak Birliktelik Kuralı Analizi uygulanmış ve ürünler arasındaki ilişkiler ortaya çıkarılmıştır.

Anahtar Kelimeler: Birliktelik kuralı analizi, market sepeti analizi, veri madenciliği.

ABSTRACT

M.Sc. Thesis

ASSOCIATION RULES AND MARKET BASKET ANALYSIS: A CASE STUDY IN RETAIL SECTOR

Pınar YAZGAN

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Industrial Engineering

Supervisor: Assist. Prof. Dr. Ali Osman KUSAKÇI
2015, 76 pages

Thanks to the rapidly developing technology, companies and businesses can easily store their data. Storage of each transaction performed forms data sets. Useful informations has to be obtained from the steadily growing data sets. At this stage, data mining is of paramount importance.

Firstly, in this study, the basics of data mining, its stages, application areas and types of basic algorithms are discussed.

Secondly, a comprehensive review of "Association Rules" algorithms, as one of the main tools of data mining, is presented. Considering the strenghts and weaknesses of the presented algorithms, Apriori algorithm is preferred for application.

Lastly, association rule analysis is applied using Apriori Algorithm on data of one of the retail chains in Turkey and relations between products are revealed. Furthermore, the implicaitons of the analysis are discussed in detail.

Keywords: Association rule mining, data mining, market basket analysis, association rules.

THANKS

I owe endless thanks to my supervisor Asst. Prof. Ali Osman Kusakci for his leading in my research and his invaluable contributions with his knowledge and experience to help me to overcome the difficulties that I encountered.

Also I offer my deepest love and respect to my mother Muzeyyen Yazgan who were always with me during all stages of my thesis.

Pınar Yazgan

İSTANBUL, 2016

FIGURES

TABLES

SYMBOLS AND ABBREVIATIONS

Ct	Import in the period t
Yt	Income of period t
Ut	The margin of error for the period
[P(H/X)]	Posterior Probability
[P(H)]	Prior Probability
MBA	Market basket analysis
ARM	Association rule mining

1. INTRODUCTION

1.1 The Subject And Scope

Markets, companies and other organizations with the help of rapidly evolving information technology (IT) in recent years, have the opportunity to store huge data. Data series are constructed by keeping record of all operations within the institution. Evaluation of day by day growing datasets and extracting useful information from the data, is of paramount importance.

Today, IT allows to obtain and keep a huge amount of data. However, the bulk data comes along a difficult task: analyzing this huge data and obtaining the correct information. Obviously, making an analysis by observing the tables with thousands of lines and columns and making useful conclusions is impossible. This necessitates to employ computer technologies. Finding patterns, trends and abnormalities and summarizing as simple models in datasets is one of the most important issue in the information age (Gancheva, 2013).

Data mining is discovery of relations and rules which is significant, potentially useful and making predictions about future through large amounts of available data using computer programs (Bükey, 2014).

One of the data mining application areas which is increasingly widespread in use in many sectors is Market Basket Analysis (MBA). In MBA, relationship and the rules are obtained taking advantage of the customer, product and sales information in a retail store. MBA, obtaining of products' sales relationship with another product and finding out the association rules, increases the profits of companies.

Association rules provide the generation of future predictions discovering objects acting together in a sales transaction data and the relationship between objects.

To achieve these rules, since the beginning of the 90s many algorithms have been developed. There are advantages relative to each other under different circumstances and different working methods of these algorithms.
Implementation of the merge, pruning methods and scanning database and revealing the association relationship between objects with help of minimum support price represent general logic of algorithms (Gürgen, 2008).

In this thesis, concepts related to data mining and especially basic algorithms employed to generate association rules in the market basket analysis have been considered in details. In addition, an application MBA is conducted with Apriori algorithm to identify the association rules between items sold in a retail store. Furthermore, the obtained rules are discussed and further implications are highlighted. In application; receipts were collected during 6 months from the supermarket and the relationship between the items in these receipts has been found with using Apriori algorithm.

1.2 Purpose And Importance

Big markets, businesses and other organizations have been collecting various types of data based on their purposes and structures in parallel with the development of information systems and technologies.

There is a need to discover meaningful and efficient templates and rules from stored and distributed data of large volumes in many areas such as shopping industry, banking transactions, governmental organizations as well as many others. Data mining finds out unclear, previously unknown but useful knowledge from available bulk data.

Altough the presence of association rules in a database is obvious from data mining perspective, extracting them is a highly difficult task. Once extracted, it allows revealing and summarizing various relationships. For example; determination of consumer tendencies to buy what type of products and services increases the sales, which results in increase of company’s profit.

Association rules and sequential patterns which allow identification of purchasing trends are frequently used under the name of Market Basket Analysis (MBA) in data mining for marketing purposes.
In addition to the MBA, these techniques are preferred in various areas, such as medicine, engineering, economics, and finance, where it is necessary to determine the subtle relationships hidden in entries of the dataset.

Association rules, discover the relationship between objects and make estimates for the future in sales of the considered items (Sherdiwala and Khanna, 2015).

Accordingly; the aim of the thesis is to investigate the basic concepts, methods and techniques within knowledge discovery process in data mining on databases, examination and comparison of the discovery process of association rules and the algorithms used for detection of these rules in MBA. As a result, associations between items may be used for increasing sales and revenue.

2. LITERATURE REVIEW

This section presents a literature review on different techniques for ARM with a special attention on MBA applications.

ARM algorithms can be classified into three main classes: (1) Frequent itemset mining, (2) Sequential pattern mining, (3) Structured pattern mining. Figure 2.1. shows various association rule mining algorithms with sub branches developed since the first introduction of ARM algorithms.

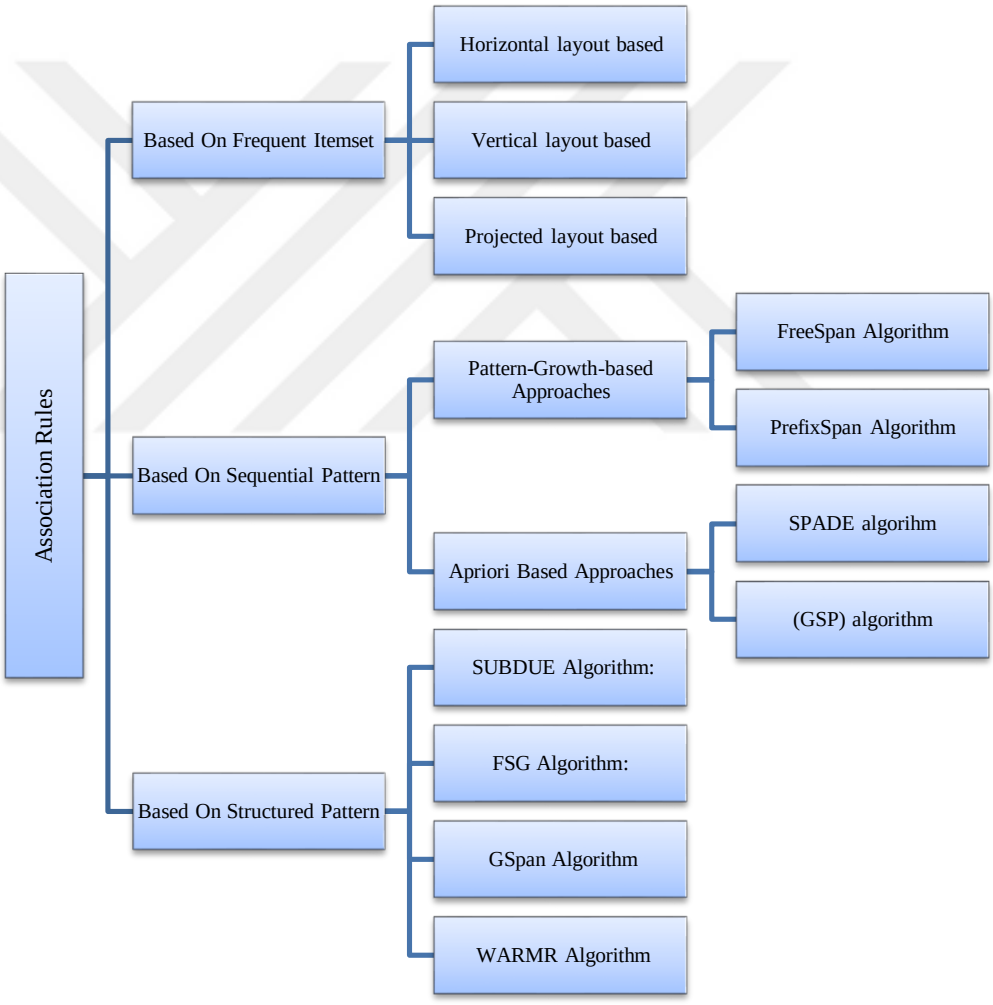

Figure 2.1 Association rule mining algorithms.

2.1 Frequent Itemset Mining

Today, frequent itemset mining is one of the most important tools employed on transactional database and has a major role in many data mining tasks to discover patterns such as classifiers, correlations, clusters, association rules, sequences. It aims to optimize the process of finding patterns more efficiently.

A frequent itemset is a pattern that occurs greater than the minimum support in the database. Some strategies are used to generate frequent itemsets are:

● Reduce the number of candidates (M) with using of pruning techniques to decrease M.

● Decrease the number of transactions (N) with using direct hashing and pruning (DHP) and vertical-based mining algorithms.

● Decrease the number of comparisons (NM) with using efficient data structures for storing transactions or candidates.

Possible applications of algorithms based on frequent item sets approach are:

● Develop arrangement of products on a catalog's pages, in shelves etc.

● Product bundling, support cross-selling.

● Technical dependence analysis, fraud detection etc. (Borgelt, 2012).

We can illustrate the main steps of finding frequent itemsets on a transaction database through the following example. Assume we have a list of 10 transaction and frequent itemsets given in Table 2.1 and Table 2.2.

Transaction Database
1:{milk, diaper, bread}
2:{beer, coke, diaper}
3:{milk, coke, bread}
4:{ milk, coke, diaper, bread}
5:{milk, bread}
6:{milk, coke, diaper}
7:{beer, coke}
8:{milk, coke, diaper, bread}
9:{beer, coke, bread}
10:{milk, diaper, bread}

Table 2.1 Illustratıve Transaction Databas

0 item	1 item	2 items	3 items
0: 10	{milk}:7 {beer}:3 {coke}:7 {diaper}:6 {bread}:7	{milk,coke}: 4 {milk,diaper}: 5 {milk,bread}: 6 {beer,coke}: 3 {coke,diaper}: 4 {coke,bread}: 4 {diaper,bread}: 4	{milk,coke,diaper}: 3 {milk,coke,bread}: 3 {milk,diaper,bread}: 4

Table 2.2 Frequent Itemsets

The minimum support is selected as s_{min} = 3 or σ_{min} = 0.3 = %30. According to Table 2.2 following can be concluded;

● There are 2^5 = 32 possible item sets over this 5 different products given in these 10 transactions. B={milk, beer, coke, diaper, bread}

● For the specified minimum support level, there are 16 frequent item sets but only 10 transactions.

Many algorithms for mining frequent itemsets have been introduced over the years like horizontal layout based algorithms, vertical layout based algorithms and projected layout based algorithms. These algorithms are shown in Table 2.5

2.1.1 Algorithms for mining from horizontal layout database

In this database, each row of database represents a transaction which has a transaction identifier (TID).

One example of horizontal layout dataset is shown in table 2.3(Gupta and Garg, 2011).

TID	ITEMS
T1	I1, I2, I3, I5, I7
T2	I1, I5, I6, I7
T3	I1, I3, I4, I6, I7
T4	I2, I4, I5
T5	I6, I7

Table 2.3 Example of horizontal layout database

Many techniques have been proposed to mine frequent patterns from horizontal data format such as Apriori Algorithm, Direct hashing and pruning (DHP), Partitioning algorithm, Sampling algorithm, Dynamic Itemset Counting (DIC), Continuous Association Rule Mining Algorithm (CARMA).

2.1.1.1 Apriori algorithm

Apriori algorithm has been developed by Agarwal and Srikant in 1994. It is one of the most famous of all ARM algorithms. Apriori is designed to work on databases including transactions and can be used to produce all frequent itemsets.

Along with Apriori, AprioriTid and AprioriHybrid algorithms have been suggested by Agrawal and Ramakrishnan in 1994. AprioriTid is executed equivalently well as Apriori in small problems, but when implemented to big-scale problems performance of the algorithm decreases (Agrawal and Srikant, 1994). On the other hand, AprioriHybrid performs better than Apriori in almost all cases.

Many improvements have been made on the Apriori algorithm in order to increase its effectiveness and efficiency (Zaki et al., 1996). Apriori algorithm is not sufficient for redundancy and candidate set generation. However, it formed the basis for many algorithms.

Ji et al. in 2006 presented a development on the existing Apriori algorithm. Unwanted candidate set generations are the disadvantage of existing Apriori algorithm. Modifying the pruning technique decreased the candidate set generation process in improved Apriori algorithm. A separate set is developed with less frequent items. This improvement has notable advantages over the standart Apriori Algorithm (Ji et al., 2006).

Huang in 2007 has changed Apriori Algorithm without large candidate set generation (Huang, 2007). Wu et al. in 2009 proposed Improved Apriori Algorithm (IAA) to decrease the number of scans on the data and redundant operations while frequent itemsets and association rules are produced.

This algorithm has a new count-based method for pruning process. In this algorithm, all the frequent itemsets are found in given data sets using genetic algorithm (Wu et al, 2009).

Along with the rapid increase in log data, there was a need for handling logging data. Shao et al. in 2009 suggested 3D-Apriori algorithm.
3D-Apriori algorithm has main features as attribute data discretion and spatial predicate extraction for generation of association rule. 3D-Apriori interprets the logging data and enhances efficiency of association rules behind the logging data transformation (Shao, 2009).

Sharma and Sufyan in 2012 performed a probabilistic analysis of Apriori algorithm to discover frequent itemsets and association rules in a transaction database. It contains a single database scan and limit unsuccessful candidate sets. The concept of recursive medians is used in the algorithm as an Inverted V-Median Search Tree (IVMST). The recursive medians compute the dispersion of each itemsets in the transaction list and the maximum number of common transactions for any two itemsets. Using the above mentined procedures, they presented a time efficient algorithm to find frequent itemsets (Sharma and Sufyan, 2012).

Wang et al. in 2010 improved the efficiency of data mining in large transaction database by applying Fast-Apriori algorithm. According to the authors, data mining engine could be derived using an integration of various mining algorithms, cluster analysis, regression analysis, classification and other techniques. An engine, obtains queries from users, searchs the memory to show suitable results to the user. They applied a fast approach into existing Apriori algorithm for getting quick responses. This fast algorithm has better performance than Apriori algorithm (Wang et al, 2010).

Zeng et al. (2011) concentrated on time and space complexity of Apriori algorithm and optimize the complexities. The Hash Mapping Table (HMT) and Hash Tree methodologies were used to optimize time and space complexity. HMT and Hash tree store transactions and can locate the itemsets easily (Zeng et al, 2011). Data collection and evaluation processes are comparatively faster than traditional Apriori algorithm. Xiaohui (2012) proposed a new kind of ARM algorithm and presented an improved

Apriori algorithm. The improved algorithm can decrease the input-output operation of mining process and reduce times of database searching, saving storage space required during application of algorithm (Xiaohui, 2012). This is much more efficient than the traditional algorithms in mining association rule.

Enhanced scaling Apriori was suggested by Prakash and Parvathi in 2010. This method is an improved Apriori algorithm to limit candidate set number while producing association rules and overall execution time (Prakash and Parvathi, 2010).

In 2008 Kamrul et al. presented a novel algorithm, named as Reverse Apriori Frequent Pattern Mining. This algorithm works efficiently and produces large frequent itemsets and reduces the number of items until it takes the largest frequent itemsets (Kamrul et al., 2008).

2.1.1.2 Direct hashing and pruning (DHP) algorithm

DHP algorithm was recommended by Park et al. in 1995 to decrease the number of candidates in the early passes and the size of database. DHP employs a hashing technique aiming to restrict the number of candidate itemsets, efficiently generates large itemsets and reduces the transaction database size (Park et al, 1995).

Another hash-based approach for mining frequent itemsets was improved by Wang and Chen in 2009. The information of all itemsets are fitted into a structure by using fixed hash-based technique. This method summarizes the data information by using a hash table for predicting the number of the non-frequent itemsets and speeds up the mining process (Wang and Chen, 2009).

2.1.1.3 Partitioning algorithm

Partitioning algorithm was discovered by Savasere et al., in 1995 which is based on idea of partitioning of database in n parts to find the frequent elements so that memory problems for large databases can be solved since database is divided into n parts (Savasere et al,1995).

This algorithm decreases database scan to generate frequent itemsets but time for computing the frequency of candidate generated in each partitions increases.
On the other hand, it reduces the I/O overhead and CPU overhead for most cases significantly.

2.1.1.4 Dynamic itemset counting algorithm (DIC)

DIC algorithm was created by Brin et al. in 1997 for database partitions into intervals of a fixed size to reduce the number of transitions through the database. This algorithm aims to find large itemsets which uses fewer candidate itemsets than approaches based on sampling and uses fewer passes over the data than traditional algorithms.

Also, DIC algorithm presents a new method of generating implication rules. These rules are standardized based on both the antecedent and the consequent (Brin et al, 1997).

2.1.1.5 Sampling algorithm

Sampling algorithm was presented by Toivonen in 1996. In this algorithm, a sample of itemsets R is taken from the database instead of whole database D. This algorithm reduces the database activity for finding association rules as it requires only a subsample of the database scanned (Toivonen, 1996). This algorithm is suitable for any kind of databases, although it sometimes cannot give accurate results.

2.1.1.6 Continuous association rule mining algorithm (CARMA)

CARMA was proposed in 1999 by Hidber to compute large itemsets online. This algorithm uses method to limit the interval size to 1. The user can change minimum confidence, minimum support and the parameters during the first scan of the transaction sequence (Hidber, 1999). CARMA out-performs DIC and Apriori on low support thresholds.

2.1.1.7 Split and merge algorithm (SAM)

SAM algorithm was introduced by Borgelt and Wang in 2009. It finds frequent item sets with a split and merge technique where the data is represented as an array of transactions. The traversal order for the prefix tree and the horizontal representation form of the transaction database can be combined.
In each step, two subproblems formed with a merge step and a split step in two conditional database (Borgelt and Wang, 2009).

2.1.1.8 PRICES algorithm

This algorithm was developed by Wang and Tjortjis in 2004. This algorithm first recognizes all large itemsets and creates association rules. It reduces large itemset generation time by logical operations and scanning database (Wang and Tjortjis, 2004). It is an efficient method and ten times as fast as Apriori algorithm.

2.1.2 Algorithms based on vertical layout database

In vertical layout data set, each column contains an item, followed by a TID list.
An example of vertical layout database set is shown in Table 2.4 (Gupta and Garg, 2011).

ITEM	TID_List
I1	T1, T2, T4, T5
I2	T2, T4, T5,T6
I3	T1, T2, T4
I4	T3, T4, T5
I5	T1, T2, T3

Table 2.4 Example of vertical layout database

2.1.2.1 Equivalence class transformation algorithm (ECLAT)

ECLAT algorithm was created by Zaki in 2000 for exploring frequent itemsets in a transaction database. It uses vertical layout database. Each item utilizes intersection based method to calculate the support. Support of an itemset can be calculated by intersecting of any two subsets.

Confidence is not calculated in this algorithm (Zaki, 2000). The algorithm finds the elements using depth first search. It scans the database only once.

2.1.3 Algorithms for mining from projected layout based database

This kind of database uses divide and conquer strategy to mine itemsets. It counts the support more efficiently then based on Apriori algorithms. The projected layout consists of record id separated by column. Tree Projection algorithms may work based on two kinds of ordering: breadth-first and depth-first (Neelima et al, 2014).

2.1.3.1 FP_growth algorithm

FP-growth method has been devised for mining of the complete set of frequent itemsets without candidate generation by Han et al. in 2000.
FP-growth method is an effective tool to mine long and short frequent patterns (Han et al, 2000).

FP-growth has several benefits over other methods:

- Creating a highly compact FP-tree which is smaller than the original database,
- Implementing a pattern growth method to prevent costly candidate generation,
- Saving the costly database scans in the subsequent mining processes,
- And working in a divide-and-conquer way and decreasing the size of the subsequent conditional pattern bases and conditional FP-trees.

Extensions and many alternatives were implemented to the FP-Growth approach: Depth-first frequent itemset generating algorithm (Agarwal et al., 2001), H-Mine algorithm (Pei et al., 2007); discovering top-down and bottom-up traversal of such trees in pattern-growth mining by Liu et al. in 2002; and prefix-tree-structure for efficient pattern growth mining by Zhu and Grahne in 2003.

2.1.3.2 H_mine algorithm

H-Mine was developed by Pei et al. in 2007 which was created using in-memory pointers. H-mine uses an H-struct new data structure for mining (Pei et al, 2007).
In large databases, it firstly makes a partitioning of the database and mines the partitions in main memory using H-struct.

It benefits of this data structure and dynamically adjusts links in the mining process and runs very fastly in memory-based settings. H-mine has demonstrated a good performance for various types of data. However, execution time is larger than other algorithms because of partitioning (Pei et al, 2007).

Table 2.5 shows frequent ItemSet mining algorithm.

Horizontal Layout Based			
Study	**Authors / Year**	**Method**	**Advantages**
Fast Algorithms For Mining Association Rules.	Agrawal, Ramakrishnan, 1994	Combining the best features of Apriori and AprioriTid, AprioriHybrid algorithm.	AprioriHybrid performs better than Apriori in almost all cases.
An Effective Hashbased Algorithm for Mining Association Rules.	Park et al., 1995	DHP algorithm (Direct Haching and Pruning)	Restricts the number of candidate itemsets and reduce the transaction database size
An Efficient Algorithm For Mining Association Rules In Large Databases.	Savasere et. al., 1995	Partitioning algorithm	This algorithm decreases database scan to create frequent itemsets. Reduces the I/O overhead and CPU

Dynamic Itemset Counting And Implication Rules For Market Basket Data.	Brin et al., 1997	DIC(Dynamic itemset counting) algorithm	Decreases the number of transitions through the database and use fewer candidate itemsets than approaches based on sampling.
Sampling Large Databases For Association Rules.	Toivonen. 1996.	Sampling Algorithm	Reduces the database activity for finding association rules, less scan or time.
Online Association Rule Mining.	Hidber, 1999	CARMA(Continuous Association Rule Mining Algorithm) algorithm	Out-performs Apriori and DIC on low support thresholds. Use memory more efficiently
SAM: A Split And Merge Algorithm For Fuzzy Frequent.	Borgelt and Wang, 2009	SAM(Split and Merge Algorithm)	This algorithm can be implemented on external storage or relational databases easily
PRICES: An Efficient Algorithm For Mining Association Rules.	Wang and Tjortjis, 2004	PRICES Algorithm	Reduces large itemset generation time. It is ten times as quick as Apriori in some cases
Vertical Layout Based			
Study	**Authors / Year**	**Method**	**Advantages**
Scalable Algorithms For Association Mining.	Zaki, 2000	Six new algorithms combining these features (ECLAT (Equivalence CLAss Transformation), MaxEclat, Clique, MaxClique, TopDown, and AprClique)	Minimizes I/O costs by making only a small number of database scans, decreases computation costs

Projected Layout Based			
Mining Frequent Patterns Without Candidate Generation.	Han et al., 2000	FP-growth method	Saves the costly database scans in the subsequent mining processes and decreases the size of the subsequent conditional pattern bases and conditional FP-trees.
HMine: Fast And Space Preserving Frequent Pattern Mining In Large Databases.	Pei et al., 2007	H-Mine algorithm	H-mine has an great performance for different kinds of data and a polynomial space complexity.

Table 2.5 Frequent itemset mining algorithm

2.2 Sequential Pattern Mining

Sequential pattern mining discovers frequent subsequences as patterns in a sequence database. Ordered elements or events are found in a sequence database. For example: <a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)> Table 2.6 shows a sequence database.

SID	Sequence
5	<a(abc)(ac)d(cf)>
10	<(ad)c(bc)(ae)>
15	<(ef)(ab)(df)cb>
20	<eg(af)cbc>

Table 2.6 Sequence database

There are several applications of sequential pattern mining:

- Customer shopping sequences: A customer can make several next purchases, e.g., buying a PC and Antivirus tools and some software, followed by buying a memory card, and finally buying a printer and some office papers.
- Medical treatments, natural disasters.
- Weblog click streams, telephone calling patterns.

- Science and engineering processes.
- Gene structures and DNA sequences.

Sequential pattern mining can be classified into two major groups: (1) Apriori-based Approaches, and (2) Pattern-Growth-based Approaches.

2.2.1 Apriori based approaches (the candidate generation-and-test approach)

The candidate generation-and-test approach is an extension of the Apriori-based frequent pattern mining algorithm to sequential pattern analysis (Pei et al 2004).
Two main methods have been developed based on this idea: (1) GSP, a horizontal format-based sequential pattern mining method, (2) and SPADE, a vertical format-based method.

2.2.1.1 Generalized sequential patterns algorithm (GSP)

GSP is a horizontal data format based sequential pattern mining algorithm suggested by Agrawal and Srikant in 1996. It contains a sliding time window, time constraints and user-defined taxonomies. In this work, GSP uses the downward-closure property of sequential patterns and adopts a multiple pass, candidate generate-and-test approach (Srikant and Agrawal, 1996).

2.2.1.2 Sequential pattern discovery using equivalent classes algorithm (SPADE)

SPADE algorithm is an Apriori-Based Vertical Data Format algorithm represented by Zaki in 2001. The algorithm decomposes the original problem into smaller sub-problems which can be easily solved in main memory using efficient lattice search techniques and simple join operations (Zaki, 2001).

2.2.2 Pattern-growth-based approaches

These approaches provide efficient sequential pattern mining in large sequence databases without candidate generation.

Two main Pattern-Growth algorithms are Frequent pattern-projected Sequential Pattern Mining (FREESPAN) (Han et al., 2000) and Prefix-projected Sequential Patterns Mining (PrefixSpan) (Pei et al., 2001).

2.2.2.1 Frequent pattern-projected sequential pattern mining (FREESPAN)

FREESPAN algorithm is proposed by Han et al. in 2000 for the purpose of reducing efforts of candidate subsequence generation. This algorithm uses frequent items to recursively project sequence databases into a set of smaller projected databases. This work showed that FREESPAN mines the complete set of patterns and runs more efficiently and faster than Apriori-based GSP algorithm (Han et al, 2000).

2.2.2.2 prefix-projected sequential patterns mining (PrefixSpan)

This is a pattern-growth approach to sequential pattern mining, was developed by Pei et al. in 2001. PrefixSpan works in a divide-and-conquer way. This algorithm projects recursively a sequence database into a set of smaller projected databases and reduces the number of projected databases using a pseudo projection technique (Pei et al., 2001).

SPADE, GSP and PrefixSpan have been compared by Han et al. in 2004. PrefixSpan has better performance than GSP, FREESPAN, and SPADE and consumes smaller memory space than GSP and SPADE.

Table 2.7 shows sequential pattern mining algorithms.

Apriori Based			
Study	**Authors / Year**	**Method**	**Advantages**
Mining Sequential Patterns: Generalizations And Performance Improvements.	Srikant,, Agrawal ,1996	Generalized Sequential Patterns (GSP)	GSP is much faster than the Apriori All algorithm. It guarantees finding all rules that have a user-specified minimum support.

SPADE: An Efficient Algorithm For Mining Frequent Sequences.	Zaki, 2001	SPADE algorithm	SPADE outperforms the best previous algorithm. Problems can be solved in main memory easily.
Pattern-Growth-based			
FREESPAN: Frequent Pattern-projected Sequential Pattern Mining.	Han et al. ,2000	FREESPAN (Frequent pattern-projected Sequential Pattern Mining)	FREESPAN mines the complete set of patterns and runs more efficiently and faster than GSP algorithm.
Mining Sequential Patterns By Pattern-growth: The PrefixSpan Approach.	Pei et al., 2004	PrefixSpan algorithm	PrefixSpan has better performance than the apriori based algorithm FREESPAN, GSP and SPADE.

Table 2.7 Sequential pattern mining algorithm

2.3 Structured Pattern Mining

Complicated commercial and scientific applications need to resolve more complicated patterns than sequential patterns and frequent itemsets. For example; sophisticated patterns consist of trees, lattices, and graphs. Graphs play a major role in modelling sophisticated structures. They are used in various applications such as, chemical informatics, text retrieval, video indexing, bioinformatics, web analysis and computer vision. Frequent substructures can be discovered in a collection of graphs. Washio and Motoda in 2003 provided a survey on graph-based data mining (Washio and Motoda, 2003).

Several methods have been developed for mining interesting subgraph patterns from graph datasets such as greedy search based approaches like SUBDUE, mathematical graph theory based approaches like FSG and GSPAN, inductive logic programming (ILP) based approaches like WARMR etc. These algorithms are shown in Table 2.10.

2.3.1 SUBDUE algorithm

SUBDUE algorithm is a graph-based relational learning system which is developed by Holder et al. in 1994 developed over the years (Holder et al., 1994).
SUBDUE produces a smaller number of substructures in graph datasets by finding subgraphs and can efficiently discover best compressing frequent patterns.
This algorithm is very efficient for finding recurring subgraphs in a single large graph (Ketkar et al., 2005).

2.3.2 Frequent subgraph discovery algorithm (FSG)

FSG, algorithm is proposed by Karypis and Kuramochi in 2004 for finding frequently occurring subgraphs in large graph datasets. This algorithm can be used to explore recurrent patterns in spatial, scientific and relational datasets. This work showed that FSG is efficient for finding all frequently occurring subgraphs in datasets that contain over 200,000 graph transactions and scales (Karypis and Kuramochi, 2004).

2.3.3 Graph-based substructure pattern mining algorithm (GSPAN)

GSPAN algorithm discovers frequent substructures without candidate generation. This algorithm can be used for mining all kinds of frequent substructures including sequences, trees, and lattices. GSPAN algorithm mines frequent subgraphs more efficiently than others (Yan and Han, 2002). Also it outperforms FSG algorithm in mining larger frequent subgraphs in a bigger graph set with lower minimum supports.

2.3.4 Inductive logic programming algorithm (WARMR)

WARMR, a powerful Inductive Logic Programming (ILP) Algorithm was presented by King et al. (2001). WARMR is the first ILP data mining algorithm to be used to chemoinformatic data. WARMR extends APRIORI to discover frequent queries in data by using rules to generate the candidates from frequent queries and mines Association Rules in Multiple Relations (ARMR's) (King et al., 2001). WARMR is better than previous algorithms for frequent pattern discovery.

Table 2.8 shows structured pattern mining algorithms.

Structured Pattern Mining			
Study	**Authors / Year**	**Method**	**Advantages**
Substructure discovery in the SUBDUE system.	Holder et al., 1994.	SUBDUE algorithm.	It is very efficient for finding recurring subgraphs in a single large graph.
An Efficient Algorithm for Discovering Frequent Subgraphs.	Karypis and Kuramochi , 2004.	FSG (Frequent SubGraph discovery) Algorithm	It is efficient for finding all frequently occurring subgraphs in datasets containing over 200,000 graph transactions and scales.
Graph-Based Substructure Pattern Mining.	Yan and Han, 2002.	GSPAN Algorithm	It mines frequent subgraps more efficiently with lower minimum supports.
WARMR: A Data Mining Tool For Chemical Data.	King et al, 2001.	WARMR (Inductive Logic Programming) Algorithm.	It is better than previous algorithms for frequent pattern discovery.

Table 2.8 structured pattern mining algorithms

3. DATA MINING

3.1 What Is Data Mining?

Today, thanks to rapidly developing technology and software, datas are stored very quickly. As a result, size of the stored data is increasing day by day. Big data brings some problems along with. Many times, when they are not managed well, the amount of data stored on database grows exponenetially. Hence, the complexity of data increases. Consequently, need for better analysis techniques also increases.

Analysis of thousands of records cannot be done manually and needs to be done automatically. Data mining has an significant function serving for this purpose. Data mining is finding relationships and rules through large amounts of data for making predictions about future using software. It has been defined as "*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."* (Frawley et al., 1992; Fayyad et al., 1996).

In other words, data mining is the application that search large data stores for exploring trends and patterns.

The key features of data mining are (Manjunath, 2015):

- Creation of useful information,
- Discovery of patterns automatically,
- Focus on large databases and data sets,
- Estimation of likely outcomes.

3.2 History Of Data Mining

From past to present datas were always interpreted and information was asket to get. Today, data mining is used in many areas. But the birth of the term of data mining dates back to the 1960s. In this term, were given such names as data fishing, data dredging Instead of data mining.

The concept of the database and storage of data took place in the world of technology. At the end of 1960, simple computer have been developed by scientists (Carbone, 2000).

In 1970s, Relational Database Management System applications began to be used. With developed database management systems, it is possible to store and query petabytes and terabytes of data. Furthermore, expert systems have been developed based on simple rules and simple machine learning has been provided. Also, data warehouses allow users to analyse the information on different databases. In 1975, John Henry Holland wrote a book, Adaptation in Natural and Artificial Systems, on genetic algorithms. In this book, the theoretical foundations and exploring application were presented (Li, 2015).

Database management systems has become widespread in the 1980s and has been implemented in various areas. During the last three decades, companies created databases which consist of very large amount of data about their customers, competitors and products. SQL database query language or similar language can be used to access datas. HNC trademarks the phrase “database mining” for product protection called DataBase Mining Workstation. It was used to create neural network models, now do not exist. In 1989 The term “Knowledge Discovery in Databases” (KDD) was invented by Gregory Piatetsky-Shapiro (Li, 2015).

In 1990s, studies and publications has started on how useful informations can be found from databases in which the amount of data is increasing exponentially. The term “data mining” appeared in the database community. In 1992, the first software for data mining was implemented. Vladimir N. Vapnik, Isabelle M. Guyon and Bernhard E. Boser recommended a development on the original support vector machine to recognize patterns and analyze data used for classification and regression analysis. In 1993, Gregory Piatetsky-Shapiro started Knowledge Discovery Nuggets (KDnuggets). KDnuggets.com have a much wider audience now (Li, 2015).

In the 2000s, data mining continuously improved and it has been implemented in almost all areas and increased interest in this field. Historical development of data mining, is shown in Figure 3.1. (Savas et al., 2012).

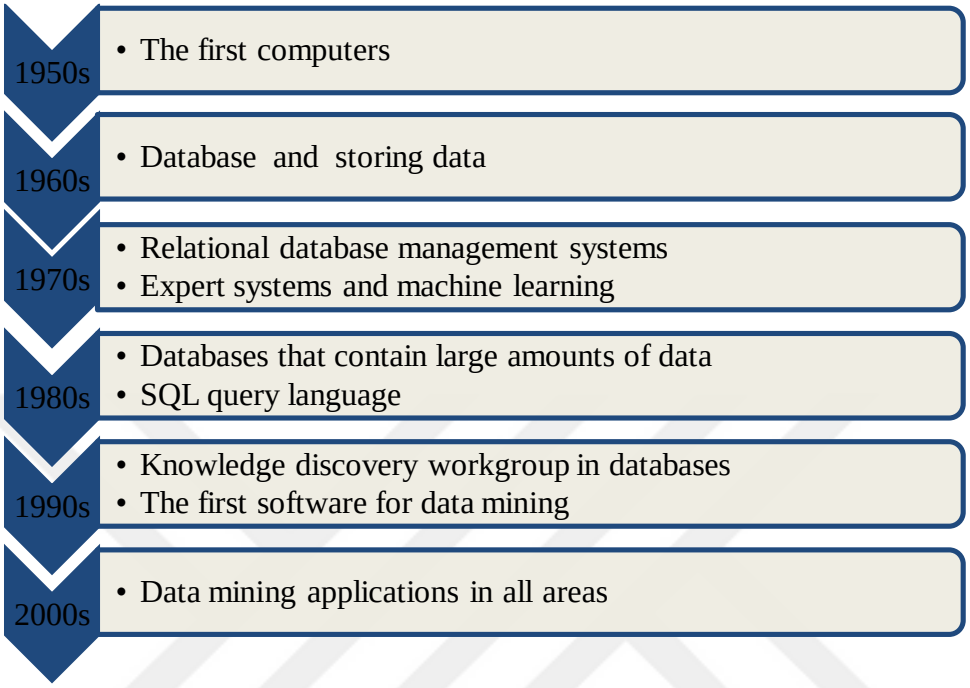

Figure 3.1 The historical process of data mining (Savas et al., 2012).

3.3 Data Mining Process

Data mining is a process at the same time (Figure 3.2).

The steps of the data mining process are generally as follows (Shearer, 2000):

1. Defining the problem
2. Preparation of the data
3. The establishment and evaluation of the model
4. Using model
5. Monitoring model

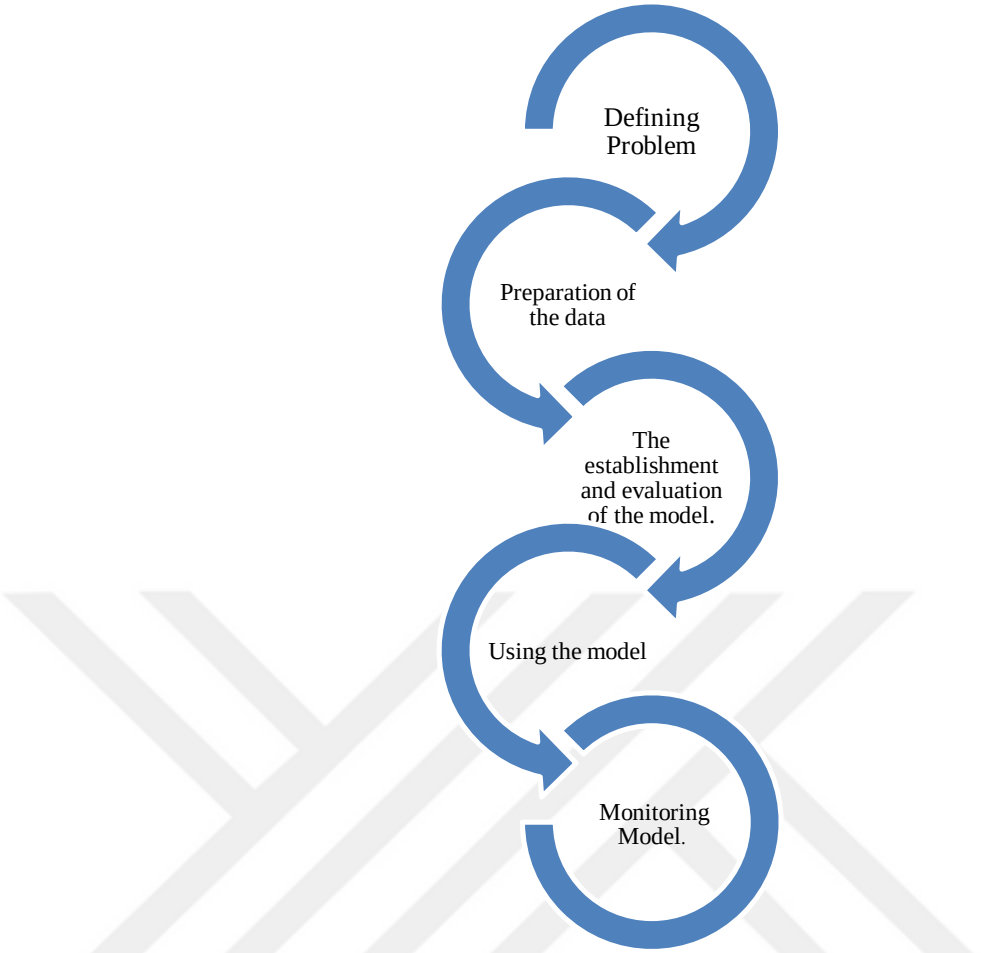

Figure 3.2 Data mining process (Shearer, 2000)

3.3.1 Defining the problem

The first requirement for success in data mining study is the identification that application will be made for which company purpose. Company objectives should focus on the problem and be expressed in a clear way. Also, estimates relating to costs and benefits should be specified.

3.3.2 Preparation of data

Preparation of data phase is composed of collection, assessment, consolidation and cleaning, selecting and transformation.

3.3.2.1 Collection

In this step, the data source to collect data is determined. Organization's own data sources and databases such as census, weather, central bank or data marketing companies can be utilized.

3.3.2.2 Assessment

The degree of compatibility of collected data should be reviewed and evaluated in this step because models that will provide best results can only be based on good data.

3.3.2.3 Consolidation and cleaning

In this step, the data collected from different sources is collected in a single database and a pruning process is conducted.

3.3.2.4 Selection

In this step, data selection is done depending on the model to be established. It means selection of dataset that will be used for training of dependent, independent variables and model. Variables which can cause a weight reduction of other variables in the model such as Sequence number, identification number should not be in the model.

3.3.2.5 Transformation

In a model developed for the estimation of credit risk, the use of separate debt and revenue data is preferred instead of precalculated rate such as debt to income. Also, algorithm used in the model, plays a major role in representation of the data.

3.3.3 Establishment and evaluation of the model

In this phase, the most appropriate model is found for described problems. Data preparation and model building stages, are iterative process until the emergence of the best model.

3.3.4 Using the model

Model that was established and its validity was accepted can be applied directly or can be used as lower part of another application. For example, established models can be used directly in business applications such as risk analysis, credit assessment, fraud detection or can be embedded into an application that will provide order.

3.3.5 Monitoring model

Changes occur in the characteristics of the entire system and datas that are produced by system. For this reason, established models need to be monitored and revised continuously. Graphs showing the differences between predicted and observed variables are very useful for monitoring results.

3.4 Data Mining Areas

With advanced technology, datas can be collected very quickly, stored, processed and made available to the institutions as information.
Today, quick access to information is very important especially in the business world where requires decision making especially to provide immediate and maximum profit. Many studies have been made on distributed and large-volume data sets. They have been focused on data mining to satisfy fast and reliable information.

Below are listed the areas of data mining.

Retail/Marketing

• Determination of consumption trends of consumers and purchasing habits,
• Estimating the response to the e-mail campaigns,
• Definition of the relationship between demographic characteristics of consumers,
• Performing market analysis prediction of response to be given to a product to be released.

Banking

• Detection of fraud cases that may occur in credit card use,

• Identification of loyal customers to the bank,

• Create card usage profile, estimating changes in customer card use,

• Determining the costs of credit card users group,

• Finding hidden correlations between different financialndicators,

• Establish specific rules using historical market datas (Akpınar, 2000).

Economy

• Identification of economic trends and irregularities,

• Creation of economic policies for the country's economy.

Education

• Examination training models and student achievement situations,

• Determination of state increasing success in education,

• Produce estimates for productivity growth.

Medicine

• Estimation of patient behavior,

• Identification of successful medical therapy performed on different diseases,

• Estimation of potential disease.

Health Care and Insurance

• Conducting analysis of money paid through the insurance policy,

• The estimation of which customers will get a new insurance policy,

• Determining the behavior patterns of risky customers,

• Identifying fraudulent behavior.

Security

• Determination of the pros and cons propaganda pages with scanning of Internet pages,

• Determination of terrorist activities

•Following the communication tools.

Transportation

• Decide on the distribution list, determining the vehicles of distribution channels,

• Making of load pattern analysis and determination of the loading state.

3.5 The Problems in Data Mining (Savas et al., 2012)

Major problems can occur in data environments where big data is found. Data mining systems can lead to incorrect results working wrong. Therefore, problems which prevent working correctly of data mining systems must be solved.

Potential issues in data mining applications:

Residual Data:

They are unnecessary attributes in the sample set which is used to achieve the desired result in the problem. It may appear in many operations.

Uncertainty:

It relates to the degree of noise in the data.

Null Value:

It can be any attribute values that are not included in the primary key. It is not equal to any value.

Dynamic Data:

Online databases are dynamic and these content changes constantly. This is very inconvenient for knowledge discovery methods.

Missing Data:

Missing data is another major challenge to be faced. What to do when missing data are:

- The average of the variables can be used in place of missing data.
- Appropriate value can be used based on existing data.
- Records with missing data can be removed.

Using Different Types of Data:

Applications can use different data types such as categorical data types, integer, floating point numbers, multimedia data, data including geographic informations, ordinal datas from many sources.

Database size:

Database size is increasing rapidly. Algorithms should be used very carefully in large samples.

3.6 Data Mining Algorithms

Models used in data mining are examined under two main categories including predictive and descriptive.

3.6.1 Predictive models

In predictive models, it is aimed to develop a model utilizing the results for known data and estimate the value of the results for data sets. Predictive models are regression and classification models.

3.6.1.1 Regression

Regression analysis is an analysis method that enables us to find the cause and effect relationship between variables.

With regression analysis, the answers to the following questions is searched:

- Is there a relationship between independent and dependent variables?
- What is the strength of this relationship?
- What kind of a relationship between the variables?
- How to be estimated future value of the dependent variable?
- A special variable or variables group.
- What is the impac of a special variable or variables group on other variables? (Ada, 2012)

Regression is divided into 2 main groups as single variable and multi variable according to number of independent variables.

It includes a single variable. It examines the relationship between an independent variable and a dependent variable. Line equation representing this relationship is formulated.

$$Y = a + bx \quad (3.1)$$

Regression models have one dependent variable and multiple independent variables.

$$Y=\beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u \quad (3.2)$$

3.6.1.2 Classification

The algorithm discovers relationships between the attributes to predict the outcome. It can be predicted a certain outcome with reference to original data or based on a model of that data.

There are many classification methods such as Decision Trees, Bayesian Classification, Backpropagation, Support Vector Machines, K-nearest Neighbour, Neural Networks, Genetic Algorithms, Time Series Analysis.

3.6.1.2.1 Decision trees

A decision tree is a one of the predictive machine-learning models. It evaluates all the possible outcomes based on the data and provides ease to understand the problem to decision maker with use of symbols such as line, square, circle.
Decision trees are widely used in operations research. It can help identification of target, preferences, risks, earnings of a business management.

A decision tree includes 3 types of nodes:

- Decision nodes: Represent decision variables and it is represented by squares.
- Chance nodes

Represent events that can take certain values with certain probabilities and it is represented by circles.

●End nodes

It indicates the final result for branch that it considered this point as destination point. It is represented by triangles.

Figure 3.3 shows new decision tree.

https://upload.wikimedia.org/wikipedia/commons/a/ad/Decision-Tree-Elements.png

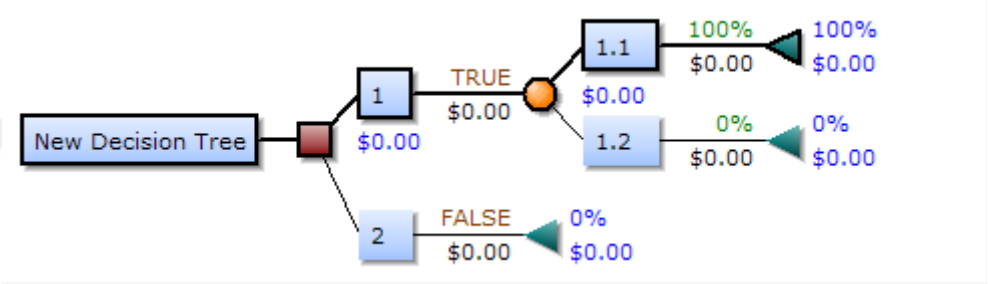

Figure: 3.3 New decision tree

https://upload.wikimedia.org/wikipedia/commons/a/ad/Decision-Tree-Elements.png

3.6.1.2.2 Naive bayes classification algorithm

Naive Bayes classification algorithm is a classification / categorization algorithm named after Thomas Bayes. It aims to identify the category of the class that the data submitted to the system with defined by a series of calculations based on Naïve Bayes classification principles of probability. Bayesian classification ensures practical learning algorithms, prior knowledge and observed data can be combined. Bayesian Classification ensures a helpful perspective to understand and evaluate many learning algorithms (Chai, 2002).

Bayes' Theorem is discovered by Thomas Bayes. Bayesian probability interprets the concept of probability. Probabilities divided into 2 groups:

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

H is some hypothesis and X is data tuple.

In Bayes' Theorem,

$$P(H/X)= P(X/H)P(H) / P(X) \quad (3.3)$$

Naive Bayes theorem can handle real, discrete data and streaming data well. Naive bayes theorem can be used in areas such as data mining, biomedical engineering, identfying medical of diseases or abnormalities, the classification of the graphs electrocardiogram (ECG), the separation of graphics electroencephalography (EEG), genetic researchs, the identification spamming, parsing text and product classification.

3.6.1.2.3 Time series algorithm

Time series are series which are shown distribution of observations according to the time. They report the values of a variable observed in certain time frame. For example; advertising expenditures by years and Purchases of goods by years are shown as example for time series.

The following table is an example of time-series data. It shows imports and incomes of the company by years (Tarı, 1999). The model obtained based on these data;

Ct= import in the period t

Yt= income of period t

Ut=the margin of error for the period

$$C_t=b_0 + b_1 Y_t+U_t \quad (3.4)$$

Table 3.1 shows imports and incomes of the company by years (Tarı, 2002).

YEAR	Import	Income
1986	7,56	39,36
1987	12,35	58,56
1988	20,47	100,58
1989	33,76	170,41
1990	58,75	287,25

Table 3.1 Imports and incomes of the company by years (Tarı, 2002).

Time series consist of many techniques such as Moving Average Method, Simple Average Method, Exponential Smoothing Method, Trend Analysis Method.

3.6.1.2.4 Genetic algorithms

Genetic algorithm is a search and the optimization method which is working in a manner similar to survival of the better pricipal partially observed in nature. The basic principles of genetic algorithms have been proposed for the first time by John Holland in University of Michigan in 1970s. Genetic algorithms are search methods based on genetics and natural selection (Fraser, 1957; Bremermann, 1958; Holland, 1975).

Genetic algorithms provide successful results for network design problems, pathfinding problems, social and economic planning problems, applications of artificial intelligence, expert systems, engineering design etc.

3.6.1.2.5 Artificial neural networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems.
ANN was developed for the aim to automatically perform capabilities such as creation and discovery of new knowledge without any help. Figure 3.4 shows artificial neural network layer.

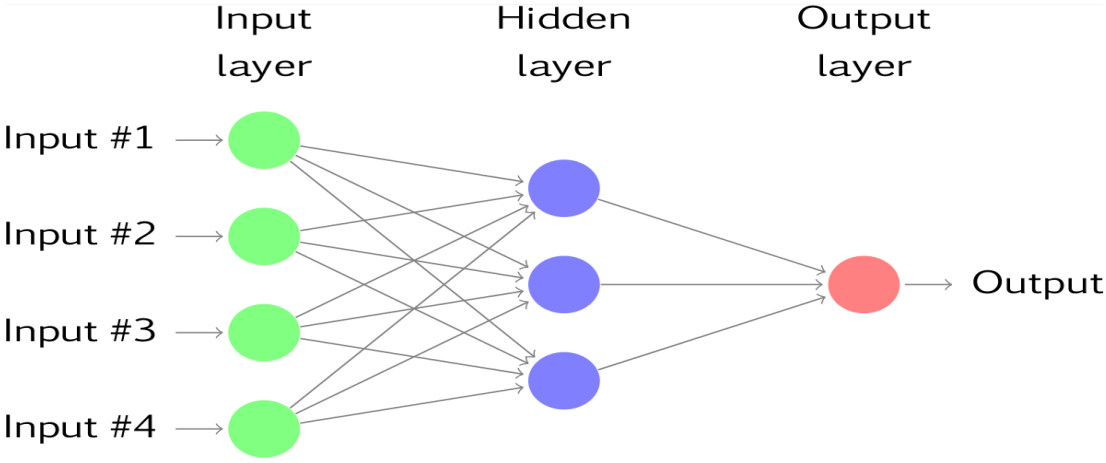

Figure 3.4 Artificial neural network layer
(https://www.otexts.org/sites/default/files/fpp/images/nnet2.png)

ANNs are used in many fields such as classification, modelling and prediction. Some of those are (Cayiroglu, 2014):

Space: Flight simulation, automatic pilot applications, controls and components of error.

Defense: Weapon routing, selecting targets, radar, sonar sensor systems, signal processing, image processing, etc.

Banking: Improving credit applications, customer analysis and investment budget estimates.

Security: Fingerprint Identification, credit card frauds detection, retinal scanning, face matching, etc.

Health: breast cancer early diagnosis and treatment, EEG, ECG, MRI, quality improvement, drug effects analysis, blood analysis etc.

Finance: Valuation, market performance analysis, budget estimation, goal setting, etc.

Production: Product design, identification of machine wear, durability analysis, quality control, business charts etc.

3.6.2 Descriptive models

Descriptive models show the main features of the data. In descriptive models, the identification of patterns of existing data that can be used to guide decision-making is provided. Descriptive models are grouped as clustering, and association rules models.

3.6.2.1 Clustering method

The process of grouping objects with their similarities is called clustering. The main purpose of cluster analysis is grouping objects(units) based on their characteristics. The result of a cluster analysis are shown as the coloring of the squares into three clusters.

Figure 3.5 shows cluster analysis.

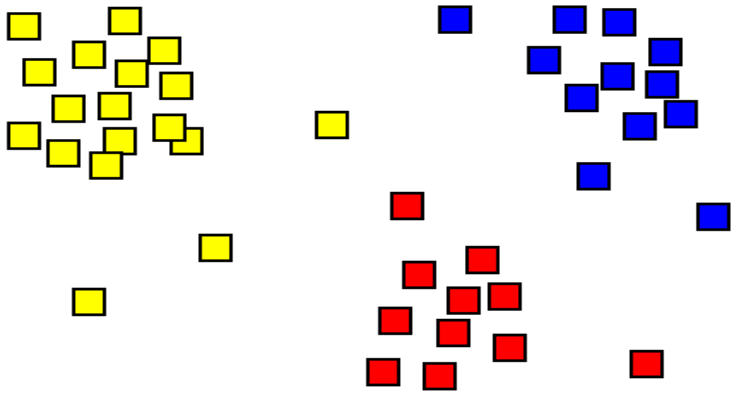

Figure 3.5 Cluster analysis

Clustering Analysis are used in different disciplines such as medicine, psychology, biology, sociology, educational sciences, economics, engineering, marketing, data mining.

4. ASSOCIATION RULE MINING

Association rules have been an attractive research topic in data mining area and attract more attention of researchers with the help of gradually increasing computational power achieved. Can be used for discovering frequent patterns, associations, correlations between data.

Association rule mining (ARM) may be helpful in decision making task with finding such relationships among the attributes in database. For example: *"It finds the new useful rules in the sales transaction database, which reflects the customer purchasing behaviour patterns, such as the impact on the other goods after buying a certain kind of goods*" (Sherdiwala and Khanna, 2015).

Some of the areas in which association rules have been used are:

1. Market Basket Analysis:

MBA is one of the most typical application areas of association rules. When a customer buys any product, what other products s/he puts in the basket with some probability may be determined by applying association rules. When determined which products are normally purchased together store managers may organize the shelves accordingly. Thus, customers can reach these products more easily. This, in turn, results with an increase in sales rates and developing effective sales strategies.

2. Medical Diagnosis:

Using association rules in medical diagnosis can be beneficial for helping physicians to cure patients. Serban et al.(2006) has suggested a technique based on relational association rules that helps to determine the possibility of illness in a certain disease.

3. Protein Sequences:

Proteins are sequences with 20 types of amino acids. Each protein consists of a unique 3-dimensional structure with amino -acid sequence.

Gupta et al.(2006) have discovered the nature of associations between amino acids in a protein.

This is the first systematic study to find global associations between amino acids. Knowledge of these association rules or constraints helps artificial proteins synthesis.

4. Census Data:

Censuses give general statistical information on society. The information about economic and population census can be estimated in planning public business (for setup new shopping malls, factories or banks), public services (funds, health, education, transport). (Malerba et al, 2001). Malerba et al. in 2001 proposed a method for the discovery of spatial association rules including spatial relations in census data.

5. Customer Relationship Management (CRM) Of Credit Card Business:

CRM helps for increasing the cohesion between the bank and credit card customers with identifing the preference of different services, products and customer groups according to their liking or preference. Chen et al. in 2005 classified customers into clusters to identify high-profit, gold customers and provided as early as possible services and products wanted by the customers.

4.1 Commonly Used Terms

In MBA literature, some words and terms are commonly used. Before giving some details of the employed algorithms, it is necessary to give short definitions of them may help the reader to build a well-structured background.

•Itemset: An itemset is collection of one or more items. A k-itemset is an itemset that contains k number of items. Example: {Milk, Bread, Butter}

•Frequent itemset: This is an itemset whose support is greater than or equal to a min-support.

•Candidate set: A set of itemsets that need to test for seeing if they adapt a specific requirement.

The two important basic measures of association rules are support(s) and confidence(c).

Support

Support defines how often a rule is applicable to a given data set. The rule X ⇒Y holds with support s if %s of transactions in D contain X ∪Y.

$$S=\frac{\sigma(X\cup Y)}{\#\ of\ trans.} \quad (4.1)$$

In table 4.1, a simple case with five instances is used to demonstrate how the support of a certain X=>Y rule can be calculated.

TID	Items	Support=Occurence /Total Support
1	Bread, Butter, Peanut	Total Support=5
2	Bread,Butter, Milk	Support{Bread,Butter}=2/5=%40
3	Butter,Peanut	Support{Butter,Peanut}=3/5=%60
4	Bread,Peanut	Support{Bread, Butter,Peanut}=1/5=%20
5	Butter,Peanut,Milk	

Table 4.1 Example of support measure

Confidence

Confidence defines how frequently items in Y appear in transactions that contain X.The rule X ⇒Y holds with confidence c if %c of the transactions in D that contain X also contain Y.

$$C=\frac{\sigma(X\cup Y)}{\sigma(X)} \quad (4.2)$$

In table 4.2, a simple case with five instances is used to demonstrate how the confidence of a certain X=>Y rule can be calculated.

TID	Items	Given X⇒Y Confidence=Occurence{Y}/Occurence{X}
1	Bread, Butter, Peanut	
2	Bread,Butter, Milk	Confidence{Bread⇒Butter}=2/3=%66
3	Butter,Peanut	Confidence{Butter⇒Peanut}=3/4=%75
4	Bread,Peanut	Confidence{BreadButter⇒Peanut}=1/2=%50
5	Butter,Peanut,Milk	

Table 4.2 Example of confidence measure

Lift

Lift measures how much more often X and Y occur together than expected if statistically independent.

$$\text{Lift } (X \rightarrow Y) = \frac{Support\ (X \wedge Y)}{Support(X) * Support(Y)} \qquad (4.3)$$

4.2 Creation of Association Rules

The main objective of mining association rules is creating rules from Transaction Group D. Support value of the rules must be greater than or equal to specified minimum support (minSUP). Confidence value of the rules must be must be greater than or equal to specified minimum confidence value (minCONF).

A Brute-Force approach for mining association rules is to list possible all rules and compute the support and confidence for every possible rule. Rules which have support and confidence values that smaller than minSUP and minCONF threshold values. This approach is very expensive because there are many rules that can be extracted from a dataset that contains d item is 3^d-2^{d+1}+1 (Kupar et al, 2005).

Creation of Association Rules Consists Of Two Steps:

1. Determine Common Items

Item sets whose support value equal to or greater than the minimum support are created. Support value of each object must be greater than previously defined minimum support value, min_support.

Given d items, we have 2d possible itemsets.2d-1 sets of items can be created from d items. This step determines performance of algorithms.

2. Creation of Rules

Create rules that have confidence value is greater than or equal to minconf. These rules should provide minSUP and minCONF status.

4.3 Successfull Points Of Market Basket Analysis

Successfull points of MBA are summarized as:

Produce understandable and easy results.

Can work on the data with different sizes.

Calculations which are required for basket analysis is simpler than another methods such as genetic algorithms, neural networks.

4.4 Failed Points Of Market Basket Analysis

MBA has some failed points such as:

As the size of the problem grows, the required calculations increases exponentially.

It ignores very rare items in log.

The most accurate method of basket analysis is produced in situations where all products have approximately the same frequency in records.

Support and confidence threshold values limit the number of generated rules. In the case of a very low threshold value user can face with the danger of losing the care rules.

4.5 Types Of Association Rules

Different types of association rules based on types of values handled, levels of abstraction involved and dimensions of data involved as seen in the figure 4.1. below.

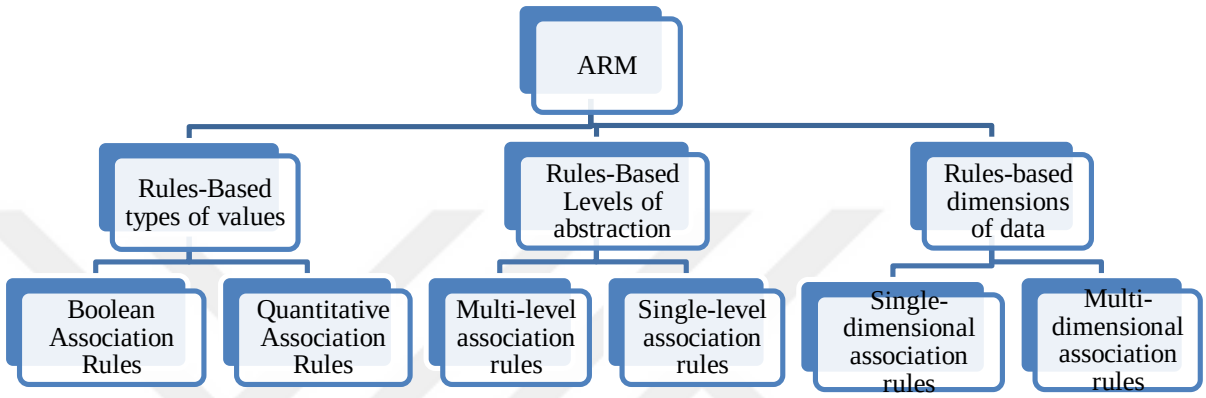

Figure 4.1 Types of association rules

4.5.1 Types of values handled

Types of values handled have 2 types such as Boolean Association Rules and Quantitative Association Rules.

Boolean Rules that concern associations between the presence or absence of items. e.g. "buys A" or "does not buy A". buys(x, " DBMiner ") ^ buys(x, "DMBook") ⇒ buys(x, "SQLServer") [%0.2, %60]

Quantitative rules concern associations between attributes or quantitative items. age(x, "35..45") ^ income(x, "45..50K) ⇒ buys(x, "PC") [%1, %85] (Slimani, 2014).

4.5.2 Levels of abstraction involved

Levels of abstraction involved divide into 2 groups such as single-level association rules and multi-level association rules.

Associations between attributes or items at the same abstraction level.

i.e., at the same level of hierarchy.

Cola, Chips ⇒ Bread [%0.5, %56]

Associations between attributes or items at different abstraction level.

i.e, at different levels of concept hierarchy.

Cola:Pepsi, Chips:Doritos:Barbeque ⇒ Bread [%0.1, %78]

4.5.3 Dimensions of data involved

Dimensions of data involved consists of 2 types such as single-dimensional association rules and multidimensional association rules.

Items in the rule refer to only one predicate or dimension. Transactional data is used. Items in a rule belong to the same transaction.

buys(X, milk) ⇒ buys(X,bread)

It consists of 2 predicates or dimensions.

Items in the rule refer to two or more predicates or dimensions,

Two types of these rules are:

- Interdimensional rules: same predicates/ attributes may not be repeated in a rule age(X,"19-26") ∧ occupation(X,"student") ⇒ buys(X,"coke")
- Hybrid-dimensional rules: same predicates/ attributes can be repeated in a rule age(X,"19-26") ∧ buys(X, "popcorn") ⇒ buys(X, "coke")

5. A CASE STUDY OF MBA ON A SUPERMARKET CHAIN

There are several applications of association rules in data mining. In this thesis, products associated with each other have been determined by the application of association rules to the dataset consist of shopping records in a supermarket. Data have been analyzed and association rules have been found between products by employing Apriori algorithms.

In MBA application, 170810 *receipts* from the supermarket store are obtained. The data set consists of transactions at checkout counters of the market chain collected for a period of 6 month. These records are transferred to the software program "KNIME" and the potential relationships between products are investigated by using Apriori algorithm.

Because of privacy principles, store or product brand name are not mentioned in the study.

5.1 Dataset

During a 6-month period, 170810 receipts were collected. Considering these 170810 receipts, it was observed that the supermarkt's hyeracical classsification of the product is highly contraversial. In other words, some entries in the dataset were grouped into some subclasses where, for instance, kaşar cheese and white cheese were classified into different classes. Thus, a preprocessing of the dataset was necessary where some subclasses were combined together to build new higher level classes. After the preprocessing stage, 36 main product groups have been determined.

Before we introduce the methods used and the results we obtain, we should note that the results of this study must be considered with some cautiusness:

- Data are collected during a specific time interval will give different results in different periods.

● Campaigns and promotions of products will affect purchases in period of collected of receipts.

● Shopping habits will vary due to target audiences and region of the markets where the study was performed.

5.2 Methods And Algorithms Used

An association rule mining technique of MBA is used in finding the relationship between the products forming the dataset we considered. Apriori algorithm was preferred because of the superiority to other algorithm despite its simplicity. Additionally, it is widely used in several other studies as mentioned in review chapter. KNIME program was used due to the properties such as ease of use and simple interface.

5.2.1 Apriori algorithm

Apriori algorithm has been developed by Agarwal and Srikant in 1994. It is one of the most famous of all association rule mining algorithms. Apriori is designed to operate on databases containing transactions and can be used to generate all frequent itemset. Table 5.1 shows apriori algorithm code. (Jajodia and Wijesekera, 2005)

```
1)  L_1 = {large 1-itemsets};
2)  for ( k = 2; L_{k-1} ≠ ∅; k++ ) do begin
3)      C_k = apriori-gen(L_{k-1}); // New candidates
4)      forall transactions t ∈ D do begin
5)          C_t = subset(C_k, t); // Candidates contained in t
6)          forall candidates c ∈ C_t do
7)              c.count++;
8)      end
9)      L_k = {c ∈ C_k | c.count ≥ minsup}
10) end
11) Answer = ⋃_k L_k;
```

Table 5.1 Apriori algorithm's pseudo code (Jajodia and Wijesekera, 2005)

Each transaction consists of set of items. Considering a threshold C, the Apriori algorithm defines the item sets which are subsets of at least C transactions in the database.

Apriori makes multiple passes over the database. The first pass of the algorithm simply counts item occurrences to identify the large 1-itemsets, L_1. Then, it produces the candidate itemsets in C_1 and saves the frequent itemsets in L_1.

A subsequent pass, say pass k, has two phases. In first phase, the algorithm generates the candidate itemsets in C_k from large itemsets L_{k-1} using the apriori-gen function. This function takes as argument L_{k-1}, the set of all large (k-1)-itemsets. In the join step, This function first joins L_{k-1}.

Next, it generates all (k-1)-subsets from the candidate itemsets in C_k and deletes all candidate itemsets from C_k where some (k-1)-subset of the candidate itemset is not in the frequent itemset L_{k-1}. Then, it scans the transcational database and examines C_k for determining which candidates are frequent and the support for each candidate itemset in C_k. Saves the frequent itemsets in L_k. The algorithm terminates when L_k becomes empty. Figure 5.1. shows the steps of Apriori algorithm (Strickland, 2015).

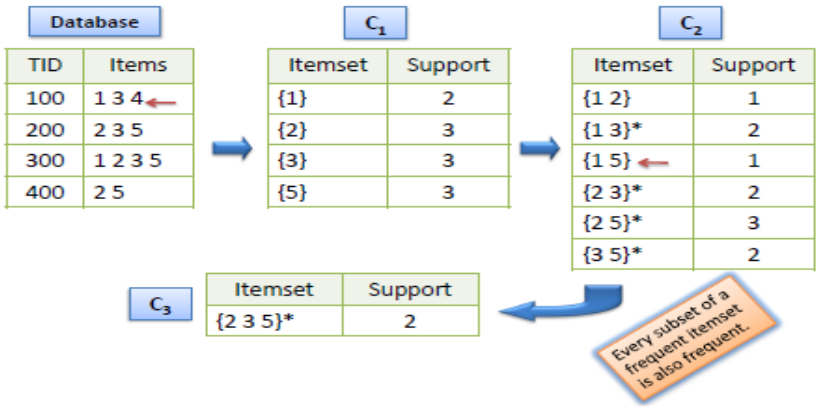

Figure 5.1 Apriori algorithm steps (Strickland, 2015)

5.2.2 Apriori algorithm solution steps

In this part, the solution procedure of Apriori algorithm is shown and association rules are found on an illustrative example to demonstrate how the algorithm works.
In Table 5.2 part of data set is shown.

ID	Bakery Products	Cheese	Meat Products	Hot Drinks	Dairy Products
1	0	1	1	0	0
2	1	0	0	0	1
3	0	0	1	0	1
4	1	1	1	1	0
5	1	0	0	0	0
Sum	3	2	3	1	2

Table 5.2 Part of data set

In the illustrative data set above, Ik={Bakery Products, Dairy Products, Meat Products, Cheese, Hot Drinks} repesents product set where sample size is D=5Firstly, the products are arranged in alphabetical order as in Table 5.3.

ID	Bakery Products	Dairy Products	Meat Products	Cheese	Hot Drinks
1	0	0	1	1	0
2	1	1	0	0	0
3	0	1	1	0	0
4	1	0	1	1	1
5	1	0	0	0	0
Sum	3	2	3	2	1

Table 5.3 Sorted data set

Next, the supports of products are found. Support of Bakery Products: 3/5=0,6The same process is repeated for other products and all support values are found.
Thus, C1 candidate set is obtained as in Table 5.4.

Item Set	Support
Bakery Products	0,60
Dairy Products	0,40
Meat Products	0,60
Cheese	0,40
Hot Drinks	0,20

Table 5.4 C1 candidate set

For example, minimum support threshold is 0.20. There is no product which does not satisfy the minimum support threshold. L1 is generated in this way as shown in Table 5.5.

Item Set	Support
Bakery Products	0,60
Dairy Products	0,40
Meat Products	0,60
Cheese	0,40
Hot Drinks	0,20

Table 5.5 L1 large item set

Supports of dual combination of products in L1 are calculated and C2 is generated. Support value for {Bakery Products, Dairy Products } = 1/5= 0.2 and confidence value for { Bakery Products, Dairy Products } = 1/3 = 0,33.

After finding support and confidence values of other combinations C2 table is formed as follows, Table 5.6.

Item set	Support	Confidence
Bakery Products, Dairy Products	0,20	0,33
Bakery Products, Meat Products	0,20	0,33
Bakery Products, Cheese	0,20	0,33
Bakery Products, Hot Drinks	0,20	0,33
Dairy Products, Meat Products	0,20	0,50
Dairy Products, Cheese	0	0
Dairy Products, Hot Drinks	0	0
Meat Products, Cheese	0,40	0,66
Meat Products, Hot Drinks	0,20	0,33
Cheese, Hot Drinks	0,20	0,50

Table 5.6 Dual combinations for C2 candidates, support and confidence values

Duals which do not provide support threshold are removed from C2 and L2 large item set is created. L2 item sets which provide support threshold is presented in Table 5.7.

Item set	Support	Confidence
Bakery Products, Dairy Products	0,20	0,33
Bakery Products, Meat Products	0,20	0,33
Bakery Products, Cheese	0,20	0,33
Bakery Products, Hot Drinks	0,20	0,33
Dairy Products, Meat Products	0,20	0,50
Meat Products, Cheese	0,40	0,66
Meat Products, Hot Drinks	0,20	0,33
Cheese, Hot Drinks	0,20	0,50

Table 5.7 L2 large itemsets

New support values are determined to get the triple combinations from L2 and C3 is generated. L3 large items sets consist of C3 candidate set as in 5.8.

Item Set	Support	Confidence
Bakery Products, Meat Products, Cheese	0,20	0,33
Bakery Products, Meat Products, Hot Drinks	0,20	0,33
Bakery Products, Cheese, Hot Drinks	0,20	0,33
Meat Products, Cheese, Hot Drinks	0,20	0,33

Table 5.8 C3 candidate set

All item sets are above the support threshold value in candidate set C3. Thus, L3 large itemset is created as in 5.9.

Item Set	Support	Confidence
Bakery Products, Meat Products, Cheese	0,20	0,33
Bakery Products, Meat Products, Hot Drinks	0,20	0,33
Bakery Products, Cheese , Hot Drinks	0,20	0,33
Meat Products, Cheese, Hot Drinks	0,20	0,33

Table 5.9 L3 Large itemset

L4 large itemsets as in 5.11 consist of C4 candidates as in 5.10.

Item Set	Support	Confidence
Bakery Products, Meat Products, Cheese, Hot Drinks	0,20	0,33

Table 5.10 C4 candidate set

Aplying same procedure as before, L4 can be obtained as in Table 5.11.

Item Set	Support	Confidence
Bakery Products, Meat Products, Cheese, Hot Drinks	0,20	0,33

Table 5.11 L4 large itemset

Finally, association rules were found. Thus, fort the given dataset following rules can be generated.All of these rules are given in the table 5.12 below.

Rule	Antecedent	Consequent	Support %	Confidence %
1	Bakery Products	Dairy Products	0,20	0,33
2	Bakery Products	Meat Products	0,20	0,33
3	Bakery Products	Cheese	0,20	0,33
4	Bakery Products	Hot Drinks	0,20	0,33
5	Dairy Products	Meat Products	0,20	0,50
6	Meat Products	Cheese	0,40	0,66
7	Meat Products	Hot Drinks	0,20	0,33
8	Cheese	Hot Drinks	0,20	0,50
9	Bakery Products	Meat Products	0,20	0,33
10	Bakery Products	Meat Products	0,20	0,33
11	Bakery Products	Cheese, Hot Drinks	0,20	0,33
12	Meat Products	Cheese, Hot Drinks	0,20	0,33
13	Bakery Products	Meat Products, Cheese, Hot Drinks	0,20	0,33

Table 5.12 All of rules obtained with Apriori algorithm

For instance, the obtained rule for Bakery Products => Dairy Products indicates that; the probability of observing the concurrent sales of Bakery Products and Dairy Products in total receipt transactions is %20 and customers purchasing Bakery Products will also buy Dairy Products with probability of %33.

5.2.3 Application areas of apriori algorithm

Apriori algorithm was used in many areas such as medicine, education, banking, e-commerce, telecommunication, finance, marketing, tourism.

Medicine

Duru (2005) developed a software, DMAP, used for exploring the social status of the diabetics. This software was executed on a database consists of 66 patients records with purpose to analyze the diabetics.

Abdullah et al. (2008) researched associations between diagnosis and treatments in their work. This work showed that this algorithm is beneficial for finding the large item sets and thus generating the association rules inmedical billing data.

Zhang et al. (2014) used Apriori Algorithm to find frequent itemsets in a database of medical diagnosis, and generates strong association rules in a medical database.
This method is very good for processing and analyzing the data of drug treatment and disease prevention in the medical area. So, this work can help doctors in medical diagnosis.

Education

Angeline(2013) used Apriori to extract the set of rules and analyzes the given data to categorize the student based on their academical performance.

This work helps to predict the performance of the student, define the average and below average students and to develop their performance. This performance enhancement will also help students to get placement in various industries according to their success.

Ahmed et al. (2009) identified the patterns in matching organization and student using Apriori Algorithm. This study aims to extract the historical placement pattern which will be a useful guideline for student matching and future organization.

Banking

Yang(2013) improved the bank customer segmentation accuracy by using Apriori algorithm. This work shows that this algorithm can be better than the traditional algorithms to improve the customer classification accuracy.

Aggelis (2003) used Apriori algorithm for the investigation association rules between products and services a bank offers. These rules aim at the continuous improvement of bank services and products to approach new customers.

E- Commerce

Revathi and Geetha (2015) proposed a modified Apriori Algorithm to develop the efficiency and accuracy of the frequent data item and classify the users based on the usage of the web. This algorithm reduced scanning time and shortened the computation time.

Sharif et al. (2005) demonstrated the development of e-commerce system by using Apriori algorithm in their work. This system has increased the potential of a good customer relationship management and give a new idea for making other e-commerce system.

Telecommunication

Jiang (2011) introduced Apriori Algorithm and used it for mining the frequency of mobile phone services, and referrals to salespersons. It helps mobile telecommunicaiton company to make decision on services planning.

Finance

Xu and Zhang (2009) analyzed the financial income of specific city in a given period. Association rules are mined using Apriori Algorithm.

Th'r motivation was finding the knowledge and rules ensuring a scientific basis for decision-making during the evaluation of the financial situation in certain region.

Marketing

Gancheva (2013) used Apriori Algorithm to mine association rules to provide valuable information about co-occurrences and co-purchases of products.

This information can be used to take decisions about marketing activity such as promotional support, inventory control and cross-sale campaigns.

Gürgen (2008) determined relations with each of the products in receipts of one of the supermarket chains in Turkey for 7 days using Apriori Algorithm. It was aimed to increase product sales and revenue.

Tourism

Liu and Fan in 2013 analyzed tourist behaviour with Apriori algorithm. They mined out rules with the association rule algorithm. As a consequence, they could make suggestions to tourists by offering better travelling experiences and services.

Mu et al. (2009) improved recommendation method for tourism information service based on improved Apriori method. They generated recommendation itemset consists of tourist destinations or scenic spot records.

The collaborative filtering recommendation was selected as promising future technologies based on the user modeling.

5.2.4 Program

As mentioned before, KNIME program is used in this study. The first version of KNIME was released in 2006 by a team of developers from a Silicon Valley software company specializing in pharmaceutical applications.

KNIME is used by life sciences banks, telcos, publishers, consulting firms, car manufacturer, and various other industries. KNIME has hundreds of modules for data integration (file I/O, database nodes supporting all common database management systems), data transformation (converter, filter, combiner)s, methods for data analysis and visualization. Additional plugins integrate methods for image mining, text mining, time series analysis.

KNIME allows modeling, analyzing, and merging different data sets. It offers a visual environment to get results in a short time offering a simple interface.

Figure 5.2 shows a screenshot of KNIME.

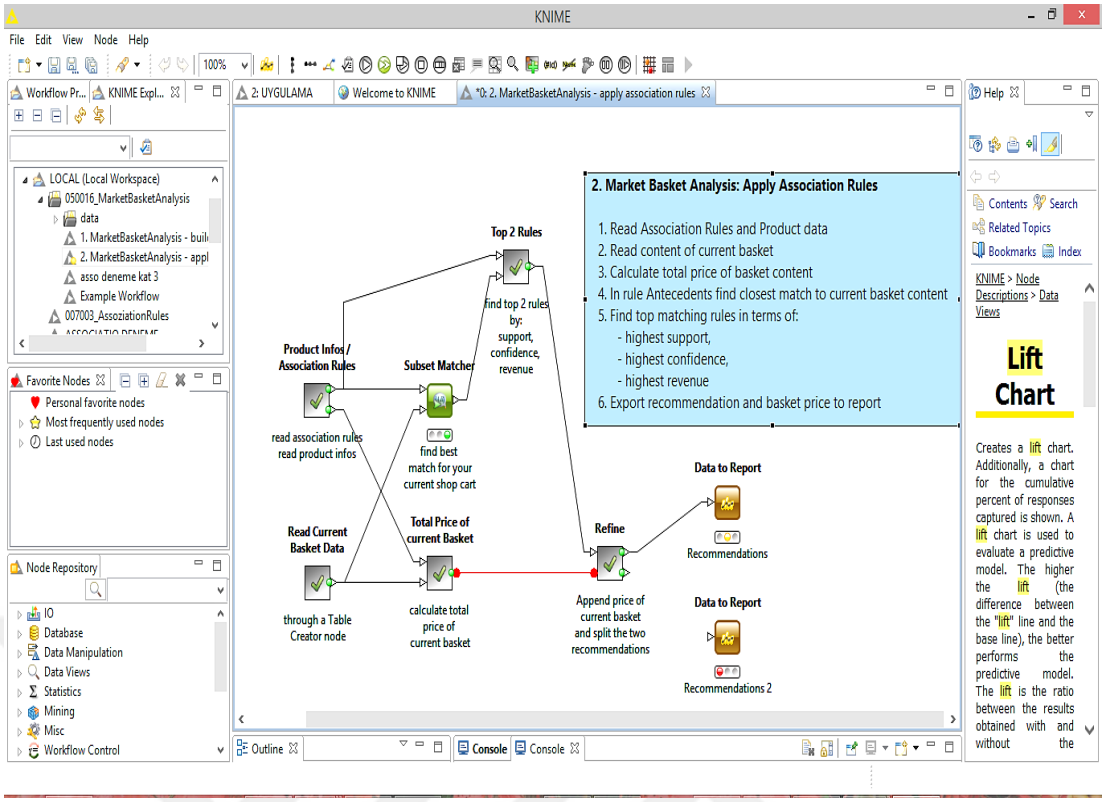

Figure 5.2 A screenshot of KNIME (wikipedia)

5.3 Goals Of Case Study

There are subtle relationships between entries on a dataset which cannot be identified at first glance. Product sales data is a classical example where mining of these relationships is necessary. Relationships between different products in market basket can be found using association rules of data mining methods. Therefore, the goals of this study can be stated as follow;

- finding the product data sets related to each other in data collection subject to this study,
- arranging shelf layout according to the relationship of the products and increasing sales amounts and sales revenue,
- taking advantage of this relationship in campaigns and promotions which were applied to products for sales process.

6. RESEARCH RESULTS AND DISCUSSION

In the database subject to this study, there are 36 product groups. The following figure 6.1. shows that 10 product groups which are the most frequently sold product groups.

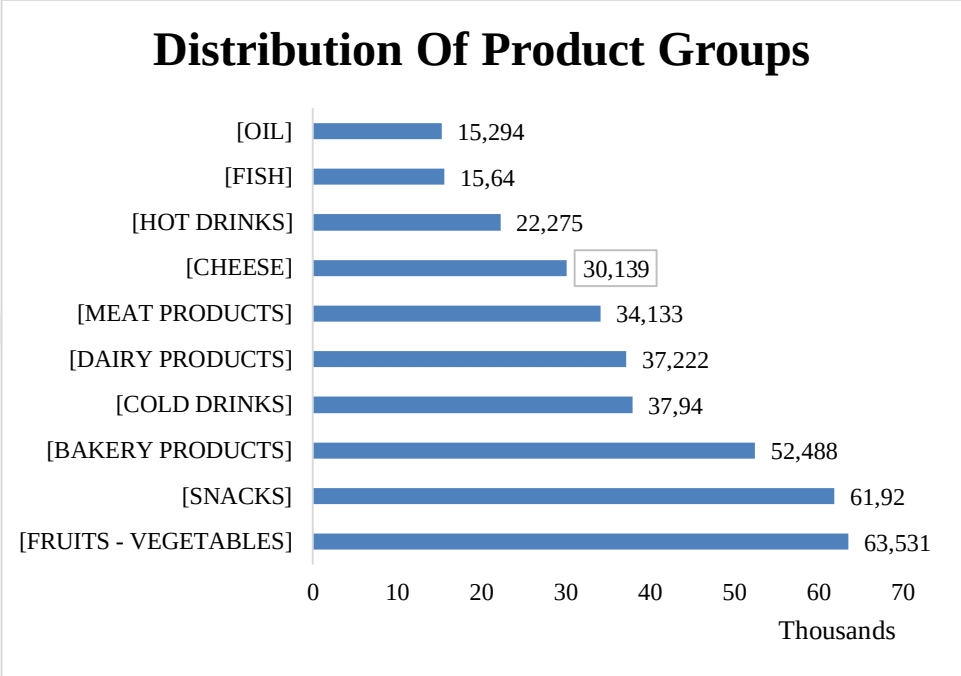

Figure 6.1 Distribution of product groups

As seen from the table, the most frequently product groups sold are FRUITS_VEGETABLES (63,531), SNACKS (61,920), BAKERY PRODUCTS(52,488) etc.

Association rules for 36 product groups are obtained using Apriori algorithm. To limit ourselves with the most valuable association rules, the minimum support value is assigned as %10 and the minimum confidence value is assigned as %20.

As a result of the analysis, 11 association rules are obtained in dataset. These rules are shown in the table 6.1.

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule 0	0.102	0.283	1.297	DAIRY PRODUCTS	<---	[SNACKS]
rule 1	0.102	0.47	1.297	SNACKS	<---	[DAIRY PRODUCTS]
rule 2	0.103	0.278	1.274	DAIRY PRODUCTS	<---	[FRUITS - VEGETABLES]
rule 3	0.103	0.474	1.274	FRUITS - VEGETABLES	<---	[DAIRY PRODUCTS]
rule 4	0.107	0.294	0.957	BAKERY PRODUCTS	<---	[SNACKS]
rule 5	0.107	0.347	0.957	SNACKS	<---	[BAKERY PRODUCTS]
rule 6	0.108	0.298	1.343	COLD DRINKS	<---	[SNACKS]
rule 7	0.108	0.487	1.343	SNACKS	<---	[COLD DRINKS]
rule 8	0.124	0.341	0.917	FRUITS - VEGETABLES	<---	[SNACKS]
rule 9	0.124	0.332	0.917	SNACKS	<---	[FRUITS - VEGETABLES]
rule 10	0.128	0.418	1.123	FRUITS - VEGETABLES	<---	[BAKERY PRODUCTS]
rule 11	0.128	0.418	1.123	BAKERY PRODUCTS	<---	[FRUITS - VEGETABLES]

Table 6.1 Association rules for product groups.

Some association rules created for the product groups are as follows;

Rules for SNACKS => DAIRY PRODUCTS

Possibility of seen together of SNAKCS and DAIRY PRODUCTS is %10.2 and Customers of SNACKS get DAIRY PRODUCTS with %28.3 probability.

Rules for DAIRY PRODUCTS => SNACKS

Possibility of seen together of DAIRY PRODUCTS and SNAKCS is %10.2 and Customers of DAIRY PRODUCTS get SNACKS with %47 probability.

Rules for FRUITS - VEGETABLES=> DAIRY PRODUCTS

Possibility of seen together of FRUITS - VEGETABLES and DAIRY PRODUCTS is %10.3 and Customers of FRUITS - VEGETABLES get DAIRY PRODUCTS with %27.8 probability.

Rules for COLD DRINKS=> SNACKS

Possibility of seen together of COLD DRINKS and SNACKS is %10.8 and Customers of COLD DRINKS get SNACKS with %48.7 probability.

We conduct another trial where we look for more subtle relations between product groups. In the second trial, the minimum support value is decreased to %5 while the minimum confidence value is kept same as %20.

As a result of the second analysis, 52 association rules are obtained in the data set. These rules are shown in the table 6.2.

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule 0	0.05	0.665	3.768	CHEESE	<---	[OLIVE]
rule 1	0.05	0.284	3.768	OLIVE	<---	[CHEESE]
rule 2	0.051	0.493	1.603	DAIRY PRODUCTS	<---	[BAKERY PRODUCTS, FRUITS - VEGETABLES]
rule 3	0.051	0.567	1.525	FRUITS - VEGETABLES	<---	[BAKERY PRODUCTS, DAIRY PRODUCTS]
rule 4	0.051	0.396	1.819	BAKERY PRODUCTS	<---	[DAIRY PRODUCTS, FRUITS - VEGETABLES]
rule 5	0.053	0.429	1.395	FRUITS - VEGETABLES	<---	[BAKERY PRODUCTS, SNACKS]
rule 6	0.053	0.497	1.336	BAKERY PRODUCTS	<---	[SNACKS, FRUITS - VEGETABLES]
rule 7	0.053	0.413	1.139	SNACKS	<---	[BAKERY PRODUCTS, FRUITS - VEGETABLES]
rule 8	0.054	0.523	1.407	DAIRY PRODUCTS	<---	[SNACKS, FRUITS - VEGETABLES]
rule 9	0.054	0.434	1.989	FRUITS - VEGETABLES	<---	[SNACKS, DAIRY PRODUCTS]
rule 10	0.054	0.519	1.431	SNACKS	<---	[DAIRY PRODUCTS, FRUITS - VEGETABLES]
rule 11	0.054	0.414	1.113	FRUITS - VEGETABLES	<---	[HOT DRINKS]
rule 12	0.057	0.284	1.28	MEAT PRODUCTS	<---	[COLD DRINKS]
rule 13	0.057	0.256	1.280	COLD DRINKS	<---	[MEAT PRODUCTS]
rule 14	0.065	0.298	1.34	DAIRY PRODUCTS	<---	[COLD DRINKS]
rule 15	0.065	0.292	1.34	COLD DRINKS	<---	[DAIRY PRODUCTS]
rule 16	0.066	0.506	1.397	SNACKS	<---	[HOT DRINKS]
rule 17	0.072	0.323	1.05	BAKERY PRODUCTS	<---	[COLD DRINKS]
rule 18	0.072	0.233	1.05	COLD DRINKS	<---	[BAKERY PRODUCTS]
rule 19	0.073	0.337	1.687	DAIRY PRODUCTS	<---	[MEAT PRODUCTS]
rule 20	0.073	0.368	1.687	MEAT PRODUCTS	<---	[DAIRY PRODUCTS]
rule 21	0.074	0.42	1.365	CHEESE	<---	[BAKERY PRODUCTS]
rule 22	0.074	0.241	1.365	BAKERY PRODUCTS	<---	[CHEESE]
rule 23	0.077	0.355	2.014	CHEESE	<---	[DAIRY PRODUCTS]
rule 24	0.077	0.439	2.014	DAIRY PRODUCTS	<---	[CHEESE]
rule 25	0.078	0.39	2.212	CHEESE	<---	[MEAT PRODUCTS]
rule 26	0.078	0.442	2.212	MEAT PRODUCTS	<---	[CHEESE]
rule 27	0.078	0.216	1.224	CHEESE	<---	[SNACKS]
rule 28	0.078	0.444	1.224	SNACKS	<---	[CHEESE]
rule 29	0.08	0.215	0.97	FRUITS - VEGETABLES	<---	[COLD DRINKS]
rule 30	0.08	0.361	0.97	COLD DRINKS	<---	[FRUITS - VEGETABLES]
rule 31	0.081	0.224	1.119	MEAT PRODUCTS	<---	[SNACKS]
rule 32	0.081	0.406	1.119	SNACKS	<---	[MEAT PRODUCTS]
rule 33	0.082	0.409	1.33	BAKERY PRODUCTS	<---	[MEAT PRODUCTS]
rule 34	0.082	0.266	1.33	MEAT PRODUCTS	<---	[BAKERY PRODUCTS]
rule 35	0.09	0.412	1.34	DAIRY PRODUCTS	<---	[BAKERY PRODUCTS]
rule 36	0.09	0.292	1.34	BAKERY PRODUCTS	<---	[DAIRY PRODUCTS]
rule 37	0.09	0.243	1.375	CHEESE	<---	[FRUITS - VEGETABLES]
rule 38	0.09	0.511	1.375	FRUITS - VEGETABLES	<---	[CHEESE]
rule 39	0.098	0.489	1.315	FRUITS - VEGETABLES	<---	[MEAT PRODUCTS]
rule 40	0.098	0.263	1.315	MEAT PRODUCTS	<---	[FRUITS - VEGETABLES]
rule 41	0.102	0.283	1.297	DAIRY PRODUCTS	<---	[SNACKS]
rule 42	0.102	0.47	1.297	SNACKS	<---	[DAIRY PRODUCTS]
rule 43	0.103	0.474	1.274	DAIRY PRODUCTS	<---	[FRUITS - VEGETABLES]
rule 44	0.103	0.278	1.274	FRUITS - VEGETABLES	<---	[DAIRY PRODUCTS]
rule 45	0.107	0.294	0.957	BAKERY PRODUCTS	<---	[SNACKS]
rule 46	0.107	0.347	0.957	SNACKS	<---	[BAKERY PRODUCTS]
rule 47	0.108	0.298	1.343	COLD DRINKS	<---	[SNACKS]
rule 48	0.108	0.487	1.343	SNACKS	<---	[COLD DRINKS]
rule 49	0.124	0.341	0.917	FRUITS - VEGETABLES	<---	[SNACKS]
rule 50	0.124	0.332	0.917	SNACKS	<---	[FRUITS - VEGETABLES]
rule 51	0.128	0.345	1.123	FRUITS - VEGETABLES	<---	[BAKERY PRODUCTS]
rule 52	0.128	0.418	1.123	BAKERY PRODUCTS	<---	[FRUITS - VEGETABLES]

Table 6.2 Association rules for product groups.

Some new association rules created for the product groups are given as;

Rules for OLIVE => CHEESE

Possibility of seen together of OLIVE and CHEESE is %5 and

Customers of OLIVE get CHEESE with %66.5 probability.

Rules for CHEESE => OLIVE

Possibility of seen together of CHEESE and OLIVE is %5 and

Customers of CHEESE get OLIVE with %28.4 probability.

Rules for BAKERY PRODUCTS, FRUITS-VEGETABLES => DAIRY PRODUCTS

Possibility of purchasing together of BAKERY PRODUCTS, FRUITS-VEGETABLES and DAIRY PRODUCTSis %5.1 and

Customers of BAKERY PRODUCTS, FRUITS-VEGETABLES get DAIRY PRODUCTS with %49.3 probability.

7. CONCLUSIONS AND RECOMMENDATIONS

In this thesis, association rules and frequently used data mining techniques were disscussed in details. Furthemore, basic algorithms used for association rules were given in details.

Additionally, several ARM methods are summarized and a comprehensive literature review was given where pros and cons of each approach is highlighted. Considering the findings of review section, Apriori algorithm was selected as the most suitable tool for ARM.

It was applied on the receipts collected from one of the supermarket chains operating in the retail sector in Turkey. The aim of the thesis is to find meaningful relationships between products. Taking advantage of this relationship, identification of effective advertising campaigns, and promotions as well as arrangement of shelf layoutwas made possible. This may also increase the sales volume and sales revenue of the retail store.

Receipts collected during 6 months from the market in accordance with association rules were transferred to a computer program for analysis. For analysis, KNIME program which is widely used for ARM was preferred.

After examining the distribution of products, association rules were found with scanning the dataset with Apriori Algorithm. The operation was performed on product groups and related statements were made by presenting obtained rules in tables.

According to the results, FRUITS_VEGETABLES and SNACKS were purchased more frequently than the other product groups. Knowledge obtained as a result of this work can be used to make more effective shelf arrangement or organize different product campaigns. Products which have higher confidence value than the other products must be placed on shelves closely such as COLD DRINKS and SNACKS, DAIRY PRODUCTS and FRUITS AND VEGETABLES, DAIRY PRODUCTS and SNACKS.

The top six highest confidence levels belonged to COLD DRINKS, SNACKS, DAIRY PRODUCTS, FRUITS AND VEGETABLES, BAKERY PRODUCTS and MEAT PRODUCTS. Table 7.1 shows the confidence matrix of these six product groups.

		Items					
Consequent	**Confidence**	COLD DRINKS	SNACKS	DAIRY PRODUCTS	FRUITS VEGETABLES	BAKERY PRODUCTS	MEAT PRODUCTS
	COLD DRINKS		%29.8				%25.6
	SNACKS	%48.7		%47	%33.2	%34.7	%40.6
	DAIRY PRODUCTS		%28.3		%27.8		%33.7
	FRUITS VEGETABLES		%34.1	%47.4		%41.8	%48.9
	BAKERY PRODUCTS				%41.8		%40.9
	MEAT PRODUCTS	%28.4	%22.4	%36.8	%26.3	%26.6	

Table 7.1 Confidence matrix of top six product groups

According to results, we may propose a shelf layout model for 300 square meters store. Also, Top 6 products can be placed in the store as depicted in Figure 7.1.

Figure 7.1 Shelf layout in 300 square meters store

Thus, store managers can check more frequently and keep an appropriate level of stocks of these products that establish the obtained relations.

At this point, the constraints of this study should not be overlooked such as, performing of this study with products in the store where the research was done, collecting of data during a specific time interval will give different results in different periods, affecting purchases in period of collected of receipts by campaigns and promotions of products,

varying of shopping habits due to target audiences and region of the markets where the study was performed.

In future research, a dataset including product groups and customer profile may be studied so that an individual marketing strategy for each customer can be generated. Association rules obtained for products can be grouped according to the customer profile. In this way, each customer interested in different products is targeted by obtains by establishing different campaigns.

REFERENCES

Abdullah, U., Ahmad, J., Ahmed A., 2008. Analysis of Effectiveness of Apriori Algorithm In Medical Billing Data Mining 2008 International Conference on Emerging Technologies 18-19 October, Rawalpindi, 327-331.

Ada, N,. 2012. Regresyon Analizi-Bilgisayar Destekli İstatistiki Yöntemler, Access Date: 10.02.2016. http://ormanweb.sdu.edu.tr/dersler/scarus/regresyon.pdf.

Agarwal, R. C., Aggarwal, C. C., Prasad, V. V. V., 2001. A tree projection algorithm for generation of frequent item sets. Journal of parallel and Distributed Computing, 61(3), 350-371.

Aggelis, V., Christodoulakis, D., 2003. Association Rules and Predictive Models for e-Banking Services" , 1st Balkan Conference on Informatics, Salonica.

Agrawal, R., Srikant, R., 1994. Fast Algorithms For Mining Association Rules, In Proceedings of 20th International Conference on Very Large Data Bases, 487–499.

Ahmed, A. M., Norwawi, N. M., Ishak, W., 2009. Identifying Student and Organization Matching Pattern Using Apriori Algorithm for Practicum Placement, International Conference on Electrical Engineering and Informatics,5-7 August , Selangor, 28 – 31.

Akpınar, H., 2000. Veri Tabanlarında Bilgi Keşfi Ve Veri Madenciliği, Istanbul University Faculty of Management Journal, 29, 1-22.

Angeline, M. D., 2013. Association Rule Generation for Student Performance Analysis Using Apriori Algorithm, The Standard International Journals (The SIJ) Transactions on Computer Science Engineering & its Applications (CSEA), 1(1), 12-16.

Anonym, 2014. Bayes Theorem Access Date: 12.02.2016. http://www.milefoot.com/math/stat/prob-bayes.htm.

Borgelt, C., 2004. Frequent Pattern Mining , Intelligent Data Analysis and Graphical Models Research Unit European Centre for Soft Computing c/ Gonzalo Guti´errez Quir´os s/n, 33600 Mieres, Spain.

Borgelt, C., 2012. Frequent Item Set Mining, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 437-456.

Borgelt, C., Wang, X., 2009. SaM: A Split and Merge Algorithm for Fuzzy Frequent Item Set Mining. Proc. 13th Int. Fuzzy Systems Association World Congress and 6th Conf. of the European Society for Fuzzy Logic and Technology, Lisbon, 968-973.

Bremermann, H. J., 1958. The evolution of intelligence. The nervous system as a model of its environment, Technical Report No. 1, University of Washington, Department of Mathematics, Seattle, WA.

Brin, S., Motwani, R., Ullman, J. D., Tsur, S., 1997. Dynamic Itemset Counting And Implication Rules For Market Basket Data, In ACM SIGMOD Record, 26(2), 255-264.

Bükey, F. Ö., 2014. Data Mining Applications In Customer Relationship Management And A Comparative, Marmara University, Department of Industrial Engineering, Ph.D. THESIS, 172p, Istanbul.

Carbone, P., 2000. What Is The Origin Of Data Mining? Access Date: 05.02.2016. http://www.taborcommunications.com/dsstar/00/1031/102347.html.

Cayiroglu, I., 2014. İleri Algoritma Analizi-5 Yapay Sinir Ağları, Acces Date: 15.02.2016. http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-5.Hafta-YapaySinirAglari.pdf.

Chai, K., Chieu, H., Ng, H. B., 2002. Bayesian Online Classifiers for Text Classification and Filtering, Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, 97-104.

Chen, R. S., Wu, R. C., Chang, C. C., Chen, J. Y., 2005. Data Mining Application in Customer Relationship Management Of Credit Card Business, In Proceedings of 29th Annual International Computer Software and Applications Conference, Washington, 39-40.

Duneja, E., Sachan, A. K., 2012. A Survey on Frequent Itemset Mining with Association Rules, International Journal of Computer Applications 46(23), 18-24.

Duru, N., 2005. An Application of Apriori Algorithm on a Diabetic Database, Conference: Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 398-404.

Fraser, A. S., 1957. Simulation of genetic systems by automatic digital computers. II: Effects of linkage on rates under selection, Austral. J. Biol. Sci. 10: 492–499.

Gancheva, V., 2013. Market Basket Analysis Of Beauty Products, Erasmus University Rotterdam, Erasmus School of Economics,M.Sc. Thesis, 94p, Rotterdam.

Grabmeier, J., Rudolph, A. 2002. Techniques of Cluster Algorithms in Data Mining, Data Mining and Knowledge Discovery, 6, 303-360.

Grahne, G., Zhu, J., 2003. Efficiently Using *Prefix*-trees in Mining Frequent Itemsets. In IEEE ICDM'03 Workshop FIMI'03, Melbourne, Florida, USA.

Gupta, B., Garg, D., 2011. A Taxonomy of Classical Frequent Item set Mining Algorithms. International Journal of Computer and Electrical Engineering, 3(5), 695-699.

Gupta, N., Mangal, N., Tiwari, K., Mitra, P., 2006. Mining Quantitative Association Rules in Protein Sequences, In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining –AUSDM., Berlin, 273-281.

Gürgen, G., 2008. Birliktelik Kuralları İle Sepet Analizi Ve Uygulaması. Marmara University, Social Sciences, M.Sc. Thesis, 104, Istanbul.

Han, J., Pei, J., Yin, Y., 2000. Mining Frequent Patterns without Candidate Generation. In Proc. ACM SIGMOD Intl. Conference on Management of Data., New York, 1-12.

Han, J., Pei, J., Feng, X., 2004. From Sequential Pattern Mining to Structured Pattern Mining: A Pattern-Growth Approach. Yan University of linois at Urbana-Champaign, Urbana, IL 61801, U.S.A. State University of New York at. Buffalo 19(3), 257-279.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., Hsu, M. C., 2000. FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining. In Proc. 2000 ACM SIGKDD Int. Conf. Know ledge Discovery in Databases (KDD'00), Boston, MA, 355-359.

Hidber, C., 1999. Online association rule mining. In Proc. Of the 1999 ACM SIGMOD International Conference on Management of Data, 28(2), 145–156.

Holder, L. B., Cook, D. J., Djoko, S., 1994. Substructure Discovery In The Subdue System, In: Proceeding of the AAAI'94 workshop knowledge discovery in databases (KDD'94), Seattle, WA, 169–180.

Holland, J. H., 1975. Adaptation in Natural and Artificial Systems, University of Michigan Press, 183, Michigan.

Huang, L. J., 2007. FP-growth Apriori algorithm's Application in the Design for Individualized Virtual Shop on the Internet, In Machine Learning and Cybernetics, 2007 International Conference , Vol. 7, Hong Kong, 19-12 August, 3800-3804.

Jajodia, S., Wijesekera, D., 2005. Data and Applications Security XIX: 19th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Storrs, CT, USA, 7-10 August.

Jiang, M. H., 2011. Mining Mobile Telecommunication Services Using Apriori Algorithm, Advanced Materials Research, 2369-2373, 2011.

Kamrul, S., Mohammad, K., Hasnain, A., 2008. Reverse Apriori Algorithm for Frequent Pattern Mining, Asian Journal of Information Technology, 524-530.

Ketkar, N. S., Holder, L. B., Cook, D. J., 2005. Subdue: Compression-Based Frequent Pattern Discovery in Graph Data, The First International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations, Chicago, 21-24 August, 71-76.

Khanna, S. O., Sherdiwala, K. B., 2015. Association Rule Mining: An Overview Access Date: 15.02.2016, http://oaji.net/articles/2015/1250-1428333425.pdf.

King, R., Srinivasan, A., Dehaspe, L., 2001. WARMR: A Data Mining tool for chemical data. J. of Computer-Aided Molecular Design , 15,173–181.

Kupar, V., Steinbach, M., Tan, P. N., 2005. Introduction To Data Mining. Pearson Addison Wesley, 769p, London.

Kuramochi, M., Karypis, G., 2004. An Efficient Algorithm for Discovering Frequent Subgraphs, IEEE Trans. Knowl. Data Eng., 16(9), 1038-1051.

Ji, L., Zhang, B., Li, J., 2006. A New Improvement on Apriori Algorithm, Computational Intelligence and Security, 2006 International Conference on. Vol. 1. IEEE, 840-844.

Li, R., 2015. History Of Data Mining. Access Date: 05.02.2016. http://rayli.net/blog/data/history-of-data-mining/.

Liu, D. S., Fan, S. J., 2013. Tourist Behavior Pattern Mining Model Based on Context, Hindawi Publishing Corporation Discrete Dynamics in Nature and Society, (2013),12.

Liu, J., Wang, K., Tang, L., Han, J., 2002. Top down fp-growth for association rule mining , Springer Berlin Heidelberg, 334-340.

Malerba, D., Esposito, F., Lisi, F., Appice, A., 2001. Mining Spatial Association Rules, In Census Data. In Proceedings of Joint Conf. on "New Techniques and Technologies for Statistcs and Exchange of Technology and Know-how". 541-550.

Manjunath, K. V., 2015. Data Mining Techniques for Anti Money Laundering, International Journal of Advanced Research in Science, Engineering and Technology, 2(8), 819-823.

Mu, Z., Yi, C., Xiaohong, Z., Junyong, L., 2009. Study On The Recommendation Technology For Tourism Information Service, In Computational Intelligence and Design, ISCID'09. Second International Symposium Vol. 1, 12-14 Dec, Changsha, 410-415.

Neelima, S., Satyanarayana, N., Murthy, K. P. A., 2014. Survey on Approaches for Mining Frequent Itemsets. IOSR Journal of Computer Engineering (IOSRJCE), 16(4), 31-34.

Palagin, D., 2009. Mining Quantitative Association Rules in Practice. Royal Institute of Technology, School of Computer Science and Communication, M.Sc. Thesis, 73p, Stockholm.

Park, J. S., Chen, M. S., Yu, P. S., 1995. An Effective Hash-based Algorithm For Mining Association Rules, SIGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data, New York. 175-186.

Pei, J., Han, J., Mortazavi-Asl, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M., 2001. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. 17th International Conference on Data Engineering (ICDE), Washington, 215.

Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., Yang, D., 2007. H-Mine: Fast And Space Preserving Frequent Pattern Mining In Large databases. IIE transactions, 39(6), 593-605.

Prakash, S., Parvathi, R. M. S., 2010. An enhanced Scalling Apriori for Association Rule Mining Efficiency, European Journal of Scientific Research, 39, 257-264.

Revathi, R., Geetha, M., 2015. Re-Modified Apriori Algorithm in E-Commerce Recommendation System, International Journal of Innovative Research in Computer and Communication Engineering, 3(7), 6737- 6744.

Savas, S., Topaloglu, N., Yılmaz, M., 2012. Veri Madenciliği ve Türkiye'deki Uygulama Örn., İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21, 1-23.

Savasere, A., Omiecinski, E., Navathe, S., 1995. An Efficient Algorithm Formining Association rules in large databases, Proceedings of the 21th International Conference on Very Large Data Bases, San Francisco, 432-444.

Serban, G., Campan, A., Czibula, I. G., 2006. A Programming Interface For Medical diagnosis Prediction. Studia Universitatis, "Babes-Bolyai", Informatica, LI(1), 21-30.

Shao, X., 2009. The application of improved 3dapriori three dimensional association rules algorithm in reservoir data mining, Computational Intelligence and Security CIS '09. International Conference on , 11-14 Dec, Beijing, 64-68.

Sharif, M. N. A., Ching, N. M., Bakri, A., Zakaria, N. H., 2005. Using a Priori Algorithm for Supporting an e-Commerce System, Journal of Information Technology Impact, 5(3), 129-138.

Sharma, V., Sufyan, B. M. M., 2012. A Probabilistic Approach to Apriori Algorithm, International Journal of Granular Computing, Rough Sets and Intelligent Systems, 2(3), 225-243.

Shearer, C., 2000. The Crisp-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, 5, 13-23.

Slimani, T., 2014. Class Association Rules Mining based Rough Set Method, International Journal of Engineering and Technology (IJET), 6, 2786-2794.

Srikant, R., Agrawal, R., 1996. Mining sequential patterns: Generalizations and performance improvements. In Proc. 5th Int. Conf. Extending Database Technology (EDBT'96), Avignon, France, 3-17.

Strickland, J., 2015. What the heck are Association Rules in Analytics? Access Date: 20.02.2016.
http://bicorner.com/2015/07/22/what-the-heck-are-association-rules-in-analytics/

Tarafder, K. A., Khaled, S. M., Islam, M. A., Islam, K. R., Feroze, H., Ferdaus, A. A., 2008. Reverse Apriori Algorithm for Frequent Pattern Mining, Asian Journal of Information Technology, 7(12), 524-530.

Tarı, R., 2002. Ekonometri, Alfa Yayını, 403, Istanbul.

Toivonen, H., 1996. Sampling Large Databases For Association Rules, In 22th International Conference on Very Large Databases (VLDB'96), Bombay, 134-145.

Voznika, F., 2007. Viana L. Data Mining Classification. Access Date: 10.02.2016.
https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

Wang, E. T., Chen, A. L., 2009. A Novel Hash-based Approach For Mining Frequent Itemsets Over Data Streams Requiring Less Memory Space, Data Mining and Knowledge Discovery, 19(1), 132-172.

Wang, C., Tjortjis, C., 2004. PRICES: An efficient algorithm for mining association rules, in Lecture Notes Computer Science, 3177, 352-358.

Wang, H., Ji, X., Xue, Y., Liu, X., 2010. Applying Fast-Apriori Algorithm to Design. Data Mining Engine, In Proc. of International Conference on System Science, Engineering Design and Manufacturing Informatization, Yichang, 12-14 Nov., 63-65.

Washio, T., Motoda, H., 2003. State Of The Art of Graphbased Data Mining, ACM SIGKDD Explorations Newsletter, 5(1), 59-68.

Wu, H., Lu, Z., Pan, L., Xu, R., Jiang, W., 2009. An Improved Apriori-based Algorithm for Association Rules Mining. 6th International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, 4-16 August, 51-55.

Xiaohui, L., 2012. Improvement of Apriori algorithm for association rules. World Automation Congress (WAC), 24-28 June, Mexico, 1-4.

Xu, Z., Zhang, R., 2009. Financial revenue analysis based on association rules mining, Computational Intelligence and Industrial Applications. PACIIA 2009. Asia-Pacific Conference on, Hangzhou, 28-29 Nov, 220 – 223.

Yalcin, N., 2014. Genetik Algoritmalar, Access Date: 14.02.2016. http://bm.bilecik.edu.tr/Dosya/Arsiv/duyuru/genetik_algoritmalar.pdf

Yan, X., Han, J., 2002. Gspan: Graph-based substructure pattern mining. In ICDM'02: 2nd IEEE Conf. Data Mining, Urbana, 721-724.

Yang, G. X., 2013. The Research of Improved Apriori Mining Algorithm in Bank Customer Segmentation, Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering, 22-23 March , Zhengzhou, 3174- 3177.

Zaki, M. J., Ogihara, M., Parthasarathy, S., Li, W. 1996. Parallel data mining for association rules on shared-memory multi-processors. InSupercomputing, Proceedings of the 1996 ACM/IEEE Conference on, 43-43.

Zaki, M. J., 2000. Scalable algorithms for association mining. Knowledge and Data Engineering, IEEE Transactions on, 12(3), 372-390.

Zaki, J. M., 2001. SPADE: An efficient algorithm for mining frequent sequences. Journal Machine Learning, 42(1-2), 31-60.

Zeng, Z., Yang, H., Feng, T., 2011. Using HMT and HASH_TREE to Optimize Apriori Algorithm, International Conference on Business Computing and Global Informatization, 29-31 July, Shanghai 412-415.

Zhang, W., Ma, D., Yao, W., 2014. Medical Diagnosis Data Mining Based on Improved Apriori Algorithm, Journal Of Networks, 9(5), 1339-1345.

CURRICULUM VITAE (CV)

Name Surname : Pınar YAZGAN

Birth Place And Date: ISTANBUL, 24/10/1990

Marital Status : (Single)

Foreign Language : English, Spanish

E-mail : pinar.yazgan@iticu.edu.tr

Education

High School : Istanbul Köy Hizmetleri Anadolu Lisesi, 2009

Bachelor : Bahceşehir University, Faculty Of Engineering, Software Eng.

Master : Istanbul Commercial University,
Institute of Natural Science, Industrial Engineering Dept.

Work Experience

ACRON Bilişim 2014

OBASE Bilgisayar ve Danışmanlık Hizm. Tic. A.Ş. 2014--…(continue)

Publications

Oguz, B., Isitman, A. C., Bayrak, M., Yazgan, P., 2012. Mobilized Patient Record Management Systems, 2, 235-238.