**REPUBLIC OF TURKEY**
**YILDIZ TECHNICAL UNIVERSITY**
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# DATA ANALYTICS IN INDUSTRIAL APPLICATIONS

**ÖZGÜRDENİZ DÖĞER**

**MSc. THESIS**
**DEPARTMENT OF INDUSTRIAL ENGINEERING**
**PROGRAM OF SYSTEMS ENGINEERING**

**ADVISER**
**ASSOC. PROF. DR. VİLDAN Ç. ÖZKIR**

**İSTANBUL, 2018**

**REPUBLIC OF TURKEY**

**YILDIZ TECHNICAL UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# DATA ANALYTICS IN INDUSTRIAL APPLICATIONS

A thesis submitted by Özgürdeniz DÖĞER in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 22.06.2018 in Department of Industrial Engineering, Systems Engineering Program.

**Thesis Adviser**

Assoc. Prof. Dr. Vildan Ç. ÖZKIR

Yıldız Technical University

**Approved By the Examining Committee**

Assoc. Prof. Dr. Vildan Ç. ÖZKIR

Yıldız Technical University _____

Asst. Prof. Dr. Ganimet Nilay YÜCENUR, Member

Beykent University _____

Asst. Prof. Dr. Nezir AYDIN, Member

Yildiz Technical University _____

## ACKNOWLEDGEMENTS

Firstly, I'd like to thank my advisor Assoc. Prof. Dr. Vildan Özkır who has shared her knowledge and experience. She is always kind and supportive for her students.

Also, I'd like to thank my previous advisor Assoc. Prof. Dr. Tuğba Efendigil who helped me lot between the beginning of the thesis and middle of the thesis.

I would like to also thank to all my other teachers of my master degree education, they are all kind and I'm lucky being student of those.

Thanks to my parents who have been supporting me and giving their love whole my life.

June, 2018

Özgürdeniz DÖĞER

## TABLE OF CONTENTS

Page

LIST OF SYMBOLS ..... vi

LIST OF ABBREVIATIONS ..... viii

LIST OF FIGURES ..... x

LIST OF TABLES ..... xii

ABSTRACT ..... xiiii

ÖZET ..... xiiiii

CHAPTER 1

INTRODUCTION ..... 1

1.1 Literature Review ..... 2
1.2 Objective of the Thesis ..... 10
1.3 Hypothesis ..... 10

CHAPTER 2

METHODS ..... 11

2.1 Clustering Methods ..... 11
2.1.1 Distance Measures ..... 13
2.1.2 Expectation Maximization Algorithm ..... 14
2.1.3 K-Means Algorithm ..... 16
2.1.4 Hierarchical Clustering Algorithm ..... 19
2.2 Classification Methods ..... 22
2.2.1 Decision Tree ..... 22
2.2.2 Naïve Bayes ..... 25
2.2.3 Support Vector Machine ..... 26
2.2.4 Evaluation of Classification Results ..... 27

CHAPTER 3

PROPOSED METHODOLOGY ..... 31

3.1 Data Exploration Phase ..... 32
3.2 Data Preparation Phase ..... 32
3.3 Clustering Phase ..... 33

3.4 Classification Phase ..... 34
3.5 Preparing Decision Support Tool ..... 35

CHAPTER 4

APPLICATION ..... 36

4.1 Data Exploration Phase ..... 36
4.2 Data Preparation Phase ..... 36
4.3 Clustering Phase ..... 39
4.3.1 Expectation Maximization ..... 19
4.3.2 K Means Algorithm ..... 41
4.3.3 Hierarchical Clusterer ..... 42
4.3.4 Summary of Clustering ..... 43
4.4 Classification Phase ..... 44
4.4.1 Support Vector Classifier and Prediction Results ..... 45
4.4.2 Naive Bayes Classification and Prediction Results ..... 46
4.4.3 C 4.5 Classification and Prediction Results ..... 48
4.5 Preparing Decision Support Tool ..... 49

CHAPTER 5

CONCLUSION ..... 50

REFERENCES ..... 53

CURRICULUM VITAE ..... 53

## LIST OF SYMBOLS

| | |
|---|---|
| acc | Accuracy |
| d(i,j) | Distance |
| err | Error Rate |
| fm | F-Measure |
| fn | False negative |
| fp | False positive |
| n | Number of diversified aim values |
| P | Number of diversified aim values |
| p | Precision |
| p | Observation values |
| r | Recall |
| $S_i$ | Class |
| sn | Sensitivity |
| sp | Specificity |
| tn | True negative |
| tp | True positive |
| w | Weight |
| $x_j$ | Attribute |
| $\theta_k$ | Weight of each distribution |
| $\Pi_k$ | Weight of each distribution |
| $\widehat{\mu_k}$ | The data points for the means |
| $\widehat{\sigma}_k^2$ | Overall data variance |
| D() | Distance function |
| β | Coefficient |
| α | Coefficient |
| γ | Coefficient |
| μ | Vector |

| | |
|---|---|
| $\langle \cdot, \cdot \rangle$ | Dot product |
| $\xi_i, \xi_i^*$ | Slack Variables |

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ANFIS | Adaptive Neuro-Fuzzy Inference System |
| B2B | Business to Business |
| CA | Conjoint Analysis |
| CART | Classification and Regression Tree |
| CRM | Customer Relationship Management |
| DSC | Differential Scanning Calorimetry |
| DT | Decision Tree |
| ELM | Extreme Learning Machine |
| EM | Expectation Maximization |
| FA | Factor Analysis |
| GBM | Gradient Boosting Machines |
| GBTs | Gradient-Boosted Trees |
| GM | Geometric-Mean |
| HC | Hierarchical Clustering |
| IT | Information Technologies |
| LCA | Latent Class Analysis |
| LMT | Logistic Model Tree |
| LR | Logistic Regression |
| LRFM | Length, Recent transaction time, Buying Frequency and Monetary |
| LSSVM | Least Square Support Vector Machine |
| LSVM | Linear Support Vector Machine |
| MLR | Multi Lineer Regressor |
| MSE | Mean Square Error |
| RBF | Radial Basis Function |
| RF | Random Forest |
| RFM | Recency, frequency and monetary |
| SDP | Software Defect Prediction |

| | |
|---|---|
| SOM | Self Organizing Map |
| SEER | Surveillance, Epidemiology, and End Results |
| SVC | Support Vector Clustering |
| SVR | Support Vector Regressor |
| TSS-RFE-MRMR | Time Series Segmentation, Recursive Feature Elimination and Minimum Redundancy Maximum Relevance |

## LIST OF FIGURES

Page

Figure 2.1 An example of a dendrogram in hierarchical clustering…….....19
Figure 2.2 Decision Tree……..23
Figure 3.1 Process Flow Chart…….....31
Figure 4.1 Discretized Data Instance……..….....38
Figure 4.2 Cluster members distribution of Expectation Maximization algorithm….....40
Figure 4.3 Expectation Maximization Visualization of Scatter……40
Figure 4.4 Cluster members distribution of k means algorithm……41
Figure 4.5 K Means Visualization of Scatter …….....42
Figure 4.6 Cluster members distribution of hierarchical clusterer……43
Figure 4.7 Hierarchical Clustering Visualization of Scatter ……..43

## LIST OF TABLES

Page

Table 1.1 Literature Review for Customer Segmentation……..…….…….4
Table 1.2 Classification and Prediction…….8
Table 2.1 Confusion Matrix for Binary Classification…….……...28
Table 2.2 Threshold Metrics for Classification Evaluations…….……..29
Table 4.1 Example of initial dataset ……...…….…...37
Table 4.2 Example of pre-processed dataset …….…....37
Table 4.3 Discretized Pre-processed dataset which includes 320 resellers and 98…..…38
Table 4.4 Cluster member numbers and percentage of Expectation Maximization Algorithm…….……...39
Table 4.5 Cluster member numbers and percentage of K Means Algorithm…….….41
Table 4.6 Cluster member numbers and percentage of hierarchical clusterer……..…42
Table 4.7 Evaluation of Training Dataset with Support Vector Classifier……..45
Table 4.8 The Result of Prediction Model with Support Vector Classifier……45
Table 4.9 Evaluation of Training Dataset with Naïve Bayes Classification……...46
Table 4.10 The Result of Prediction Model with Naïve Bayes Classification…….…...47
Table 4.11 Evaluation of Training Dataset with C4.5 Classification…….….48
Table 4.12 The Result of Prediction Model with C4.5 Classification…….…48

# ABSTRACT

## DATA ANALYTICS IN INDUSTRIAL APPLICATIONS

Özgürdeniz DÖĞER

Department of Industrial Engineering

MSc. Thesis

Advisor: Assoc. Dr. Vildan ÖZKIR

Information Technology Distributors need to develop methods for understanding their dealers purchasing behavior regarding to product groups in order to prepare their business plan correctly. So that correct customer segmentation is seriously important for those companies. This study provides customer management framework that is provided in order to increase IT distributors sales. Basicly, this framework includes 3 phases which are data preparation, effective clustering and prediction. Real application of proposed method is provided for one of the biggest IT distibutor company which is in Istanbul. Firstly, K-means, expectation maximization algorithm (EM) and hierarchical clustering methods ,which are proper for the data set, are applied IT dealers sales data. As a result of clustering process, it is observed that hierarchical clustering provides better performance compared to others. This study also provides decision support tool that suggests right product groups in order to increase IT distributors sales with classification method. Proposed method uses clustering results as preprocess of prediction process. Sales potential of dealers is determined regarding to pruduct groups with classification algorithms that are Naive Bayes, Support Vector Classification and C4.5 algorithms. Finally the business plans that are created via proposed framework have potential to yield results.

**Keywords:** Data mining, clustering, classification, customer segmentation, prediction, decision support tool

# ÖZET

## ENDÜSTRİYEL UYGULAMALARDA VERİ ANALİTİĞİ

Özgürdeniz DÖĞER

Endüstri Mühendisliği Anabilim Dalı
Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. Vildan ÖZKIR

Bilgi Teknolojileri (BT) Distribütörleri, iş planlarını doğru bir şekilde hazırlayabilmeleri için bayilerinin ürün gruplarına göre satın alma karakteristiklerini anlamak ve çözümlemek üzere yaklaşımlar geliştirmeye ihtiyaç duymaktadır. Dolayısıyla, bu firmalar için müşterilerini doğru bir şekilde bölümlendirmek ciddi bir öneme sahiptir. Bu çalışma ile, distribütör firmaların satışlarını arttırması amacına hizmet edebilecek bir müşteri yönetimi çerçevesi sunulmaktadır. Bu çerçeve temel olarak; veri hazırlama, etkin kümeleme ve tahmin aşaması olmak üzere üç aşamadan oluşmaktadır. Önerilen çerçevenin gerçek bir uygulaması, İstanbul'da bulunan büyük bir BT distribütörün müşterilerinin segmentasyonu için sunulmaktadır. Öncelikle, problemin yapısına veriye uygun olan kümeleme yöntemlerinden K–ortalama, beklenti maximizasyonu ve hiyerarşik kümeleme metotları BT bayilerinin satış verilerine uygunlanmaktadır. Kümeleme aşamasının sonucunda, hiyerarşik kümeleme yönteminin diğer yöntemlere göre daha iyi performans gösterdiği gözlenmektedir. Ayrıca, BT distribütörünün satışlarını arttırmak amacıyla sınıflandırma yöntemi kullanılarak bayilerine doğru ürün grubunu önermesi gerektiğine dair bir karar destek sistemi sunmaktadır. Önerilen çerçeve ile, kümeleme sonuçları tahminleme sürecinin ön işleme süreci olarak kullanılmaktadır. Sınıflandırma algoritmaları olan Naif Bayes, Destek Vektör Sınıflandırması ve C4.5 algoritması ile, bayilerin ürün gruplarına göre satış potansiyelleri belirlenmektedir. Sonuç olarak, önerilen çerçeveye göz önünde bulundurularak yapılan iş planlarının yüksek potansiyel üretme olanağı tartışılmaktadır.

**Anahtar kelimeler:** Veri madenciliği, kümeleme, sınıflandırma, müşteri segmentasyonu, tahminleme, karar destek aracı

# CHAPTER 1

## INTRODUCTION

In recent years, the pace of change has increased considerably in the business world. With the evolution of customer relationship management (CRM) and computational power, companies have focused on learning the behaviors of their customers, so that, they may provide intelligent marketing services regarding the characteristics and capabilities of their customers. Intelligent marketing activities have the potential to create high value for both customers and companies.

The analytical CRM means analyzing customers behaviors and self givenness in order to help the company's strategies for customer segmentation. For instance, analytical CRM can support an organization so as to more effectually allocate resources and discriminating better for the most lucrative part of customers [1]. Hence, analytical tools of CRM assist companies in terms of product positioning through specifications of customers. For correctly determining customers' needs and selling the product to the right customer, companies focus on successful customer segmentation practices.

Similar to the retail sector, distributor firms also adopt analytical approaches to manage their customers in a better way. IT (Information Technology) distributors should put these applications into practice more quickly when the rapidity of change and the short lifecycles of (innovative) products are considered in the sector. They sell their products through their dealer channel, which is called as reseller channel in IT market. IT distributors are continuously adding new brands or product groups their product portfolio. They require to know which, resellers have competitive advantage than the others, so that the segmented resellers may perform better in sales of new product groups. An IT distributor's reseller channel is seriously enormous and each reseller has different customer portfolio and sales characteristic. Therefore, when new product group is added in IT distributor's portfolio, they need decision support tool showing that which resellers

are able to sell new product groups, effectively. This decision support tool results in building effective sale channel and managing the sales activities more efficiently.

### 1.1 Literature Review

Customer segmentation is an essential phase for a successful customer relationship management. Customer segmentation is described as operation of splitting heterogeneous customers into homogeneous groups based on prevalent attributes regarding their diversities as well [1].

Segmentation helps companies generate various marketing strategies for each customer segment. These strategies aim at maximizing the total value created for each customer segment and for the company. Customer segmentation increases not only customers' satisfaction but also the expected profits. In addition, it strengthens the customer relationships by building profitable and faithful business interactions.

This section examines segmentation methods in literature which can be summarized in two categories: the first category focuses on statistical analysis approaches with varying data types (demographic, geographic, psychographic and behavioral), the second category focuses on clustering algorithms to create customer segmentations.

Firstly, statistical analysis approaches provide dividing base of customer into groups which are equivalent in particular way related to marketing such as interests, sex, age etc. These approaches include Recency, Frequency and Monetary (RFM), Factor Analysis (FA), Conjoint Analysis (CA), Latent Class Analysis (LCA) etc.

Chan [2] has combined a model with RFM method and genetic algorithm to segment automobile retailer market. The most profitable customers can be determined for special offer programs if the presented methodology is applied. Considering the correlation of clients' values and campaigns is the second pro of recommended methodology. Hwang et al. [3] propose an application of customer segmentation regarding three types of customer value which are customers faithfulness, todays value and futures value by using logistic regression in wireless telecommunication industry. Paker and Vural [4] applied factor analysis to segment customers for marinas. It is critically advised that marinas and marina marketing research is enlarged with extra customer segmentation work, profiling the market and cases of targeting, differentiation chances occasions and their effect on service performance of marina. Chen et al [5] combined factor analysis and k-means

methods in order to segment hot spring hotel industry. Factor analysis and two-step cluster analysis were combined by Gilbao [6] to segment Israeli mall customers. This study revealed that the majority of Israeli customers go to the mall accompanied by others. Since, this study recommends retailers should think of marketing mixes and promotional campaigns addressing groups and not individuals. Martina Rypakovaa et al. [7] present a study which has demographic, psychographic, socio/cultural, user behavior criteria by using chi square for segmenting green customers in Slovakia. Latent class segmentation which is based on conjoint data was proposed by Tabi et al. [8] that focuses on socio-demographic, psychographic and behavioral characteristics for green electricity.

Secondly; clustering analysis approaches provide collecting objects in the same groups which has similar members in anyway, it is natural grouping. These approaches include k-means, hierarchical clustering, expectation maximization, support vector clustering (SVC) etc.

Clustering analysis is unsupervised data processing tool that classifies examples into meaningful groups. So that each cluster is supposed to consist of equivalent data and supposed to be different than other clusters. Purpose of clustering is surely not estimating an aimed class variate, however basically capturing probable spontaneous grouping in dataset. So, the most general application for clustering is exploring the data and finding whole probable clusters in data. Therefore, clustering is generally used to form customer groups which are more likely each other than other groups. The algorithm or user can determine number of the clusters and it depends on the technique.

There are many studies in literature that uses integrated k-means algorithm to create clusters. Li et al. [9] propose k-means clustering model that integrated with LRFM (length, recent transaction time, buying frequency and monetary) for segmenting textile manufacturer's customers. Hong and Kim [10] integrated results of self organizing map (SOM) and k-means with a single model in order to segment customers of a company's online stores.

Liu and Ong [11] propose a model which used genetic algorithm in order to determine of clusters select variables for k-means clustering method that is applied on bank customers. Kuo et al. [12] propose clustering ensemble algorithms which is metaheuristics – based on mobile application market. Those algorithms are K-means clustering ensembles which respectively base genetic algorithm, a particle swarm optimization and an artificial bee

colony optimization. Results of experiment provides that the best algorithm is artificial bee colony-based K-means clustering ensembles for this study.

Dzobo et al. [13] propose a model that uses hierarchical clustering method to segment electricity customers. Cuadros and Domínguez [14] propose a study that uses SOM for segmenting customers in agricultural industry. Huang et al. [15] compared support vector clustering (SVC) performance with k-means and SOM methods by segmenting a drink company's customers. They find out SVC outperforms the others in marketing segmentation. Coussement et al. [16] investigate the effects of data accuracy problem with three segmentation techniques which are decision trees, RFM (recency, frequency and monetary value) and logistic regression. As result of evaluation, they recommend decision trees for customer segmentation, even though there is suspicion of data accuracy issues. Another model is proposed by Murray et al. [17] that uses k-means clustering in order to identify segments of customers with similar demand behaviors in supply chain. In addition, Benítez et al. [18] develop a dynamic clustering method which uses dynamic k-means algorithm that uses fuzzy membership function in order to segment profiles regarding to energy consumption. A soft clustering method is developed which uses latent mixed-class membership clustering approach, which is derived from Dirichlet Allocation, for segmenting an e-commerce company's customers by Wu and Chou [19].

Table 1.1 Literature Review for Customer Segmentation

| **Authors** | **Year** | **Analysis** | **Method** | **Data Characteristic** | **Application Area** |
|---|---|---|---|---|---|
| Chan | 2007 | Statistical | RFM | Demographic | Automobile retailer market |
| Hwang et al. | 2003 | Statistical | Logistic Regression | Demographic | Wireless telecommunication industry |
| Paker and Vural | 2016 | Statistical | FA | Demographic | Marinas |
| Chen et al | 2013 | Statistical | FA & K-means | Demographic | Hot spring hotel industry |
| Tabi et al. | 2014 | Statistical | CA | Demographic | Green electricity |
| Gilboa | 2008 | Statistical | FA & Two Step Cluster Analysis | Demographic | Mall |
| Martina Rypakovaa et al. | 2015 | Statistical | Chi-square test | Demographic | Green customers |

Table 1.1 Literature Review for Customer Segmentation (Cont'd)

| **Authors** | **Year** | **Analysis** | **Method** | **Data Characteristic** | **Application Area** |
|---|---|---|---|---|---|
| Dzobo et al. | 2014 | Data Mining | Hierarchical clustering | Numeric | Electricty customers |
| Cuadros and Domínguez | 2012 | Data Mining | SOM | Numeric | Agricultural industry |
| Huang et al. | 2007 | Data Mining | SVC & k-means & SOM | Numeric | Drink company |
| Liu and Ong | 2005 | Data Mining | K-means clustering integrated wih genetic algorithm | Numeric | Bank |
| Benítez et al. | 2013 | Data Mining | Dynamic k-means clustering | Numeric | Energy consumption |
| Wu and Chou | 2010 | Data Mining | Latent mixed-class membership clustering | Numeric | E-commerce |
| Kuo et al. | 2016 | Data Mining | Metaheuristic based K-means | Numeric | Mobile application |
| Li et al. | 2010 | Data Mining | K-means clustering integrated with LRFM | Numeric | Textile |
| Hong and Kim | 2011 | Data Mining | SOM and k means | Numeric | Online store |
| Coussement et al | 2012 | Data Mining | Decision trees, RFM, logistic regression | Numeric | General mail order catalogs customers |
| Murray et al. | 2015 | Data Mining | K means | Numeric | Supply Chain |

In customer segmentation, classification techniques provide essential information to manage the customer relations since they are able to process an importantly big amount of data. Classification techniques provide a basis to estimate class labels, which are categorical. Also, it classifies data with basing class labels and training dataset. The term may cover any structure in that some prediction and decision made relying on currently existing information. Classification technic is known method for recurrently making this kind of decisions in new circumstances [20].

There are many studies in the literature that focuses on investigating the comparative performance of classification methods related with application area and data types. Chen et al. [21] compared Random Forest (RF), Logistic Model Tree (LMT) and Classification and Regression Tree (CART) by using training data after using linear support vector machine algorithm (LSVM) in order to form an estimate of estimated capability of the 12 landslide conditioning effects. Generally, those three models produce rationally great performances. As results, RF model provides the best predictive ability as regards CART and LMT models.

Zhao et al. [22] compares classification and regression trees (CART) and support vector machine (SVM) models for influential factors of groundwater such as variance of the groundwater and water tank level and also amount of rainfall. This study provides that CARTM model abilities are not only generalization and better fitting, on the other hand it has serious pros in the analyzing process of this study's data characteristics and viewing of significant variables. It is an efficient procedure to estimate ground water level in earth sliding.

Mujtaba et al [23] aim to estimate automatically the reason of mortem via forensic autopsy reports text with comparing several drafts for text classification, term weighing or feature value representation, feature reduction and feature extraction. Most of classification procedure are compared and as end up in various text classification algorithms, support vector machines classifier has advantageous among decision tree, random forest, k-nearest neighbor, ensemble – voted classifier and Naïve Bayes.

Lynch et al. [24] apply a number of supervised learning techniques to SEER program (Surveillance, Epidemiology, and End Results) database for classifying lung cancer sufferer from the point of surviving. This study includes Decision Trees, Support Vector Machines (SVM), linear regression, custom ensemble and Gradient Boosting Machines (GBM). As result, the best performing is the custom ensemble and after that is GBM.

Various studies in literature that are related classification focuses on SVM. Shao et al. [25] use Time Series Segmentation, Recursive Feature Elimination and Minimum Redundancy Maximum Relevance based SVM as abbreviation TSS-RFE-MRMR in order to predict undulation of price also modeling electricity market price without considering additional effect. Cheng and Yu [26] present a model for conducting automatically cloud classification on whole sky pictures and carry out short period solar

irradiance estimation as to the classification outcomes. The all sky pictures are split into blocks for coping with the mixed cloud type provisions. As regards non block based procedure, suggested block based procedure may rise the class rate by 5% - 10%. SVM with radial basis function (RBF) kernel is used for classifying. On the other hand, multi linear regressor and support vector regressor are used for prediction. Gharagheizi et al. [27] present a model which uses least square support vector machine classification as novel technique for determining circumstances under which production happens. The study's results present that application of least square SVM approach can efficiently help completion designers for making well timed sand control plan with minimum breakdown of manufacturing.

There are different studies which combines heuristic algorithms with classification methods. Ahmet and Maheswari [28] present a churn estimation procedure based on metaheuristic which runs churn estimation on enormous telecom data. Firefly algorithms hybridized form is used as a classifier in this study. Mausa and Grbac [29] present a model which uses genetic algorithm for classification in software defect prediction. Uthayakumar et al. [30]propose a method which is a unified framework in order to quantitatively and qualitatively analyze by using Ant Colony Optimization which is successfully performed. As proof of performance analysis, this method is greater than current classifiers that are RBF, logistic regression, random forest and multilayer perceptron with regards to several factors of performance analysis. Moreover, suggested ant miner model is defined as more favorable method for bankruptcy estimation by comparison with other classic artificial intelligence and statistical procedures.

In addition, there are lots of developed models in literature. Mage et al. [31], firstly clusters Differential Scanning Calorimetry (DSC) and clusters were evaluated for identifying the most effective functional groups on behavior of thermal decomposition. Afterwards, methodical classification is developed for assigning unknown thermal behaviors molecule to specific cluster. At last, estimated models of thermal characteristics are prepared with the scope of different classes letting estimating the whole differential scanning calorimetry. Corani and Scanagatta [32], design a model that estimates various air pollution factors by using multi label classifier which is based on Bayesian networks. Furthermore, it learns its own structure with structural learning. Saez et al. [33] develop classification and prediction framework which considers usability of energy and water, monthly flow data and remote sensing measures of plant cover. Gradient boosted trees is

performed in order to classify networks of entire river. Mesquitaet al. [34] develop a software defect prediction (SDP) for improving the trustability of decision making process and also making it probable delay the last decision in regard to refused modules. Software defect prediction is also known as rejoELM and its another version is IrejoELM. Both procedures are based on the weighted extreme learning machine with rejection alternative. So this option makes it probable delay the rejected ones to another moment. Banihashemi et al. [35], build a model which combines Artificial Neural Network and Decision Tree algorithm in order to use energy optimization problems which includes qualitative and quantitative datasets. While Artificial Neural Network is used for prediction process, Decision Tree uses as classification process.

Table 1.2 Classification and Prediction

| **Authors** | **Year** | **Method** | **Data Characteristic** | **Application Area** |
|---|---|---|---|---|
| Chen et al. | 2016 | Logistic Model Tree (LMT), Random Forest (RF), Classification and Regression Tree (CART) | Numeric and Categorical | Geography |
| Shao et al. | 2016 | Recursive feature elimination, time series segmentation, and minimum redundancy maximum relevance based SVM | Numeric | Energy Industry (electric) |
| Mage et al. | 2017 | Decision Tree and Lineer Regression | Numeric | Chemistry |
| Corani and Scanagatta | 2016 | Multi label classifier based on Bayesian networks | Numeric | Environmental Science |

Table 1.2 Classification and Prediction (Cont'd)

| **Authors** | **Year** | **Method** | **Data Characteristic** | **Application Area** |
|---|---|---|---|---|
| Ahmet and Maheswari | 2017 | Hybridized form of Firefly algorithm | Numeric and Categorical | Telecom |
| Saez et al. | 2017 | Gadient-Boosted Trees | Numeric and Categorical | Environmental Science |
| Mesquitaet al. | 2016 | Rejo Extreme Learning Machine | Numeric and Categorical | Software Defect Prediction |
| Mausa and Grbac | 2014 | Genetic Algorithm | Numeric and Categorical | Software Defect Prediction |
| Banihashemi et al. | 2016 | Artificial Neural Network and Decision Tree | Numeric and Categorical | Environmental Science |
| Cheng and Yu | 2015 | SVM with RBF, Multi Lineer Regression and SVM | Categorical | Physic Science |
| Mujtaba et al. | 2017 | Ensemble-Voted Classifier, K-Nearest Neighbor, RF, Naive Bayes, Decision Tree, and SVM | Categorical | Investigation and Law |
| Lynch et al. | 2017 | SVM, Linear Regression, Custom Ensemble , Gradient Boosting Machines, and Decision Trees | Numeric and Categorical | Healthcare |
| Gharagheizi et al. | 2016 | Least Square SVM | Numeric and Categorical | Energy |
| Uthayakumar et al. | 2017 | Ant Colony Optimization, Logistic Regression, RBF, RF and Multilayer Perceptron | Categorical | Finance |

There are various studies in various areas for customer segmentation in literature such as automotive, telecommunication, logistic, hospitality, energy, agriculture, fast moving

consumer goods, finance, e-commerce, textile. However, there is lack of study for IT (information technologies) market. Besides, there are lots of studies in different area in literature which were presented for prediction and classification such as geography, energy, chemistry, telecom, environmental science, telecom, software defect prediction, physic science, investigation and law, healthcare and finance. Hence, a decision support tool is required to assist IT distributors in order to manage customer relations. Therefore, we propose a new decision support framework which consists of clustering model and prediction model with classification. By processing the sales data of an IT distributer in Turkey, we investigated the revenues of resellers regarding to product groups. Therefore, this framework can be seen as a recommendation system that proposes the right product group to the right IT reseller which may cause an increase in IT distributors sales. As result of this study, we expect that the distributor has the chance of increasing its sales of each product groups and also customer satisfaction.

### 1.2 Objective of the Thesis

Proposed model of the thesis can be used in order to develop sales of companies. Especially, the model of the thesis fits on IT distributor companies, which have B2B sales channel. Since, those companies have many product groups and also they have so many dealers in their sales channel, IT distributors sell most of the time only some specific product groups to only some specific dealers.

The aim of this thesis is to assist distributor's plans on sales development based on product group by utilizing data mining methods instead of using traditional methods which rely on managerial experience. With this purpose, a decision support tool is developed to suggest which of its existing dealers may purchase a specified product group that is wanted to increase sales by IT distributor.

### 1.3 Hypothesis

Customers decision of buying a product or not, that can be predicted by using classification algorithms on dataset, which consists of customers purchasing data of other products.

# CHAPTER 2

# METHODS

This section provides the explanation of clustering and classification methods that is utilized in the thesis. Classification and clustering are two major machine learning tools and two of the most important areas of data mining, because they have many practical applications such as medical diagnosis, fraud detection and predicting financial trends.

Classification is the task of assigning a new sample to set of previously known classes, while clustering is the task of grouping samples into clusters of similar samples. This is the reason that classification in known as supervised learning and clustering is known as unsupervised learning [36].

## 2.1 Clustering Methods

Clustering methods become beneficial when a large amount of information is required for being classified or grouped into sets of meaningful categories. Clustering data into meaningful groups is a significant duty of both artificial intelligence and data mining. Clustering methods provide a base to classify data in unsupervised structure. The performance of results generally depends on the algorithm that is used [37].

Without considering clustering type, the data mining duty of clustering endeavours for getting groupings in data is a method which data points in cluster has more similarity than data points in different clusters [38].

In order to classify groups with basing on membership of data points, clusters can be:

- Strict partitioning or exclusive clusters: All data objects pertain to only one specific cluster, that’s why it is the most usual cluster type [38].

- Overlapping clusters: In this type of cluster, cluster groups are not exclusive also all data could pertain to not only one cluster it may be more than one. For instance, company's customers can wants to be grouped as high benefit and high volume customer in the mean time [38].
- Hierarchical clusters: All sub clusters can be united in one parent cluster. For instance, the most beneficial customers cluster could be one more time split as a long time customers cluster and also a cluster that includes customers which are new with high value purchases [38].
- Stochastic or fuzzy clusters: Every data pertain to all cluster groups which have variation rating of membership between zero and one. For instance, supposing that datasets all clusters are A, B, C and D, a data point may be related with all clusters with possibilities as A= 0,5 , B=0,1 , C=0,4 and D=0. Fuzzy clustering provides that each data points have association probabilities of membership to each clusters instead of a definite relation [38].

In addition, clustering methods could be grouped according to algorithm, which is used to find clusters in whole data set. All groups have differentiation on algorithms according to what relationship they compute and evaluate in data set [38].

- Prototype based clustering: A centralized data object, which is also known as a prototype, represents each cluster in this clustering type. Generally, each clusters prototype is the centre of the cluster, therefore it is also known as centre based or centroid clustering. For instance, clustering for customer segmentation [38].
- Density clustering: Cluster can be determined according to density of region where data objects are concentrated and surrounded by low density region where data objects are rare. Every dense region can be related with a cluster and rare region can be scraped as noise. Not all data objects are clustered in this type of clustering, as noise objects is designated to none of the clusters [38].
- Hierchical clustering: This clustering creates cluster hierarchy based upon distance of each data points. Hierarchical clustering's outcome is a dendrogram which is tree schema and it illustrates different clusters at any point that is determined by user. In order to create cluster hierarchy, there are two different approaches. Former is a bottom up approximation which considers every data and clusters are united to a massive final cluster. Latter is top down approximation, which considers one cluster and iteratively split into child clusters until singular

data objects are determined as different clusters. This clustering type is helpful when size of data set is limited. Dendrogram should be cut by decision maker [38].

- Model based clustering: This type of clustering bases on stochastic distribution and statistics models. This procedure is also known as distribution based clustering. It can be considered as a grouping process, which collects the data points that pertains to same probability distribution. Therefore, every cluster could be represented by distribution model such as Poisson and Gaussian. The distribution parameter can be recursively optimized. The whole data set can be presented by a mix of distribution models with this approximation [38].

### 2.1.1 Distance Measures

Most of clustering methods are based on measuring distance between 2 observations. So that there is need for measuring distance formulas. Observation values, which has various variables, can be shown as X matrix. For instance, let a matrix $\boldsymbol{X}$ that has 3 variables and 5 observations is given as in Eq. 2.1.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_{11}} & \boldsymbol{x_{12}} & \boldsymbol{x_{13}} \\ \boldsymbol{x_{21}} & \boldsymbol{x_{22}} & \boldsymbol{x_{23}} \\ \boldsymbol{x_{31}} & \boldsymbol{x_{32}} & \boldsymbol{x_{33}} \end{bmatrix} \tag{2.1}$$

In this instance, first observation is $(\boldsymbol{x_{11}, x_{12}, x_{13}})$ and the second observation is $(\boldsymbol{x_{21}, x_{22}, x_{23}})$. The distance between these two points can be written as $\boldsymbol{d(1,2)}$. So all points distance for each is determined as $\boldsymbol{d(i,j)}$ and symmetrical D distance matrix is which is defined in [39] is as in Eq. 2.2.

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{d(1,2)} & \boldsymbol{d(1,3)} \\ \boldsymbol{d(2,1)} & \boldsymbol{0} & \boldsymbol{d(2,3)} \\ \boldsymbol{d(3,1)} & \boldsymbol{d(3,2)} & \boldsymbol{0} \end{bmatrix} \tag{2.2}$$

where $\boldsymbol{d(i,j) = d(j,i)}$.

Euclidean distance is the most used distance type and it is measured with Pythagoras theorem in 2-dimensional space [39].The Euclidean distance between a and b points is given in Eq. 2.3

$$\boldsymbol{d(i,j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}} \tag{2.3}$$

Manhattan Distance, is calculated as summarizing absolute distance between two observations is calculated as in Eq. 2.4.

$$\boldsymbol{d(i,j) = \sum_{k=1}^{p}\left(\left|x_{ik} - x_{jk}\right|\right)\ i,j = 1,2,\dots,\mathbf{n};\ \mathbf{k} = 1,2,\dots,\mathbf{p}} \quad (2.4)$$

The last commonly used distance measure is Minkowski Distance which is used when measuring distance by p variables are considered between observations values given in Eq 2.5 [39].

$$\boldsymbol{d(i,j) = \left[\sum_{k=1}^{p}\left(\left|x_{ik} - x_{jk}\right|^{m}\right)\right]^{\frac{1}{m}}\ i,j = 1,2,\dots,\mathbf{n};\ \mathbf{k} = 1,2,\dots,\mathbf{p}} \quad (2.5)$$

### 2.1.2 Expectation Maximization Algorithm

Arthur Dempster, Nan Laird, and Donald Rubin proposed Expectation Maximization algorithm in 1977. The EM algorithm is a distance-based algorithm for density prediction. EM algorithm is model based iterative algorithm in order to solve the clustering problem where the data is incomplete or considered incomplete. EM algorithm is an optimization algorithm in order to construct statistical models of the data [40].

In EM algorithm, each data instance belongs to each cluster with a certain probability. EM algorithm starts with initial predictions and iterates to find the maximum likelihood predictions for the parameters. The quality of EM algorithm becomes great when using extremely big dataset. It has been also demonstrated that EM is a good clustering method in terms of computation time and accuracy [40].

The advantages of EM is summarized by Ordonez and Omiecinski [40]:

1- It has a strong statistical basis
2- It is linear in database size
3- It is robust to noisy data
4- It can accept the desired number of clusters as input
5- It can handle high dimensionality
6- It converges fast given a good initialization

Given a set of observations, the EM algorithm targets to compute the maximum likelihood estimation of the unknown parameters of a statistical model. Therefore, the available data is supposed to be a set of observations drawn from a mixture of different probability distributions [41].

$$\boldsymbol{m(y|\Theta) = \sum_{k=1}^{K} \pi_k \phi_k(y|\theta_k)} \quad (2.6)$$

where $\boldsymbol{\Theta = \{\theta_1; \theta_2 \dots, \theta_K, \Pi_1, \dots, \Pi_K\}}$ denotes the set of parameters of the K normal distributions, with normal density given by $\boldsymbol{\phi_k}$ that is characterized by the parameters $\boldsymbol{\theta k = \{\mu k, \sigma_k^2\}}$ and $\boldsymbol{\Pi_k}$ is the weight of each distribution for the data y where $\boldsymbol{\sum_k \pi_k = 1}$ [41].

The algorithm begins with initial guesses for each unknown parameter of the Θ set. A common way to set up the initial values is to choose K random samples from the data points for the means $\boldsymbol{\widehat{\mu_k}}$ and the overall data variance as the initial value for $\boldsymbol{\widehat{\sigma}_k^2}$. For the parameters $\boldsymbol{\widehat{\mu}_k}$ an equal proportion, $\boldsymbol{1/K}$, can be initially attributed to each distribution.

After the initial guesses, the algorithm proceeds by recursively applying the following two steps [41]:

- Expectation step: For each data value, given the parameters of the mixture model, compute the posteriori probabilities of belonging to each distribution. Using the Bayes Rule, the probability that each data value, $\boldsymbol{y_i}$, $\boldsymbol{i = 1, 2, \dots, N}$, belongs to $\boldsymbol{a\Phi_k(.)}$, also called responsibilities, is given by [41]

  $$\boldsymbol{\widehat{\gamma}_{k,i}(\Phi_k|y_i) = \frac{\hat{\pi}_k \Phi_k(y_i|\theta_k)}{\sum_{i=1}^{K} \hat{\pi}_1 \Phi_k(y_i|\theta_1)}} \quad (2.7)$$

- Maximization step: The estimated probabilities on the expectation step are integrated in the maximum likelihood estimation of parameters of the normal distributions. Therefore, the updates depend on the weighted values. The aim of the maximum likelihood estimation is to compute the parameters Θ that give the best fit to the given set of independent data observations [41].

  $$\boldsymbol{\widehat{\mu}_k = \frac{\sum_{i=1}^{N} \widehat{\gamma}_{k,i}, y_i}{\sum_{i=1}^{N} \widehat{\gamma}_{k,i}}} \quad (2.8)$$

  $$\boldsymbol{\widehat{\sigma}_k^2 = \frac{\sum_{i=1}^{N} \widehat{\gamma}_{k,i}, \gamma_{k,i}^{(y_i - \mu_k)^2}}{\sum_{i=1}^{N} \widehat{\gamma}_{k,i}}} \quad (2.9)$$

  $$\boldsymbol{\widehat{\pi}_k = \sum_{i=1}^{N} \frac{\widehat{\gamma}_{(k,i)}}{N}} \quad (2.10)$$

The Expectation step and Maximization step are iterated until the convergence criterion is met (e.g., a maximum number of iterations). The algorithm's output will then give us

the estimation of the mean value, variance and mixture proportion of each normal distribution on the analyzed model [41].

There are some studies in literature that use Expectation Maximization clustering algorithm, Nilashi et al. [40] have developed the multi-criteria collaborative filtering recommender systems for hotel recommendation in online to enhance the predictive accuracy by using Gaussian mixture model with EM algorithm and Adaptive Neuro-Fuzzy Inference System. Expectation maximization is used on gene discretization for complexity simplification and better discrimination capability is employed by [42]. Ortiz-Rosario et al. [43] presented a new methodology which uses an expectation-maximization Gaussian mixture algorithm for automatic objective classification of neurons recorded in association with behavioral tasks into groups. The methodology is capable of identifying and clustering similar firing rate profiles automatically based on specific signal features.

### 2.1.3 K-Means Algorithm

Cluster analysis emerged as a major topic in the 1960's and 1970's when the monograph 'Principles of numerical taxonomy' by Sokal and Sneath (1963) motivated world-wide research on clustering methods and thereby initiated the publication of a broad range of books such as 'Les bases de la classification automatique' (Lerman 1970), 'Mathematical taxonomy' (Jardine and Sibson 1971), 'Cluster analysis for applications' (Anderberg 1973), 'Cluster analysis' (Bijnen 1973), 'Automatische Klassifikation' (Bock 1974), 'Empirische Verfahren zur Klassifikation' (Sodeur 1974), 'Probleme und Verfahren der numerischen Klassifikation (Vogel 1975), 'Cluster-Analyse-Algorithmen (Späth 1975, 1985), and 'Clustering algorithms' (Hartigan 1975). With the consequence that the basic problems and methods of clustering became well-known in a broad scientific community, in statistics, data analysis, and - in particular - in applications. [44]

The K-means clustering algorithm is described in detail by Hartigan (1975). An efficient version of the algorithm is presented here. The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that the within-cluster sum of squares is minimized. It is not practical to require that the solution has minimal sum of squares against all partitions, except when M, N are small and K = 2. We seek instead "local" optima, solutions such that no movement of a point from one cluster to another will reduce the within-cluster sum of squares [45].

The algorithm requires as input a matrix of M points in N dimensions and a matrix of K initial cluster centres in N dimensions. The number of points in cluster L is denoted by NC(L). D(I, L) is the Euclidean distance between point I and cluster L. The general procedure is to search for a K-partition with locally optimal within-cluster sum of squares by moving points from one cluster to another [45].

Step 1. For each point I ($\boldsymbol{I = 1, 2, \ldots, M}$), find its closest and second closest cluster centres, IC1($\boldsymbol{I}$) and IC2($\boldsymbol{I}$) respectively. Assign point I to cluster IC1($\boldsymbol{I}$).

Step 2. Update the cluster centres to be the averages of points contained within them.

Step 3. Initially, all clusters belong to the live set.

Step 4. This is the optimal-transfer stage:

Consider each point $\boldsymbol{I}$ ($\boldsymbol{I = 1, 2, \ldots, M}$) in turn. If cluster $\boldsymbol{L}$ ($\boldsymbol{L = 1, 2, \ldots, K}$) is updated in the last quick-transfer stage, then it belongs to the live set throughout this stage. Otherwise, at each step, it is not in the live set if it has not been updated in the last $\boldsymbol{M}$ optimal transfer steps. Let point $\boldsymbol{I}$ be in cluster $\boldsymbol{L}1$. If $\boldsymbol{L}1$ is in the live set, do Step 4a; otherwise, do Step 4b.

Step 4a. Compute the minimum of the quantity, $\boldsymbol{R2 = [NC(L) * D(I, L)^2]/[NC(L) + 1]}$, over all clusters $\boldsymbol{L}$ ($\boldsymbol{L \neq L1, L = 1, 2, \ldots, K}$). Let $\boldsymbol{L}2$ be the cluster with the smallest $\boldsymbol{R2}$. If this value is greater than or equal to $\boldsymbol{[NC(L1) * D(I, L1)^2]/[NC(L1) - 1]}$, no reallocation is necessary and $\boldsymbol{L2}$ is the new $\boldsymbol{IC2(I)}$. (Note that the value $\boldsymbol{[NC(L1) * D(I, L1)2]/[NC(L1) - 1]}$ is remembered and will remain the same for point $\boldsymbol{I}$ until cluster $\boldsymbol{LI}$ is updated.) Otherwise, point I is allocated to cluster $\boldsymbol{L2}$ and $\boldsymbol{L1}$ is the new $\boldsymbol{IC2(I)}$. Cluster centres are updated to be the means of points assigned to them if reallocation has taken place. The two clusters that are involved in the transfer of point I at this particular step are now in the live set.

Step 4b. This step is the same as Step 4a, except that the minimum $\boldsymbol{R2}$ is computed only over clusters in the live set.

Step 5. Stop if the live set is empty. Otherwise, go to Step 6 after one pass through the data set.

Step 6. This is the quick-transfer stage:

Consider each point $\boldsymbol{I}$ ($\boldsymbol{I} = \boldsymbol{1, 2, \ldots, M}$) in turn. Let $\boldsymbol{L1} = \boldsymbol{IC1}(\boldsymbol{I})$ and $\boldsymbol{L2} = \boldsymbol{IC2}(\boldsymbol{I})$. It is not necessary to check the point I if both the clusters $\boldsymbol{L1}$ and $\boldsymbol{L2}$ have not changed in the last $\boldsymbol{M}$ steps. Compute the values

$\boldsymbol{R1} = [\boldsymbol{NC}(\boldsymbol{L1}) * \boldsymbol{D}(\boldsymbol{I}, \boldsymbol{L1})\boldsymbol{2}]/[\boldsymbol{NC}(\boldsymbol{L1}) - \boldsymbol{1}]$ and R2 = $[\boldsymbol{NC}(\boldsymbol{L2}) * \boldsymbol{D}(\boldsymbol{I}, \boldsymbol{L2})\boldsymbol{2}]/[\boldsymbol{NC}(\boldsymbol{L2}) + \boldsymbol{1}]$

(As noted earlier, $\boldsymbol{R1}$ is remembered and will remain the same until cluster $\boldsymbol{L1}$ is updated.)

If $\boldsymbol{R1}$ is less than $\boldsymbol{R2}$, point I remains in cluster $\boldsymbol{L1}$. Otherwise, switch $\boldsymbol{IC1}$ ($\boldsymbol{I}$) and $\boldsymbol{IC2}$($\boldsymbol{I}$) and update the centres of clusters $\boldsymbol{L1}$ and $\boldsymbol{L2}$. Thve two clusters are also noted for their involvement in a transfer at this step.

Step 7. If no transfer took place in the last $\boldsymbol{M}$ steps, go to Step 4. Otherwise, go to Step 6. [45]

There are some studies in literature that use k means clustering algorithm, Chen et al [5] combined factor analysis and k means methods in order to segment hot spring hotel industry. Huang et al. [15] compared support vector clustering performance with k-means and self-organizing map methods by segmenting a drink company's customers. They found out it outperforms the others in marketing segmentation. A model was proposed by Murray et al [17] that uses k-means clustering in order to identify segments of customers with similar demand behaviors in supply chain.

The advantages and disadvantages of k means are:

Advantages

1. Relatively scalable and simple.
2. Suitable for datasets with compact spherical clusters that are well-separated

Disadvantages

1. Severe effectiveness degradation in high dimensional spaces as almost all pairs of points are about as far away as average; the concept of distance between points in high dimensional spaces is ill-defined
2. Poor cluster description
3. Reliance on the user to specify the number of clusters in advance
4. High sensitivity to initialization phase, noise and outliers
5. Frequent entrapments into local optima
6. Inability to deal with non-convex clusters of varying size and density [46].

### 2.1.4 Hierarchical Clustering Algorithm

Hierarchical clustering (HC) algorithms organize data into a hierarchical structure according to the proximity matrix. The results of HC usually are depicted by a binary tree or dendrogram, as depicted in Figure 2.1. The dotted line in Figure 2.1 corresponds to choosing a number of clusters based on where the tree is cut. The root node of the dendrogram represents the whole data set, and each leaf node represents a data object. The intermediate nodes thus describe the extent to which the objects are proximal to each other, and the height of the dendrogram usually expresses the distance between each pair of objects or clusters, or one object and one cluster. The best clustering results can be obtained by cutting the dendrogram at different levels. This representation provides very informative descriptions and a visualization of the potential data clustering structures, especially when real hierarchical relationships exist in the data, such as in the data from evolutionary research on different species of organisms. HC algorithms are classified primarily as agglomerative methods or divisive methods. Agglomerative clustering starts with N clusters, each of which includes exactly one object. A series of merge operations then combines all objects into the same group. Divisive clustering proceeds in an opposite way. The entire data set initially belongs to a single cluster, and then a procedure successively divides it until all clusters are singletons. For a cluster with N objects, there are $2N-1 - 1$ possible two-subset divisions, which is very computationally expensive. Therefore, divisive clustering is not commonly used in practice. We focus on agglomerative clustering in the following discussion [47].

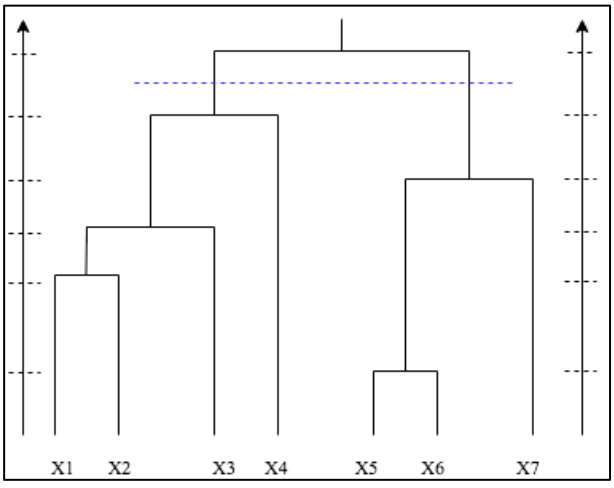


Figure 2.1 An example of a dendrogram in hierarchical clustering.

The variety of agglomerative clustering algorithms originally came about because of the different definitions for the distance between two clusters. The general formula for distance was proposed by Lance and Williams as [47]:

$$\boldsymbol{D(C_i, (C_i, C_j)) = \alpha_i D(C_l, C_i) + \alpha_j D(C_l, C_i) + \beta D(C_i, C_j) + \gamma |D(C_l, C_i) - D(C_l, C_j)|} \tag{2.11}$$

where $\boldsymbol{D}()$ is the distance function, and $\boldsymbol{\alpha_i, \alpha_j, \beta}$ and $\boldsymbol{\gamma}$ are coefficients whose values depend on the scheme used.

The formula describes the distance between a cluster $\boldsymbol{l}$ and a new cluster formed by the merging of the two cluster $\boldsymbol{i}$ and $\boldsymbol{j}$. Note that when $\boldsymbol{\alpha_i = \alpha_j = 1/2}$, $\boldsymbol{\beta = 0}$, and $\boldsymbol{\gamma = -1/2}$, the formula becomes

$$\boldsymbol{D(C_l, (C_i, C_j)) = \min(D(C_l, C_i), D(C_l, C_j))} \tag{2.12}$$

which corresponds to the single linkage method. When $\boldsymbol{\alpha_i = \alpha_j = \gamma = 1/2}$ and $\boldsymbol{\beta = 0}$, formula is

$$\boldsymbol{D(C_l, (C_i, C_j)) = \max(D(C_l, C_i), D(C_l, C_j))} \tag{2.13}$$

which corresponds to the complete linkage method [47].

Several agglomerative clustering algorithms, including group average linkage, median linkage, centroid linkage, and Ward's method, can also be constructed by selecting appropriate coefficients in the formula. Single-linkage, complete-linkage and average-linkage methods, all of which also are called graph methods, consider all points of a pair of clusters when calculating their inter-cluster distance. The other methods are called geometric methods because they use geometric centers to represent clusters and they determine the distances between clusters based on these centers [47].

Ward's linkage method's objective of minimizing the variance / dispersion within customer segments formed and its ability to consider each customer as independent i.e. each customer segment begins as its own cluster [13].

The aim in Ward's linkage method is to join customers into customer segments such that the dispersion of the final CIC estimate within the final customer segment is minimized. To do this, each customer segment begins as its own cluster. Clusters are merged if the merger results in the reduction of the dispersion of final CIC estimate within the merger formed. The difference between each customer segment within a cluster and that average

similarity is calculated and squared. The sum of the squared deviations is used as a measure of error (-L2squared dissimilarity measure) within a cluster. This means that at each merging stage of clusters the average dissimilarity of the cluster is measured. A customer segment is selected to enter the cluster if it is the customer segment whose inclusion in the cluster produces the least increase in the error.

L2 squared dissimilarity measure (L2SDM) which is described in [13]

$$\boldsymbol{L2SDM} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \sum_{\forall \boldsymbol{x_i} \in \boldsymbol{C_k}} \|\boldsymbol{X_i} - \boldsymbol{\mu_k}\|^{\boldsymbol{2}} \tag{2.14}$$

where $\boldsymbol{C_k}$ is the set of instances in cluster $\boldsymbol{k}, \boldsymbol{\mu_k}$ the vector mean of cluster $\boldsymbol{k}$. The components of $\boldsymbol{\mu_k}$ are calculated as in Eq. 2.15 [13].

$$\boldsymbol{\mu_{k,j}} = \frac{\boldsymbol{1}}{\boldsymbol{N_k}} \sum_{\forall \boldsymbol{x_i} \in \boldsymbol{C_k}} \boldsymbol{x_{i,j}} \tag{2.15}$$

Hierarchical clustering is useful when the data size is limited. A level of interactive feedback is required to cut the dendrogram tree at a given level of precision [38].

There some studies in literature which use hierarchical clustering such as Dzobo et al. (2014) offered a model that uses hierarchical clustering method to segment electricity customers. Zhang et al. (2018) proposed an unsupervised method to detect shilling attacks based on hidden Markov model and hierarchical clustering. Ma et al. (2018) used hierarchical clustering for building energy usage behaviors [46].

The advantages and disadvantages of hierarchical clustering are [46]:

Advantages

- Embedded flexibility regarding the level of granularity.
- Well suited for problems involving point linkages, e.g. taxonomy trees [46].

Disadvantages

- Inability to make corrections once the splitting/merging decision is made.
- Lack of interpretability regarding the cluster descriptors.
- Vagueness of termination criterion.
- Prohibitively expensive for high dimensional and massive datasets.
- Severe effectiveness degradation in high dimensional spaces due to the curse of dimensionality phenomenon [46].

## 2.2 Classification Methods

Classification algorithms are used in data mining to classify data based on class labels. It involves building a model using training data set, and then using the built model to assign given items to specific classes/categories. In the model building process, also called training process, a classification algorithm finds relationships between the attributes of the data and the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a new data set in which the class assignments are unknown [48].

### 2.2.1 Decision Tree

Decision tree is a classification technique, which purposes to divide data in to groups by using flowchart that looks like a tree. In this context, this model includes a root node and branch nodes. A decision tree begins from the input data is divided into different groups based upon some estimator variables pre-identified as dividing criteria, which is called the root node. After that, those divided data are spread to sub nodes as branches occurring from the root node. The data on sub nodes will be exposed to either further or no splits. The first are the interior nodes that the next data dividing is handled to shape new subgroups as child branches occurred graphically on next step. However, the second are leaf nodes which behaves corresponding data group at the present level as their latest outputs. Figure 2.2 presents a decision tree rendition used for medium yearly source energy consumption. It represents it as per unit floor of a commercial structure. The gross floor are and building usage ratio are selected as estimator factor in the root node and the interior node, in order. Energy consumption mixed data has been simplified to a hierarchy of those groups. [49]

Considerably in decision tree analysis, information entropy is an significant notion used to measure homogeneity of data group. So it is determined by [49]

$$\boldsymbol{E} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} -\boldsymbol{P_i log_2 P_i} \tag{2.16}$$

where $\boldsymbol{E}$ is the information entropy. $\boldsymbol{n}$ and $\boldsymbol{P_i}$ a are the number of diversified aim values and the possibility of dataset getting the $\boldsymbol{i^{th}}$ aim value, in order. Based on which decision tree building linking the peak of root node to every branch node is determined, the entropy is used to compute the information particle or achievement ratio [49].

As compared with different data driven approximation, decision tree's building is simple to understand and also its application doesn't contain complex calculation fund of knowledge. Notwithstanding, decision tree's inability is also obvious that the aims used in a decision tree are firstly based on anticipation. That usually causes to important variation of its estimation from the actual outcome. Furthermore, decision tree architecture is restriction on the purpose of the method is not able to handle very well with nonlinear data and time series [49].

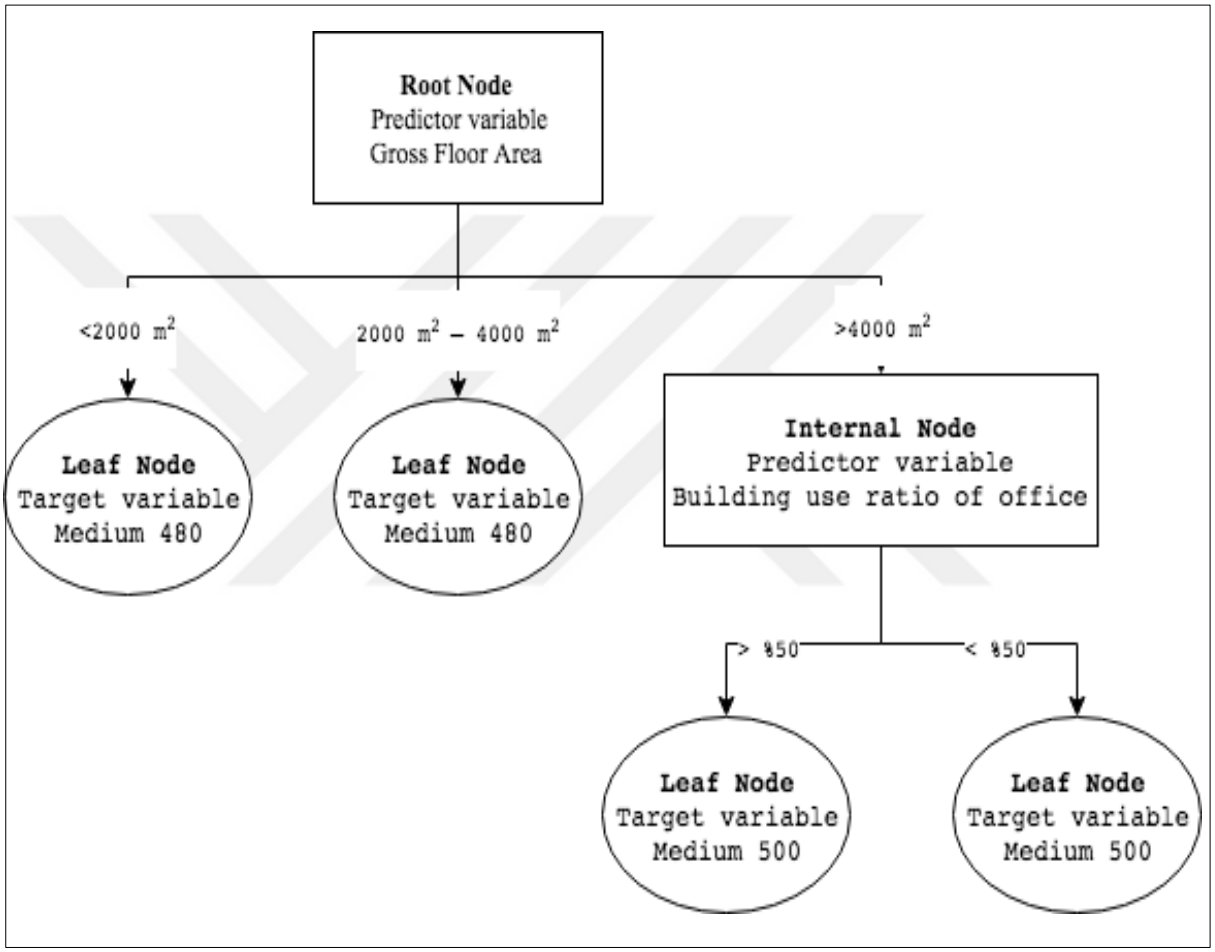



Figure 2.2 Decision Tree

**ID3 Algorithm**

ID3 computing begins with original seedling as root hub. It emphasizes via all unused attribute of the seedling determines the entropy of related attribute on each tour of the algorithm. During that time, algorithm selects the attribute that has minimum entropy value. The seedling is S, so it is divided by the chosen attribute for generating subsets of the information. This algorithm progresses to recur on each and whole items in subset and

also it considers just items which are not ever chosen before. Recursion processes could bring to stop in one those cases [20]:

- Each member in the subset belongs to same class, so that the node is transform into a leaf. Also, it is labeled with class.
- The node is transformed into a leaf and also labelled via the most prevalent class of the instances in the current subset, if there are not anymore attributes for being selected however the instances still don't belong to the same class (some of them could be positive and some of them could be negative).
- Unless there are instances in current subset, that occurs when parent set did find in order to be matching a particular chosen attributes' value. For instance; one leaf is generated and it is labelled via the most prevalent class of the instances in parent set, unless there were instance match with marks is equal or higher than 100 [20].

**C4.5 Algorithm**

In order to generate a decision tree that is extended version of ID3 measurement, C4.5 algorithm is used. C4.5 extends and improves ID3 algorithm with handling pruning trees, continues and discrete properties and missing values after building. Decision trees which are generated by C4.5 algorithm can be used in order to group and also it is often referred as statistical classifier. This algorithm generates decision trees from training dataset as like ID3 algorithm. Since C4.5 is a supervised learning algorithm, it needs training instances set that can be seen as couple, which has input object requested value or class. C4.5 analyze training test and constructs a classifier, which needs to have capacity for arranging certainly training and test cases. Test instance should have input object and shouldn't have output, so that it can estimate output value. It should be considered example training dataset as $\boldsymbol{S = S_1, S_2, \ldots S_n}$ that is already classified. Every sample of Si compose of feature vector $(\boldsymbol{x_{1,i}, x_{2,i}, \ldots, x_{n,i}})$ where xj symbolize attribute or feature of the sample. In addition, it represents the class that Si falls. This algorithm chooses an attribute that most effectively divides its samples' set into subsets in such a way that it ends up with one class or the other, at each node. Dividing requirement is information which is normalized achievement that is not symmetric measurement of P and Q probability's differentiation. Attribute, which has the maximum information achievement, is selected for making decision [20].

General working step/stages of C4.5 algorithm is as below, [20]

- Postulate every sample in the list belong to same class. C4.5 algorithm generates a leaf node so this specific class will be selected, if it is right.
- Assuming that any of the feature does not provide any information achievement. It generates a decision node high-level tree by using the prospective value if it is right.
- C4.5 algorithm generates a decision node high-level tree by using the prospective value [20].

### 2.2.2 Naïve Bayes

The Naive Bayes Classifier bases on Bayesian theorem and it is specifically used when inputs dimensionality is high. This technique is able to calculate the most probable output that based on input. Moreover, Naive Bayes Classifier could add new raw data at runtime so that it has classifier, which is a better probabilistic. It takes in consideration that absence or presence of specific attribute of a class, which is not related to the absence, or presence of any other attributes when the class variable is provided. For instance, a fruit could be taken in consideration for being a grape if it is green. This classifier takes in consideration all of those features for contributing independently to probability, which this grape is green, even if those attributes can change as regards each other or presence of other attributes of a class [20].

Naive Bayes Classifier works as below; [20]

Bayes theorem provides a way of calculating the posterior probability, $\boldsymbol{P(c|x)}$, from $\boldsymbol{P(c)}$, $\boldsymbol{P(x)}$, and $\boldsymbol{P(x|c)}$. Naive Bayes classifier considers that the effect of the value of a predictor ($\boldsymbol{x}$) on a given class ($\boldsymbol{c}$) is independent of the values of other predictors [20].

$$\boldsymbol{P(c|x)} = \frac{\boldsymbol{P(x|C)P(c)}}{\boldsymbol{P(x)}} \tag{2.17}$$

$$\boldsymbol{P(c|x)} = \boldsymbol{P(x_1|c)} \times \boldsymbol{P(x_2|c)} \times ... \times \boldsymbol{P(x_n|c)} \times \boldsymbol{P(c)} \tag{2.18}$$

- P(c|x) is following possibility of target class.
- P(c) is called former possibility of class .
- P(x|c) is known as likelihood that is the possibility of estimator of given class .
- P(x) is former possibility of estimator of class [20].

### 2.2.3 Support Vector Machine

Different favorite artificial intelligent procedure is supported vector machine, that handles with n data loggings, i.e.,$\{\boldsymbol{x_i}, \boldsymbol{Y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$, with the input $\boldsymbol{X_i} \in \boldsymbol{R^N}$ and the target $\boldsymbol{Y_i} \in \boldsymbol{R}$. (consider that $\boldsymbol{Y_i}$ also may be binary). Currently, this principle has been largely implemented for solving regression problems in order to predict a relationship between nonlinear inputs and continuous aims. As this method is called as support vector machine, it used for regression. It has become especially significant data driven approximation for estimation process [49].

The primary goal in support vector machine is building a decision function, $\boldsymbol{F(x_i)}$, by usage of a training process that is based on retrospective data. It is needed that for a given input $\boldsymbol{x_i}$, the outcome predicted by that function shouldn't deviate from the real aim $\boldsymbol{Y_i}$ larger than default threshold $\boldsymbol{\varepsilon}$. In SVR, function is mostly presents in the form of [49]

$$\boldsymbol{F(X_i)} = \langle \boldsymbol{w}, \boldsymbol{\varphi(x_i)} \rangle + \boldsymbol{b} \tag{2.19}$$

where the bias $\boldsymbol{b} \in \boldsymbol{R}$. $\langle \cdot, \cdot \rangle$ and *w* symbolizes the dot product and weight determined in $\boldsymbol{R^N}$. $\boldsymbol{\varphi(x_i)}$ is a non-linear mapping of the input space to a high dimensional feature space. In above equation $\boldsymbol{w}$ and $\boldsymbol{b}$ are two not known (unknown) and they should be predicted via optimizing the regularized risk function as minimization [49].

$$\boldsymbol{L} := \frac{\boldsymbol{1}}{\boldsymbol{2}}\|\boldsymbol{w}\|^{\boldsymbol{2}} + \boldsymbol{c}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{\xi_i} + \boldsymbol{\xi_i^*}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{n_i\xi_i} + \boldsymbol{n_i^*\xi_i^*}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{a_i}(\boldsymbol{\varepsilon} + \boldsymbol{\xi_i} - \boldsymbol{y_i} - \langle \boldsymbol{w}, \boldsymbol{\varphi(x_i)} \rangle - \boldsymbol{b}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{a_i^*}(\boldsymbol{\varepsilon} + \boldsymbol{\xi_i^*} - \boldsymbol{y_i} - \langle \boldsymbol{w}, \boldsymbol{\varphi(x_i)} \rangle - \boldsymbol{b}) \tag{2.20}$$

where $\{\boldsymbol{a_i}, \boldsymbol{a_i^*}, \boldsymbol{\eta_i}, \boldsymbol{\eta_i^*} \geq \boldsymbol{0}\}$ are the Lagrange multiplier. $\|\boldsymbol{w}\|$is the Euclidean norm. $\boldsymbol{\xi_i}, \boldsymbol{\xi_i^*}$ are two slack variables for copying with some impossible optimization restrictions. The constant $\boldsymbol{c} > \boldsymbol{0}$ identified for specifying the barter between the training error and model flatness, former is also known as over fitting and latter is also known as under fitting. It should be considered that all Lagrange multipliers are free. They are $\boldsymbol{\eta_i} = \boldsymbol{c} - \boldsymbol{a_i}$ and $\boldsymbol{\eta_i^*} = \boldsymbol{c} - \boldsymbol{a_i^*}$ and $\{\boldsymbol{a_i}, \boldsymbol{a_i^*}\}$ can be specified by corresponding double optimization [49].

$$\boldsymbol{Maximize}\ \boldsymbol{W(a_i, a_i^*)} = -\frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}}(\boldsymbol{a_i} - \boldsymbol{a_i^*})\big(\boldsymbol{a_j} - \boldsymbol{a_j^*}\big)\Big(\boldsymbol{\varphi(x_i)} \cdot \boldsymbol{\varphi}\big(\boldsymbol{x_j}\big)\Big) + \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}}(\boldsymbol{a_i} - \boldsymbol{a_i^*})\boldsymbol{y_i} - \boldsymbol{\varepsilon}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}}(\boldsymbol{a_i} + \boldsymbol{a_i^*})\ \boldsymbol{subject\ to}\ \begin{cases}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}}(\boldsymbol{a_i} - \boldsymbol{a_i^*}) = \boldsymbol{0} \\ \boldsymbol{a_i}, \boldsymbol{a_i^*} \in [\boldsymbol{0}, \boldsymbol{c}]\end{cases} \tag{2.21}$$

With the computed $\boldsymbol{a_i}, \boldsymbol{a_i^*}$ the weight $\boldsymbol{w}$ can be written a function of $\{\boldsymbol{a_i}, \boldsymbol{a_i^*}, \boldsymbol{x_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}}$. This causes to decision function in support vector regression [49].

$$\boldsymbol{F(x) = \sum_{x_i \in SV}(a_i - a_i^*)K(x, x_i) + b} \quad (2.22)$$

where $\boldsymbol{K(x, x_i) = \varphi(x), \varphi(x_i)}$. It is called as kernel function in support vector regression, having another formulas for several applications in literature e.g. $\boldsymbol{K(x, x_i) = \exp(-\gamma \|x - x_i\|^2)}$. should be indicated them in above equation doesn't suffice each inputs. Alternatively, just those with regard to $\boldsymbol{(a_i - a_i^*) \neq 0}$ are contained. Furthermore, the bias b, which is in above equation, is also calculated by those support vectors [49].

$$\boldsymbol{b = \frac{1}{N_1}\left\{\sum_{a_i \in (0,c)}\left[Y_i - \sum_{x_j \in SV}(a_J - a_j^*)K(x_i, x_j) - \varepsilon\right] + \sum_{a_j^* \in (0,c)}\left[Y_i - \sum_{x_j \in SV}(a_J - a_j^*)K(x_i, x_j) - \varepsilon\right]\right\}} \quad (2.23)$$

Herein, $\boldsymbol{N_1}$ presents number of support vectors with either $\{\boldsymbol{a_i \in (0, c), a_i^* = 0}\}$ or $\{\boldsymbol{a_i = 0,\ a_i^* \in (0, c)}\}$. Support vector regression model can be used as a estimating tool for new input $\boldsymbol{x}$, when the decision function is completely determined by the training dataset [49].

It is important to underlying the supremacy of support vector regression or more usually support vector machine. Its concept is effortlessly made a sweeping statement for other problems and it can get optimal solution globally. Its ability of handling with nonlinear connection by transferring them into high dimensional linear problem is influential. Nevertheless, the procedure is pretty running out of time for grand scale problems. Currently, so much efforts have been spending in order to develop probable direction for optimizing its calculation fast and productively [49].

### 2.2.4 Evaluation of Classification Results

**Threshold Types of Discriminator Metrics**

For binary classification problems, the discrimination evaluation of the optimal solution during the classification training can be defined based on confusion matrix as shown below [50].

Table 2.1 Confusion Matrix for Binary Classification

| | **Actual Positive Class** | **Actual Negative Class** |
|---|---|---|
| **Predicted Positive Class** | True positive (tp) | False negative (fn) |
| **Predicted Negative Class** | False positive (fp) | True negative (tn) |

Through accuracy the quality of produced solution is evaluated based on percentage of correct predictions over total instances. The complement metric of accuracy is error rate, which evaluates the produced solution by its percentage of incorrect predictions. Both of these metrics were used commonly by researchers in practice to discriminate and select the optimal solution [50].

The advantages of accuracy or error rate are, this metric is easy to compute with less complexity; applicable for multi-class and multi-label problems; easy-to-use scoring; and easy to understand by human [50]. In Table 2.2, threshold metrics for classification evaluations are summarized where each class of data; $\boldsymbol{tp_i}$ – true positive for $\boldsymbol{c_i}$; $\boldsymbol{fn_i}$ – false negative for $\boldsymbol{c_i}$; $\boldsymbol{tn_i}$ – true negative for $\boldsymbol{c_i}$; and $\boldsymbol{M}$ macro - averaging.

Table 2.2 Threshold Metrics for Classification Evaluations

| **Metrics** | **Formula** | **Evaluation Focus** |
|---|---|---|
| Accuracy (acc) | $\frac{\boldsymbol{tp+tn}}{\boldsymbol{tp+fp+tn+fn}}$ | In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. |
| Error Rate (err) | $\frac{\boldsymbol{fp+fn}}{\boldsymbol{tp+fp+tn+fn}}$ | Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated. |
| Sensitivity (sn) | $\frac{\boldsymbol{tp}}{\boldsymbol{tp+fn}}$ | This metric is used to measure the fraction of positive patterns that are correctly classified |
| Specificity (sp) | $\frac{\boldsymbol{tn}}{\boldsymbol{tn+fp}}$ | This metric is used to measure the fraction of negative patterns that are correctly classified. |
| Precision (p) | $\frac{\boldsymbol{tp}}{\boldsymbol{tp+fp}}$ | Precision is used to measure the positive patterns that are correctly predicted in a positive class. |
| Recall (r) | $\frac{\boldsymbol{tp}}{\boldsymbol{tp+tn}}$ | Recall is used to measure the fraction of positive patterns correctly classified |
| F-Measure | $\frac{\boldsymbol{2*p*r}}{\boldsymbol{p+r}}$ | This metric represents the harmonic mean between recall and precision values |
| Geometric-mean | $\sqrt{\boldsymbol{tp*tn}}$ | This metric is used to maximize the tp rate and tn rate. |

Table 2.2 Threshold Metrics for Classification Evaluations (Cont'd)

| | | |
|---|---|---|
| Averaged Accuracy | $\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i}}{l}$ | The average effectiveness of all classes |
| Averaged Error Rate | $\frac{\sum_{i=1}^{l} \frac{fp_i + fn_i}{tp_i + fn_i + fp_i}}{l}$ | The average error rate of all classes |
| Averaged Precision | $\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$ | The average of per-class precision |
| Averaged Recall | $\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$ | The average of per-class recall |
| Averaged F-Measure | $\frac{2 * p_m * r_m}{p_m + r_m}$ | The average of per-class F-measure |

**Mean Square Error (MSE)**

In general, the MSE measures the difference between the predicted solutions and desired solutions. The solution that has minimum MSE score is be preferred. The MSE is defined as below [50]:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(P_j - A_j)^2 \quad (2.24)$$

**Area Under the ROC Curve**

AUC is one of the popular ranking type metrics. The AUC was used to construct an optimized learning model and also for comparing learning algorithms. Unlike the threshold and probability metrics, the AUC value reflects the overall ranking performance of a classifier. The AUC value can be calculated as below [50].

$$AUC = \frac{S_p - n_p(n_n+1)/2}{n_p n_n} \quad (2.25)$$

where, $S_p$ is the sum of the all positive examples ranked, while $n_p$ and $n_n$ denote the number of positive and negative examples respectively. The AUC was proven theoretically and empirically better than the accuracy metric for evaluating the classifier performance and discriminating an optimal solution during the classification training [50].

Although the performance of AUC was excellent for evaluation and discrimination processes, the computational cost of AUC is high especially for discriminating a volume of generated solutions of multiclass problems [50].

# CHAPTER 3

## PROPOSED METHODOLOGY

Methodology should be started with data exploration in order to determine the need in the market. Subsequently, it should be begun data preparation phase for getting meaningful results from data mining. Because the dataset is too big and it has outliers, the results could be misleading without preprocessing.

After that, all customers are supposed to be clustered in order to segment them. So that clustering results can be used for classifying customers and applying prediction on it in order to prepare decision support tool. Afterwards, report findings are prepared in order to evaluate methods and algorithms. So that decision support tool is prepared regarding to report findings. So process flow chart is prepared as below Figure 3.1.

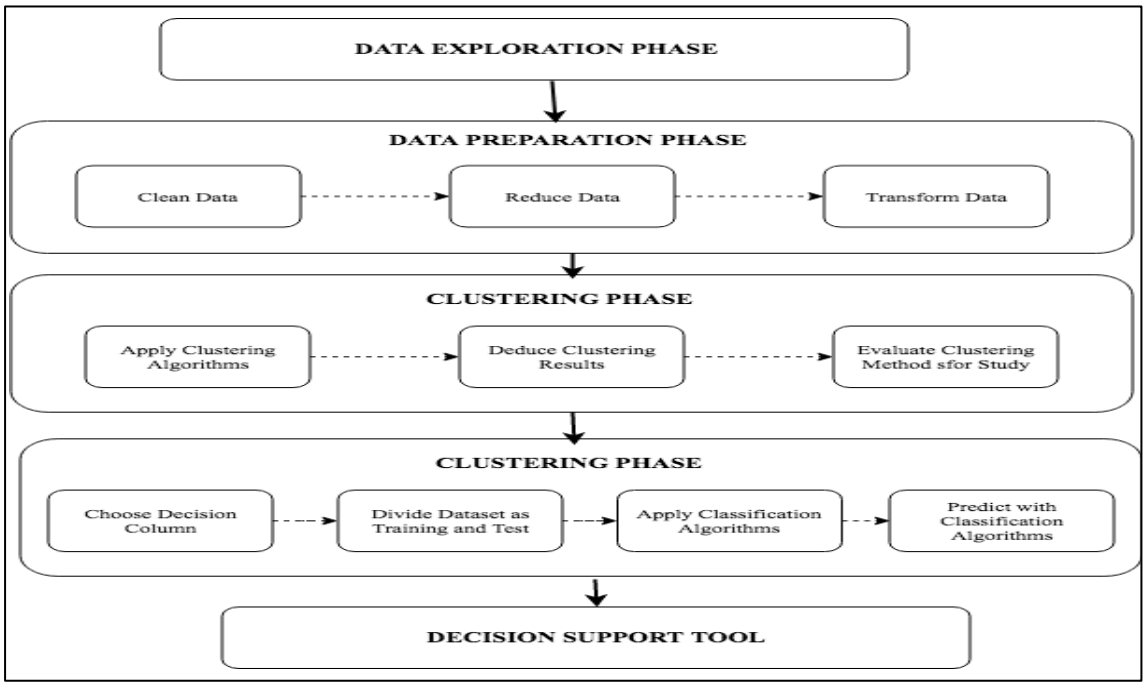



Figure 3.1 Process Flow Chart

### 3.1 Data Exploration Phase

Data exploration step begins with determining a need in market, which can be matched with more benefit. If a firm is not able to make right and proper decisions about strategies of marketing, deployment of resources, distribution of products and different areas of enterprise behavior and it is doomed in the end. In order to make correct and proper business decisions is the one of the biggest business need. Data exploration could present some simple source materials for decision making information [51].

### 3.2 Data Preparation Phase

Basically, working with data is data mining so it is significant to explore more completely what data represents also why and how it is turned into and what can be made an inference. It is difficult to determine how to get value from data mining without determining the problem. It is also important to identify the form of solution. Once to have determined how fitting solution looks like and find out or get proper data, data mining process is ready to start [51].

#### Data Cleaning

It should be started with data cleaning process if there is data in data set, which does not have any feature on demand. For instance, there could be missing, unsuitable or outlier data. Those missing and unsuitable data is called as noisy data. So the data should be cleaned first in such circumstances and it should be replaced with new which is decided by analyzer. One of the methods, which are below, can be used [39].

- The data has missing, unsuitable or outlier value that can be removed from dataset [39].
- General constant value can be used for all missing, unsuitable or outlier value. For example, “unknown” variable can be used instead. But it could be a problem to use same constant value for all variable instead of missing, unsuitable or outlier value [39].
- With calculating mean of variable, the mean can be used instead of missing, unsuitable or outlier value in this variable [39].
- It can be calculated mean of variable that belongs only one class and then it can be used instead of missing, unsuitable or outlier value [39].

- Missing, unsuitable or outlier value can be predicted with a method such as regression or decision trees and this prediction can be used instead of missing value [39].

Dataset can include some outliers, which can be positive or negative. Those outliers also should be cleaned with above methods. However, positive outliers sometimes can be demand and only negative outliers are not. In such circumstances, only negative outliers should be cleaned as changing the data with decision maker's choice.

**Data Reduction**

After that, it should be continued with data reduction, since the analyses process could be longer than we want. If it is believed that analysis results do not get changed or be improved, number of data or variables could be reduced. So data reduction can be applied as below [39].

- Data merge
- Dimension reduction
- Data compression
- Sampling
- Generalization

Unnecessary data should be removed from dataset with choosing process for data reduction. After that, it is purposed to prepare sample cluster data instead of bigger dataset. That's why this process should be done with clustering phase, when one of the clusters will be used as training dataset in order to get better result [39].

**Data Transformation**

It is not right to use the data directly in some cases. If variables' means and variations have big differences, data, which has bigger differences, has more pressure on others and seriously reduces others roles [39]. That's why dataset should be discretized for data transformation.

So that, data preparation phase is applied before clustering and classification phases.

### 3.3 Clustering Phase

The most common application of clustering is to explore the data and find all possible meaningful groups in the data. Clustering a company's customer records and attributes

can yield a few groups in such a way that customers within a group are more like each other than customers belonging to a different group [38].

In proposed methodology, dataset should be clustered and then one of the clusters should be chosen in order to be used in classification.

**Applying Clustering Algorithms**

It should be applied more than one clustering method in order to evaluate them in each other and choose the best of them for the study. So that it should be defined first which algorithms are going to be used and it can be decided as looking which of the algorithms can be cluster better the dataset.

**Deducing All Clustering Results**

After applying clustering algorithms, all clusters characteristic should be understood. Each cluster should be defined with deducing from clustering results.

**Evaluating and Choosing the Best Clustering Method for the Study**

After understanding each cluster for each clustering method, it should be chosen the best of them with looking which clustering method's result is more beneficial for the goal of the study.

### 3.4 Classification Phase

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any context in which some decision or forecast is made on the basis of presently available information [20].

Proposed methodology uses above characteristic of classification as a benefit. But classification process should be applied correctly, if not it could provide wrong and misleading prediction

**Choosing Decision Column**

After choosing clustering method, the chosen clustering method's result could be used as preprocess phase of classification. That brings benefit to prepare more effective and trustable decision support tool. Since; one of the clusters' dataset can represent training dataset better for classification process, when application focuses a specific column to

use it as decision column in classification process. Because initial dataset doesn't include decision column, it can be generated from clusters characteristic.

First, one of the columns should be decided as decision column in order to prepare training dataset for classification.

**Dividing Dataset as Training and Test Dataset**

The cluster should be divided as training and test dataset. So that, training dataset can be trained by classification algorithm and then it can be evaluate with test data.

**Applying Classification Algorithms**

Training dataset is trained by classification algorithm. After that, test dataset is predicted by classification algorithm with using training result.

It should be applied more than one classification algorithm in order to evaluate them in each other and choose the best of them.

### 3.5 Preparing Decision Support Tool

According to the results of clustering phase and prediction phase, it should be evaluated which algorithms are fit with this model. It is important because the best algorithms for this model will give more realistic results and decision suggestions in the future.

It is not considered only training classifications evaluations results to decide classification algorithm that is going to be used. Since, it is also applied one more classification algorithm for prediction. That's why, the success of prediction should be considered to get more realistic results and decision suggestions in the future.

After this evaluation, the most successful classification algorithm should be chosen. After that, decision support tool is prepared as result.

# CHAPTER 4

# APPLICATION

This study segments IT reseller companies, which are referred as customers of IT distributors. Data set includes revenue of each resellers and their product groups information. The model is created in order to cluster IT resellers into different groups regarding their differences and similarities in purchasing behavior. As a result, each cluster have their characteristic related with their purchasing behavior. One of the clusters is selected as initial dataset for prediction process after that the most sold product group is determined in the cluster. So that, this cluster is used in order to train the dataset and then prediction is applied on all customers of IT Distributor Company with using that training process.

It is beneficial to use the cluster for training process since it is easy to train initial dataset.

## 4.1 Data Exploration Phase

IT distributor company has annual sales data of whole resellers regarding to product groups. Although they have different customer segmentation, they have never clustered their customers according to purchasing behavior. So that it is needed.

Also IT distributor company needs to use the result of clustering in order to apply some prediction such as if the customer sells this product group or not. So that preparing decision support tool for this company is decided.

## 4.2 Data Preparation Phase

This study uses a dataset, which presents revenue of each resellers regarding each product groups for initial. So that variables are resellers and product groups such as server,

storage, software, keyboard etc. The initial dataset, which is numeric, includes purchasing data of 4345 resellers regarding 238 product groups. The part of initial dataset is below.

Table 4.1 Example of initial dataset

| **Customer No** | **Y Product Group** | **X Product Group** | **Z Product Group** | **ACND Product Group** | **BNAD Product Group** | **ACNB Product Group** | **…** |
|---|---|---|---|---|---|---|---|
| 82 | $ 18.7 | | | | | | ……. |
| 84 | | | | -$ 1920.0 | | $ 3.174.7 | ……. |
| 821 | | | | -$ 441.0 | | $ 4.043.6 | ……. |
| 78 | | | | | | | ……. |
| 68 | | | | | | | ……. |
| 70 | | $ 152.0 | | -$ 132.0 | | $ 2.581.1 | ……. |
| 657 | | | | | | -$ 41.0 | ……. |
| 526 | | | | -$ 583.0 | | $ 6.727.9 | ……. |
| 2595 | | | | $ 25.0 | | $ 100.0 | ……. |
| 2341 | | | | -$ 278.0 | | $ 7.006.4 | ……. |
| 617 | | | | | | -$ 166.3 | ……. |
| 588 | | | | | | $ 4.072.4 | ……. |
| 3512 | | | | | | | ……. |
| 2712 | | | | -$ 64.0 | | -$ 1.3 | ……. |
| 1778 | | | | | | | ……. |

After some pre-processing such as data cleaning and data reduction, data set is generated as 320 resellers purchasing value regarding 98 product groups. Some product groups such as motherboard types are combined each other. In addition, some product groups, which has too few revenue, are removed from the dataset. Also resellers are reduced to 320 by selecting the most revenue earned reseller. Lastly, negative outliers are removed, since it is not possible to have negative revenue. The part of pre-processed dataset is below.

Table 4.2 Example of pre-processed dataset

| Reseller no | Product Group1 | Product Group2 | Product Group3 | Product Group4 | Product Group5 | Product Group6 | Product Group7 | … |
|---|---|---|---|---|---|---|---|---|
| 82 | 0 | 0 | 0 | 11301059 | 15286.7 | 0 | 16379.31 | ……. |
| 84 | 0 | 3541413 | 0 | 0 | 0 | 1050212.,72 | 131209.33 | ……. |

Table 4.2 Example of pre-processed dataset (Cont'd)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 821 | 0 | 568967 | 0 | 0 | 0 | 973687.91 | 0 | ……. |
| 78 | 0 | 241749 | 0 | 0 | 139086 | 357815.62 | 30895.59 | ……. |
| 68 | 0 | 0 | 0 | 142034.11 | 52996,3 | 33879.21 | 6893.71 | ……. |
| 70 | 0 | 387254 | 0 | 1040.64 | 1421,09 | 798741.86 | 77294.47 | ……. |
| 657 | 0 | 34920.6 | 0 | 1067.55 | 0 | 27253.55 | 0 | ……. |
| 526 | 0 | 230678 | 0 | 0 | 0 | 252368.3 | 0 | ……. |
| 2595 | 0 | 998,01 | 0 | 88084.16 | 0 | 1074.5 | 0 | ……. |
| 2341 | 0 | 458946 | 0 | 27522.77 | 0 | 762071.91 | 0 | ……. |
| 617 | 0 | 100763 | 0 | 4154.94 | 0 | 191831.58 | 0 | ……. |
| 588 | 0 | 248746 | 0 | 9908.18 | 0 | 405646.91 | 0 | ……. |
| 3512 | 0 | 0 | 0 | 285833.7 | 0 | 25057.66 | 0 | ……. |
| 2712 | 0 | 15694.6 | 0 | 56322.79 | 0 | 17914.41 | 841025.85 | ……. |
| 1778 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ……. |

After that, pre-processed dataset is discretized in order to get better results from clustering. Bins number is defined as 10 and used equal frequency. The part of pre-processed dataset that was prepared for clustering is below. Also discretized data set instance is presented as below Figure 4.1.

Table 4.3 Discretized Pre-processed dataset which includes 320 resellers and 98 product groups.

| Customer No | Product Group 17 | Product Group 19 | Product Group 21 | Product Group 29 |
|---|---|---|---|---|
| 20 | '\'(77976.99-inf)\" | '\'(-inf-100.96]\" | '\'(-inf-102.74]\" | '\'(88545.23-inf)\" |
| 605 | '\'(865.5-1734.54]\" | '\'(-inf-100.96]\" | '\'(-inf-102.74]\" | '\'(88545.23-inf)\" |
| 1462 | '\'(865.5-1734.54]\" | '\'(-inf-100.96]\" | '\'(-inf-102.74]\" | '\'(833.02-1316]\" |
| 3695 | '\'(-inf-106.45]\" | '\'(-inf-100.96]\" | '\'(-inf-102.74]\" | '\'(-inf-105.47]\" |

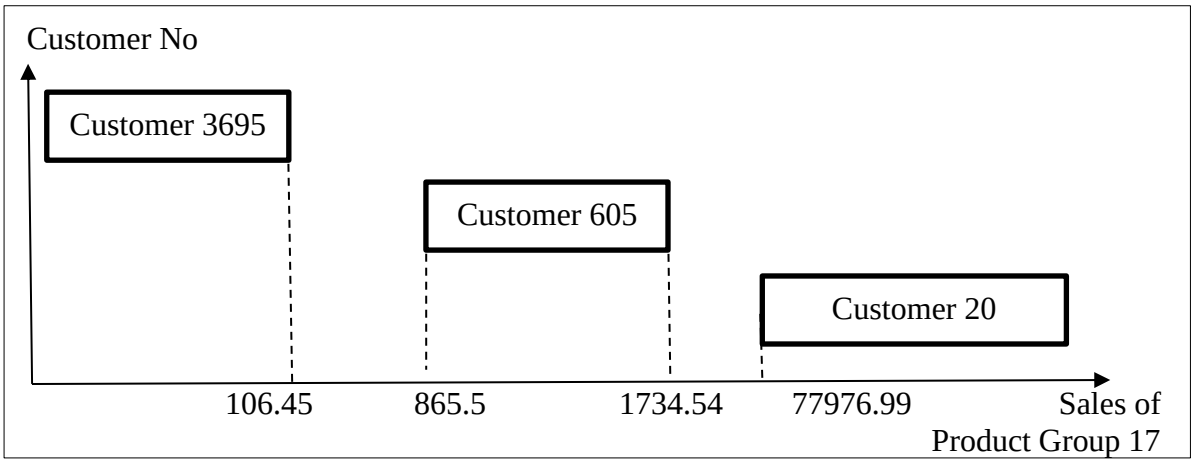



Figure 4.1 Discretized Data Instance

### 4.3 Clustering Phase

3 clustering algorithms, which are Expectation Maximization algorithm, K Means algorithm and Hierarchical Clusterer are applied in this study for segmenting the customers. Those 3 methods are decided; since Expected Maximization is robust to noisy data, K means algorithm is relatively scalable and simple, hierarchical clustering is embedded flexibility regarding the level of granularity and also well suited for problems involving point linkages. So that it is decided to choose the best of them for the study.

The number of clusters are defined as 5 in those algorithms and their results are evaluated for defining the best fitting algorithm for this study. When cluster number was defined as 3 and 4 for fitting algorithm, the result was not satisfied. Since, the most sold product groups gathered in 1 cluster and also gathered product groups were not related in each clusters, even though customers were scattered homogenously. Cluster number is not defined more than 5, as it decreases members of each clusters and it is not beneficial for classification of training dataset.

All algorithms are run in WEKA, their results are below.

#### 4.3.1 Expectation Maximization

Each clusters member numbers and percentages are similar except cluster1 so that expectation maximization algorithms results can be concluded as all customers are scattered homogeneously except Cluster 1 as shared below Table 4.4.

Table 4.4 Cluster member numbers and percentage of Expectation Maximization Algorithm

| Cluster Name | Amount | Percentage |
|---|---|---|
| Cluster 0 | 51 | 15.9% |
| Cluster 1 | 2 | 0.6% |
| Cluster 2 | 88 | 27.5% |
| Cluster 3 | 86 | 26.9% |
| Cluster 4 | 93 | 29.1% |

Figure 4.2 and Figure 4.3 illustrates distribution of members in clusters for Expected Maximization.

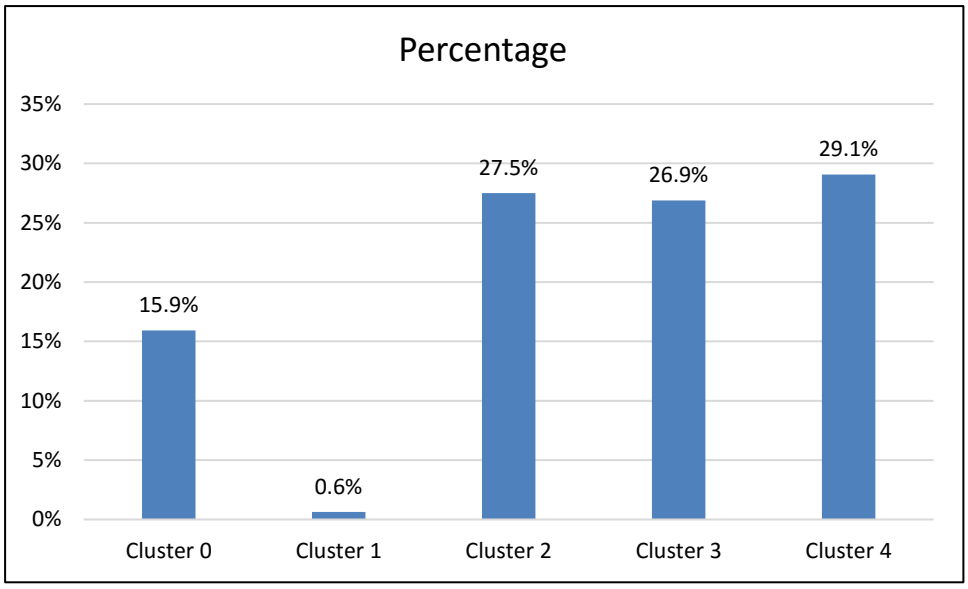



Figure 4.2 Cluster members distribution of Expectation Maximization algorithm

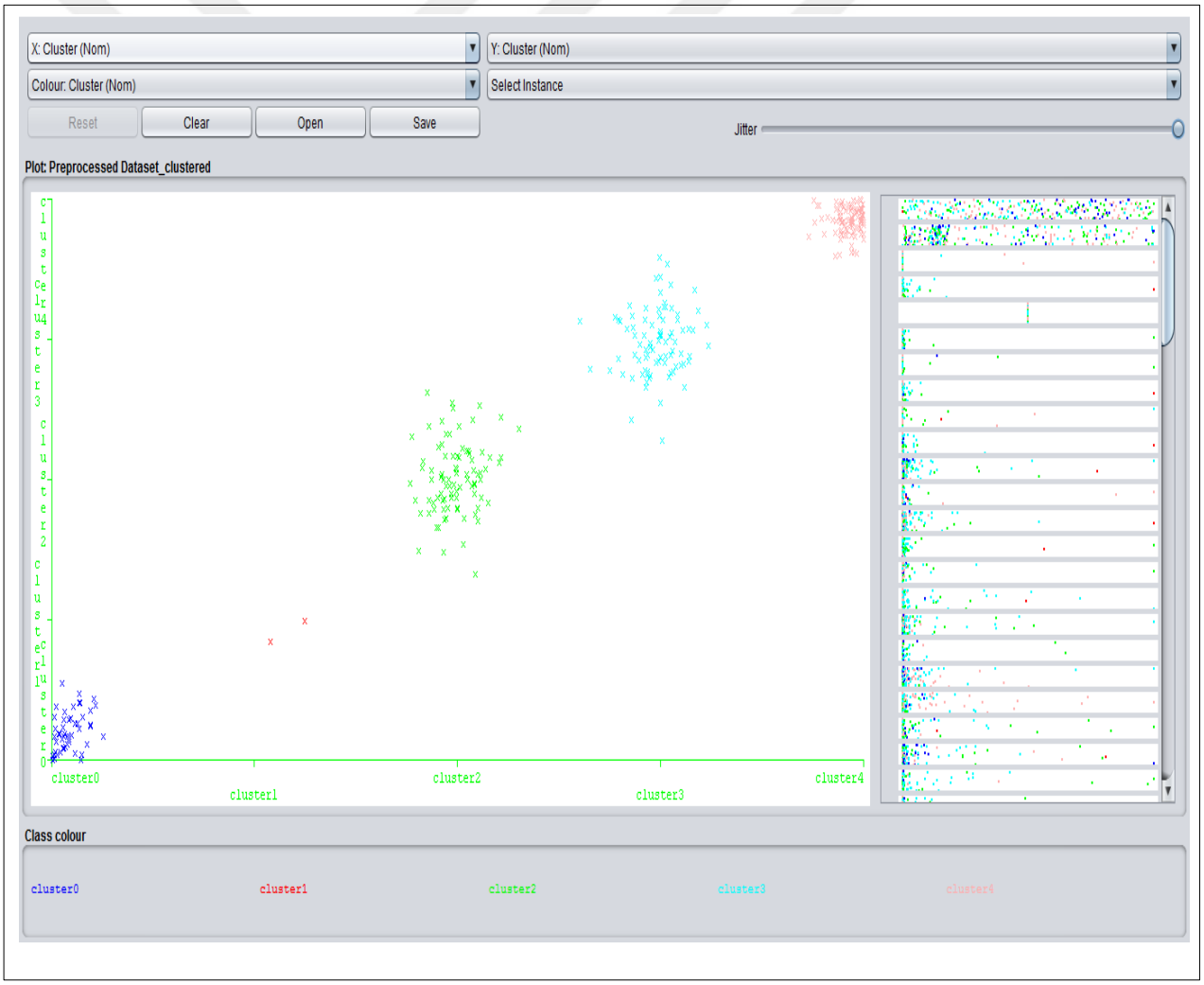



Figure 4.3 Expectation Maximization Visualization of Scatter

### 4.3.2 K Means Algorithm

As result of k means algorithm, most of all cluster members are gathered in cluster 0 and others are mostly shared by cluster 1 and cluster 4. This result doesn't have homogenous distribution of customers as seen below Table 4.5.

Figure 4.4 and Figure 4.5 illustrates distribution of members in clusters for K Means.

Table 4.5 Cluster member numbers and percentage of K Means Algorithm

| Cluster Name | Count | Percentage |
|---|---|---|
| Cluster 0 | 190 | 59.4% |
| Cluster 1 | 53 | 16.6% |
| Cluster 2 | 19 | 5.9% |
| Cluster 3 | 5 | 1.6% |
| Cluster 4 | 53 | 16.6% |

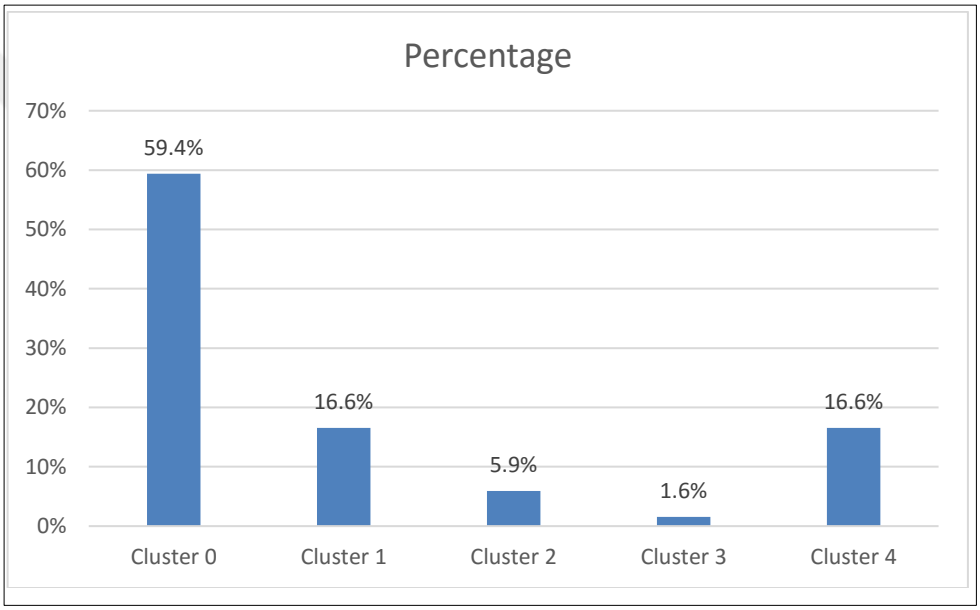



Figure 4.4 Cluster members distribution of k means algorithm

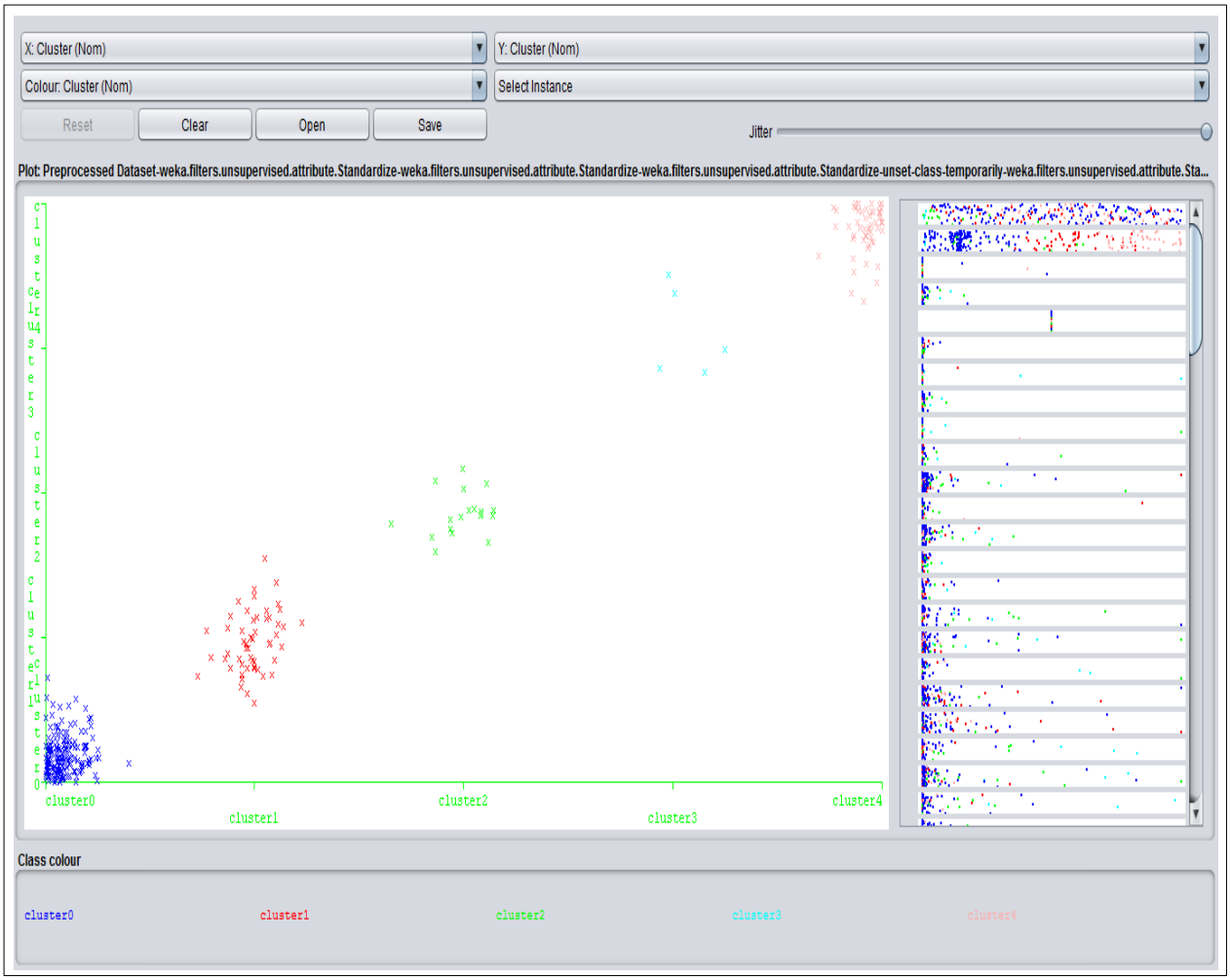



Figure 4.5 K Means Visualization of Scatter

### 4.3.3 Hierarchical Clusterer

As result of hierarchical clusterer, all the clusters members are scattered homogeneously. Member numbers of all clusters are very similar so that all 5 clusters can be used as training dataset for classification process as seen below Table 4.6.

Figure 4.6 and Figure 4.7 illustrates distribution of members in clusters for hierarchical clusterer.

Table 4.6 Cluster member numbers and percentage of hierarchical clusterer

| Cluster Name | Amount | Percentage |
|---|---|---|
| Cluster 0 | 91 | 28.4% |
| Cluster 1 | 85 | 26.6% |
| Cluster 2 | 47 | 14.7% |
| Cluster 3 | 58 | 18.1% |
| Cluster 4 | 39 | 12.2% |

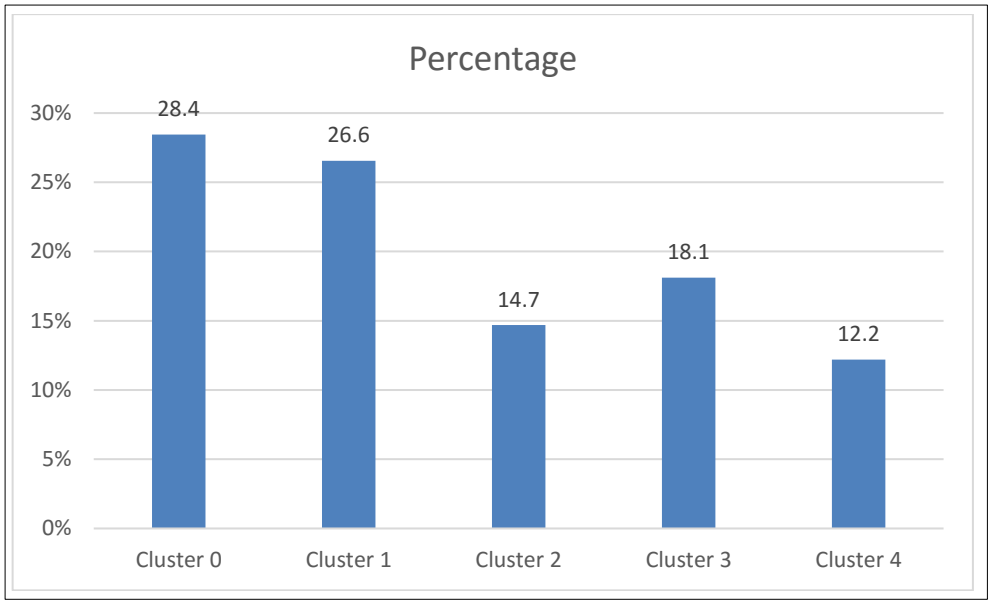



Figure 4.6 Cluster members distribution of hierarchical clusterer

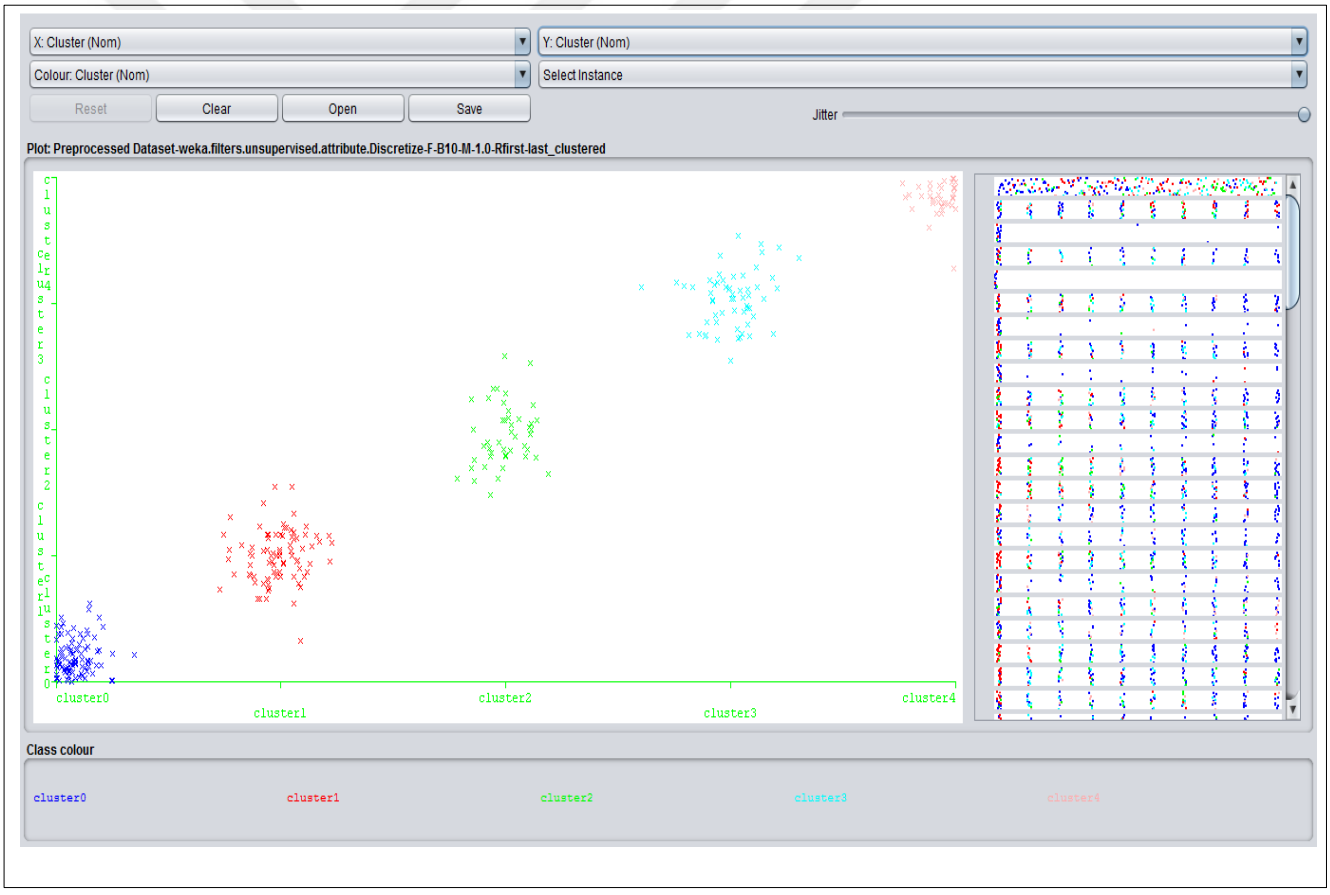



Figure 4.7 Hierarchical Clustering Visualization of Scatter

### 4.3.4 Summary of Clustering Results

Expectation maximizations distribution of all members in clusters is much better than K means algorithm according to equal distribution of members. However, hierarchical clusterers clusters members are very similar.

Results of all clusterers show that hierarchical clusterer should be chosen, if cluster members are wanted to be scattered to all cluster homogeneously. Since, it is usable for segmentation and using this result for training datasets of classification.

Hierarchical clusterers results present clusters characteristics as below:

Cluster 0's members purchase mostly various standard hardware products groups which don't need any pre sales activities such as keyboard, mouse, motherboard etc.

Cluster 1's members behavior is to purchase various software product groups such as security, business analytics etc.

Cluster 2's members buy specific product group such as printer accessorize.

Cluster 3's members purchasing behavior is to buy some design software product groups.

Cluster 4's members mostly purchase project hardware product groups which need pre sales activities and sold project based such as server and storage.

### 4.4 Classification Phase

Product Group 17 in cluster 4 is chosen, because the IT distributor company significantly needs to increase sales of Product Group 17 in their sale channel in that business term. So that, Cluster 4 is initial dataset for decision support system.

Firstly, Cluster 4 is divided as training data and test data. After that Naïve Bayes, Support Vector Classifier and C4.5 algorithms are applied on those for training and then same algorithms are used in order to predict customers decision of selling Product Group 17 which is server. As final step, the best model is chosen for decision support system.

Naïve Bayes, Support Vector Classifier and C4.5 algorithms are decided for this phase; since Naïve Bayes is able to calculate the most probable output that based on input, Support Vector Classifier has ability of handling with nonlinear connection by transferring them into high dimensional linear problem is influential, C 4.5 algorithm is Suitable for handling both categorical and quantitative values and universal for solving both classification and regression problem. So that it is decided to choose the best of them for the study.

### 4.4.1 Support Vector Classifier and Prediction Results

According to classification results, it looks classification and training success is extraordinary. Since it classifies all instances correctly, TP and under ROC area is 1 as below Table 4.7. However, prediction success is just 50% as Table 4.8.

Table 4.7 Evaluation of Training Dataset with Support Vector Classifier

| | | |
|---|---|---|
| Correctly Classified Instances | 29 | 100% |
| Kappa statistic | 1 | |
| Mean absolute error | 0 | |
| Root mean squared error | 0 | |
| Relative absolute error | 0 | |
| Root relative squared error | 0 | |
| Total Number of Instances | 29 | |
| TP Rate | FP Rate | Precision |
| 1 | 0 | 1 |
| F-Measure | MCC | ROC Area |
| 1 | 1 | 1 |

Table 4.8 The Result of Prediction Model with Support Vector Classifier

| Customer No | Product Group 93 | Product Group 94 | Product Group 95 | Product Group 96 | Product Group 97 | Real Decision | Predicted Decision | Result |
|---|---|---|---|---|---|---|---|---|
| 3434 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | yes | Wrong |
| 2447 | '\'(724.995-1402.5]\"' | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | yes | no | Wrong |
| 1474 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | yes | yes | Correct |
| 733 | '\'(-inf-142.5]\" | '\'(33267.435-inf)\" | '\'(754.94-1127.18]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | no | Correct |
| 647 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | yes | Wrong |

Table 4.8 The Result of Prediction Model with Support Vector Classifier (Cont'd)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 228 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(4828-9025.835]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | no | Correct |
| 234 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | no | Correct |
| 640 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | yes | Wrong |
| 1656 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | no | Correct |
| 637 | '\'(-inf-142.5]\" | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\'' | no | yes | Wrong |
| | | | | | | | Succes | 50% |

### 4.4.2 Naïve Bayes Classification and Prediction Results

According to classification results, it looks classification and training success is great. Since, it classifies 96.55% of instances correctly and TP Rate is 96.6% as below Table 4.9. Moreover, under ROC Area is 1. However, prediction success is 60% as Table 4.10.

Table 4.9 Evaluation of Training Dataset with Naïve Bayes Classification

| | | | | |
|---|---|---|---|---|
| Correctly Classified Instances | 28 | 0.965517 | | |
| Kappa statistic | 0.9218 | | | |
| Mean absolute error | 0.0276 | | | |
| Root mean squared error | 0.1449 | | | |
| Relative absolute error | 0.060686 | | | |

Table 4.9 Evaluation of Training Dataset with Naïve Bayes Classification (Cont'd)

| | | | | |
|---|---|---|---|---|
| Root relative squared error | 0.304683 | | | |
| Total Number of Instances | 29 | | | |
| TP Rate | FP Rate | Precision | | Recall |
| 0.966 | 0.066 | 0.967 | | 0.966 |
| F-Measure | MCC | ROC Area | | |
| 0.965 | 0.925 | 1 | | |

Table 4.10 The Result of Prediction Model with Naïve Bayes Classification

| Customer No | Product Group 94 | Product Group 95 | Product Group 96 | Product Group 97 | Real Decision | Predicted Decision | Result |
|---|---|---|---|---|---|---|---|
| 3434 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 2447 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | yes | no | Wrong |
| 1474 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | yes | yes | Correct |
| 733 | '\'(33267.435-inf)\" | '\'(754.94-1127.18]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 647 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 228 | '\'(-inf-101.4]\" | '\'(4828-9025.835]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 234 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 640 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | yes | Wrong |
| 1656 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | yes | Wrong |
| 637 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | yes | Wrong |
| | | | | | | Succes | 60% |

### 4.4.3 C4.5 Classification and Prediction Results

According to classification results, it looks classification and training success is great. Since, it classifies 96.55% of instances correctly and TP Rate is 96.6% as below Table 4.11. Moreover, under ROC Area is 1. Besides, prediction success is 80% as Table 4.12.

Table 4.11 Evaluation of Training Dataset with C4.5 Classification

| | | | |
|---|---|---|---|
| Correctly Classified Instances | 28 | 96.5517% | |
| Kappa statistic | 0.9218 | | |
| Mean absolute error | 0.046 | | |
| Root mean squared error | 0.1516 | | |
| Relative absolute error | 10.1059% | | |
| Root relative squared error | 31.8919% | | |
| Total Number of Instances | 29 | | |
| TP Rate | FP Rate | Precision | Recall |
| 0.966 | 0.066 | 0.967 | 0.966 |
| F-Measure | MCC | ROC Area | |
| 0.965 | 0.925 | 0.995 | |

Table 4.12 The Result of Prediction Model with C4.5 Classification

| Customer No | Product Group 94 | Product Group 95 | Product Group 96 | Product Group 97 | Real Decision | Predicted Decision | Result |
|---|---|---|---|---|---|---|---|
| 3434 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 2447 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | yes | yes | Correct |
| 1474 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | yes | yes | Correct |
| 733 | '\'(33267.435-inf)\" | '\'(754.94-1127.18]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | yes | Wrong |
| 647 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |

Table 4.12 The Result of Prediction Model with C4.5 Classification (Cont'd)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 228 | '\'(-inf-101.4]\" | '\'(4828-9025.835]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 234 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 640 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | yes | Wrong |
| 1656 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| 637 | '\'(-inf-101.4]\" | '\'(-inf-114.025]\" | '\'(-inf-104.275]\" | '\'(-inf-271.07]\" | no | no | Correct |
| | | | | | | Succes | 80% |

### 4.5 Preparing Decision Support Tool

The evaluation of training dataset results shows that support vector classifier is the best algorithm for this model but Support Vector Classifiers result of prediction model is not successful. Naïve Bayes and C4.5 algorithms evaluation of training data set results are really great and Support Vector Classifier is just a little bit better than them. On the other hand, C4.5 algorithms model predicts customers decision of selling Product Group 17 with 80% success and that's prediction success much more than Naïve Bayes and Support Vector Classifier. So that, C4.5 algorithm is chosen as classification and prediction process in order to maximize prediction success for support decision model.

# CHAPTER 5

## CONCLUSION

IT distributor channel consists of businesses that act as an intermediary between vendors and resellers in the distribution of software or hardware. These businesses require analytical approaches to manage their resellers in an effective manner. When the pace of change in IT industry and the short product lifecycles are considered, approaches that can be easily adapted to the firm should be encouraged.

In this thesis, a general decision aid framework is proposed to assist value-added sales processes for B2B distributor companies. The proposed framework especially fits on IT distributor companies, since they are required to manage a large number of product groups and resellers in a dynamic sales channel. IT distributors sell most of the time only some specific product groups to only some specific dealers. That's why some product groups are not sold to dealers, which have potential to purchase those product groups. However, there is an opportunity to sell those product groups to other dealers. Making effort to develop business with all dealers could cause losing focus and using human resources and finance ineffectively. Another contribution of this study, dealers that have potential to sell those product groups can be defined by decision support system so that IT distributor companies can manage their time, human resources and finance effectively instead selling same product groups to same dealers. Also, this approach may enable IT distributor companies to increase the revenue for each product group.

The aim of this thesis is to assist distributor's plans on sales development based on product group by utilizing data mining methods instead of using traditional methods which rely on managerial experience. With this purpose, a decision support tool is developed to suggest which of its existing dealers may purchase a specified product group that is wanted to increase sales by IT distributor.

Firstly, the firm's annual sales dataset is obtained regarding to product groups and dealers. Then data reduction phase is applied, ie. similar repetitive product groups are assembled into a single product group. In addition, to reduce the size of the data meaningfully, the data set is retained by the dealers that provide the majority sales of the year and the rest are removed from the dataset. After that, negative outliers were observed, and these negative outliers are set to zero in order to get a meaningful result. For getting meaningful results in the clustering and classification processes, the data set is categorized using the discretization method. The usefulness of these preprocess operations has been seen by taking meaningful results in data mining algorithms.

After the preprocessing is completed, the clustering analysis phase is performed with three different methods as expectation maximization, k-means clustering and hierarchical clustering. Dealers were segmented according to purchasing behavior by hierarchical clustering analysis, which gives the best results. It is seen that each segment has a different behavior of purchasing product groups. Within these clusters, a product group, which has strategic signification, is chosen for further analysis. The cluster is selected which sells best that product group, for using in classification process.

In the classification process, the selected cluster is divided into training dataset and test dataset. The training dataset has been classified with three different classification algorithms, which are C4.5 algorithm, support vector classifier and Naive Bayes. Then, prediction success of these methods, which are created by making prediction on the same test dataset with each algorithm, are determined. C4.5 algorithm has performed best considering estimation success.

At this point, the IT distributor can visit and discuss with the dealer whether they will buy that product group or not. In addition, promotional activities can be prepared to sell this product group. Prepared decision support tool can also be applied to different product groups and can be used to increase the sales of that product group.

Data preparation with reducing data size, removing outliers and discretizing dataset caused to get meaningful results. One of the reason to get such meaningful and beneficial outcomes is that is segmenting customers with using right clustering method and using those clusters as training and test dataset for preparing support decision system. Especially at the stage of deciding final system, dividing dataset into parts as test and training added positive effect on getting more realistic results instead of examining at

only classification evaluation results. Test dataset, which has decision result as known already, is used in removing decision results and making prediction to control how much this model is realistic.

Finally, using machine learning methods with artificial neural networks can be interesting for future studies.

## REFERENCES


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

[2] Chan C. C. H., (2008). "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer", Expert Systems with Applications, 34 (4): 2754-2762.

[3] Hwang H., Jung T. and Suh E., (2004). "An LTV model and customer segmentation based on customer value: a case study on wireless telecommunication industry", Expert System witj Applications, 26 (2): 181-188.

[4] Paker N. and Vural C. A., (2016). "Customer segmentation for marinas: Evaluating marinas as destinations", Tourism Management, 56: 156-171.

[5] Chen K. H., Liu H. H. and Chang F. H., (2013). "Essential customer service factors and the segmentation of older visitors within wellness tourism based on hot springs hotels", International Journal of Hospitality Management, 35: 122-132.

[6] Gilbao S., (2009). "A segmentation study of Israeli mall customers", Journal of Retailing and Consumer Services, 16 (2): 135-144.

[7] Rypakova M., Stefanikova L. and Moravcikova K., (2015). "Suggestion of green customer segmentation in Slovakia", Procedia Economics and Finance, 26 (sayı): 359-366.

[8] Tabi A., Hille S. L. and Wüstenhagen R., (2014). "What makes people seal the green power deal? - Customer segmentaion based on choice experiment in Germany", Ecological Economics, 107: 206-215.

[9] Li D.-C., Dai W.-L. and Tseng W.-T., (2011). "A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business", Expert Stmstems with Applications, 38 (6): 7186-7191.

[10] Hong T. and Kim E., (2012). "Segmenting customers in online stores based on factors that effect the customer's intention to purchase", Expert Systems with Applications, 39 (2): 2127-2131.

[11] Liu H.-H. and Ong C.-S., (2008). "Variable selection in clustering for marketing segmentation using genetic algorithm", Expert Systems with Aplications, 34 (1): 502-510.

[12] Kuo R., Mei C., Zulvia F. and Tsai C., (2016). "An application of a metaheuristic algorithm-based clustering ensemble method to APP customer segmentation", Neurocomputing, 205: 116-129.

[13] Dzobo O., Alvehag K., Gaunt C. and Herman R., (2014). "Multi-dimensional customer segmentation model for power system reliability-worth analysis", Electrical Power and Energy Systems, 62: 532-539.

[14] Cuadros A. J. and Dominguez V. E., (2014). "Customer segmentation model based on value generetion for marketing strategies formulation", Estudios Gerenciales, 30 (130): 25-30.

[15] Huang J.-J., Tzeng G.-H. and Ong C.-S., (2007). "Marketing segmentation using support vector clustering", Expert Systems with Applications, 32 (2): 313-317.

[16] Coussement K., Bossche F. A. V. and Bock K. W. D., (2014). "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees", Journal of Business Research, 67 (1):, 2751-2758.

[17] Murray P. W., Agar B. and Barajas M. A., (2015). "Forecasting Suply Chain Demand by Clustering Customers", IFAC-PapersOnLine, 48 (3): 1834-1839.

[18] Benítez I., Quijano A., Díez J.-L. and Delgado I., (2014). "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers", Electrical Power and Energy Systems, 55: 437-448.

[19] Wu R.-S. W. and Chou P.-H., (2010). "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach", Electronic Commerce Research and Applications, 10 (3): 331-341.

[20] Nikam S. S., (2015). "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental Journal of Computer Science & Technology, 8: 13-19.

[21] Chen W., Xie X., Wang J., Pradhan B., Hong H., Bui D. T., Duan Z. and Ma J., (2017). "A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", CATENA, 151: 147-160.

[22] Zhao Y., Li Y., Zhang L. and Wang Q., (2016). "Groundwater level prediction of landslide based on classification and regression tree", Geodesy and Geodynamics, vol. 7 (5): 348-355.

[23] Mujtaba G., Shuib L., Raj R. G., Rajandram R. and Shaikh K., (2017): "Prediction of cause of death from forensic autopsy reports using text

classification techniques: A comparative study", Journal of Forensic and Legal Medicine, 57: 41-50.

[24] Lynch C. M., Abdollahi B., Fuqua J. D., Carlo A. R. de, Bartholomai J. A., Balgemann R. N., Berkel V. H. and Frieboes H. B., (2017). "Prediction of lung cancer patient survival via supervised machine learning classification techniques", International Journal of Medical Informatics, 108: 1-8.

[25] Z. Shao, S. Yang, F. Gao, K. Zhou and P. Lin, (2017). "A new electricity price prediction strategy using mutual information-based SVM-RFE classification", Renewable and Sustainable Energy Reviews, 70: 330-341.

[26] Cheng H. Y. and Yu C. C., (2015). "Multi-model solar irradiance prediction based on automatic cloud classification", Energy, 91: 579-587.

[27] Gharagheizi F., Mohammadi A. H., Arabloo M. and Shokrollahi A., (2017). "Prediction of sand production onset in petroleum reservoirs", Petroleum, 3 (2): 280-285.

[28] Ammar A. A. and Maheswari D., (2017). "Churn prediction on huge telecom data using hybrid firefly based classification", Egyptian Informatics Journal, 18 (3) 215-220.

[29] Mauša G. and Grbac T. G., (2017). "Co-evolutionary multi-population genetic programming for classification in software defect prediction: An empirical case study", Applied Soft Computing, 55: 331-351.

[30] Uthayakumar J., Vengattaraman T. and Dhavachelvan P., (2017). "Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis", Journal of King Saud University - Computer and Information Sciences, 1319-1578.

[31] Mage L., Baati N., Nanchen A., Stoessel F. and Meyer T., (2017). "A systematic approach for thermal stability predictions of chemicals and their risk assessment: Pattern recognition and compounds classification based on thermal decomposition curves", Process Safety and Environmental Protection, 110: 43-52.

[32] Corani G. and Scanagatta M., (2016). "Air pollution prediction via multi-label classification", Environmental Modelling & Software, 80. 259-264.

[33] Saez J. P., Mande T., Larsen J., Ceperley N. and Rinaldo A., (2017). "Classification and prediction of river network ephemerality and its relevance", Advances in Water Resources, 110: 263-278.

[34] Mesquita D. P., Rocha L. S., Gomes J. P. P. and Neto A. R., (2016). "Classification with reject option for software defect prediction", Applied Soft Computing, 49: 1085-1093.

[35] Banihashemi S., Ding G. and Wang J., (2017). "Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption", Energy Procedia, 110: 371-376.

[36] Mitsa T., (2010). Temporal Data Mining, Taylor and Francis Group, London.

[37] Lasek P., (2011). Efficient Density-Based Clustering, Ph.D THESIS, Warsaw University of Technology, Faculty of Electronics and Information, Warsaw.

[38] Kotu V. and Deshpande B., (2015). Predictive Analytics and Data Mining Concepts and Practice wirh Rapidminer, First Edition, Morgan Kaufmann, Massachusetts.

[39] Özkan Y., (2008). Veri Madenciliği Yöntemleri, First Edition, Papatya, İstanbul.

[40] Nilashi M., Ibrahim O., Ithnin N. and Sarmin N. H., (2015). "A multi-criteria collaborative filtering recommender system for the tourism domain using Expectation Maximization (EM) and PCA–ANFIS", Electronic Commerce Research and Applications, 14(6): 542-562.

[41] Malafaia D., Vieira J. and Tomé A., (2016). "Adaptive Threshold Spectrum Sensing Based on Expectation Maximization Algorithm", Physical Communication, 21: 60-69.

[42] Lin H. Y., (2016). "Gene discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification" Applied Soft Computing, 48, 683-690.

[43] Rosario A. O., Adeli H. and Buford J. A., (2017). "MUSIC-Expected maximization gaussian mixture methodology for clustering and detection of task-related neuronal firing rates", Behavioural Brain Research, 317: 226-236.

[44] Bock H. H., (2008). "Origins and Extensions of The K-Means Algorithm in Cluster Analysis", Electronic Journal for History of Probability and Statistics, 4(2): sayfalar.

[45] Hartigan J. A. and Wong M. A., (1979). "A K-Means Clustering Algorithm" Journal of the Royal Statistical Society. Series C (Applied Statistics), 28 (1): 100-108.

[46] Sisodia D., Singh L., Sisodia S. and Saxena K., (2012). "A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology (IJLTET), 1: 82-87.

[47] Lam D. and Wunsch D. C., (2014) "Clustering", Academic Press Library in Signal Processing, 1: 1115-1149.

[48] Rajendiran S., (2015). Learning Classification Algorithms in Data Mining, Master of Science, California State University, Sacramento.

[49] Wei Y., Zhang X., Shi Y., Xia L., Pan S., Wu J., Han M. and Zhao X., (2018). "A Review of Data-Driven Approaches for Prediction and Classification of Building Energy Consumption" Renewable and Sustainable Energy Reviews, 82 (1): 1027-1047.

[50] Hossin M. and Sulaiman M. N., (2015). "A Review on Evaluation Metrics For Data Classification Evaluations" International Journal of Data Mining & Knowledge Management Process (IJDKP), 5 (2): 1-11

[51] Pyle D., (1999). Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco.

# CURRICULUM VITAE

## PERSONAL INFORMATION

**Name Surname** : Özgürdeniz DÖĞER

**Date of birth and place** : 29.08.1992 Isparta

**Foreign Languages** : English

**E-mail** : ozgurdeniz.doger@gmail.com

## EDUCATION

| Degree | Department | University | Date of Graduation |
|---|---|---|---|
| Undergraduate | Industrial Engineering | Istanbul University | 2014 |
| High School | Sciences | 24 Kasım Anadolu High School | 2010 |

## WORK EXPERIENCE

| Year | Corporation/Institute | Enrollment |
|---|---|---|
| 2018 | Penta Teknoloji | Business Development Specialist |

## PUBLISHMENTS

### Papers

1. Efendigil T. and Döğer Ö., (2017) "Customer Segmentation Process Using Clustering Analysis for IT Reseller Companies", 37. Yöneylem Araştırması Ve Endüstri Mühendisliği Ulusal Kongresi, 1: 124-125