ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

MSC IN INFORMATION SCIENCE PROGRAMME

APPLICATION OF DATA MINING TECHNIQUES FOR CUSTOMER SEGMENTATION IN INSURANCE BUSINESS: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

A THESIS SUBMITTED TO THE SCHOOL OF INFORMATION SCIENCE ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

DANDI MERGA GUTEMA

July, 2016

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

MSC IN INFORMATION SCIENCE PROGRAMME

APPLICATION OF DATA MINING TECHNIQUES FOR CUSTOMER SEGMENTATION IN ISURANCE BUSINESS: THE CASE OF ETHIOPIAN INSURANCE CORPORATION

BY

DANDI MERGA GUTEMA

Name and Signature of Members of Examining Board

NAME	SIGNATURE	DATE
_____	_____	_____
Chairperson, Examining Board		
_____	_____	_____
Advisor		
_____	_____	_____
Examiner		
_____	_____	_____
Examiner		
_____	_____	_____

DECLARATION

The thesis is my original, has not been presented for a degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

_____

DANDI MERGA GUTEMA

July, 2016

The thesis has been submitted for examination with our approval as university advisors.

ERMIAS ABEBE _____ _____

DEDICATION

Dedicated to my beloved father, Merga Gutema and my dear mother, Soretti Duressa.

ACKNOWLEDGEMENTS

First of all, I would like to express my heartfelt thanks to my Almighty God for his overall kindness.

Secondly, my deepest gratitude goes to my advisor, Ato Ermias Abebe, for his constructive comments and inspirations throughout this work. Without his genuine advice this research couldn't have come into an end. I am very thankful to Dr. Million Meshesha and Dr. Tibebe Beshah for their support.

Thirdly, I want to thank W/ro, Mirchaye from ICTM department of EIC at LAD, who helped me to get the dataset for this study. To be frank, it is impossible without her assistance to complete this study.

Next, I am grateful to my families-Dad, Mom, Itag, Adado, Bore, Abo and AK, who have rendered me all their care, love and encouragement. Without their sustainable support I couldn't have reached at this level of my achievement.

Then, my special thanks go to all of my friends-Benji, Abela, Lati, Joe, Wakshe, Ashe, Nati, Temu, Areb, Mame, Milli, Bito, Jewiscus, Bonsa, Dinqa, Abdi, Dave and Afe who encouraged me by sparing their time, effort and resource to support me. Horaa Bulaaa!!!

Am also indebted to my friends from Debre Markos-Abrilo, Ali, Mulesta, Kehase, Beka, Abela, Roza, Abe and Getamesay who helped, inspired and understood me when I was in trouble.

Finally, I would like to extend my gratitude to all the staff members of the Information Science. Stay blessed all! No more words!

Table of Contents

LIST OF FIGURES, TABLES AND EQUATIONS

List of Tables

List of Figures

Equations

LISTS OF ABBREVATIONS

ARFF- Attribute-Relation File Format

CLV- Customer Lifetime Value

CRISP-DM- Cross-Industry Standard Process for Data Mining

DT -Decision Tree

EIC -Ethiopian Insurance Corporation

ICTM- Information and Communication Technology Management

LAD - Life Addis District

LTV -Lifetime Value

ABSTRACT

The aim of this study is to apply data mining techniques in insurance business to build models that can segment customers based on their value. The study subject for this research is Ethiopian Insurance Corporation, which stores life insurance policy holders' data in LIFE INSIS database located at Life Addis District were selected

To meet the aforementioned objective of the study, the CRISP-DM methodology, which involves six steps was adopted to undertake data mining process and to address the business problem systematically and iteratively. During the business understanding phase, business practices of EIC life insurance were assessed using interviews with business and technical experts, and document analysis.

Through data understanding and preparation phases, information on the subject of policyholders' personal, demographic, policy coverage and transactional was taken in to account. Besides, the attributes selected were considered the degree of relevancy to develop value-based customer segmentation model using DM techniques. Accordingly, from LIFE INSIS database, 27845 records and 16 attributes were imported MS-excel**.** The data used in this study were related to one year (12 months) of customer interactions that found between August, 2011 to August, 2012 time-frame. Attributes such as **occupation ID, marital status, and sector** were removed because they showed high Missing Values. The preprocessing tasks such as handling outliers and noisy, data integration and data transformation were undertaken. And, customers' value was computed using individual policyholders' records that indicate their insured value, duration and the cost incurred attract them (agent commission) information. With consultation of experts, 7 attributes and 21622 records were included in the final datasets for modeling purpose the initial database

To build the customer segmentation models, K-means clustering algorithm and J48 decision tree algorithms of WEKA implementations were selected to discover useful patterns and to analyse the data. K-means clustering algorithm was selected since it's capable to develop models that segment customers with similar characteristics while J48 Decision tree classification technique was applied due to its quite quality and articulacy to decipher the cluster models by assigning

each record to the target variable. Besides, patterns revealed that DT models are very easy straightforward and useful to integrate with business practices, and understand the revealed clusters. As a result, the experiments made in build DT model revealed that attributes such as ***age and insured_value*** were automatically selected as best predictive attributes to split the datasets to sub-segments that have homogenous characteristics based on their value (high or low).

The results of the research pointed out that the customer segmentation models built by using the combination of classification and clustering data mining techniques are necessary for the LAD and marketing department of EIC in order to identify the valuable segments of customers and other factors underlying variations of the customers' values.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Currently, most companies are changing their approach from product centric to customer centric. In customer-centric approach, the company considers customers as assets and focuses on acquiring and retaining more customers [1]. Customer base can be seen as an asset that has a value and can help in deciding acquisition and retention. The customer base can be a source of competitive advantage for the company. In recent times, more and more companies are becoming service companies, thus enhancing the importance of the customer-centric approach.

The advent of relationship marketing along with technological advancements has led to Customer Relationship Management (CRM). Customer relationship management (CRM) is the overall process of exploiting customer-related information and using it to enhance the revenue flow from existing customers [2]. As modern economies become predominantly service- based, companies increasingly derive revenue from the creation and sustenance of long-term relationships with their customers. In such an environment, marketing serves the purpose of maximizing customer lifetime value (CLV) and customer equity, which is the sum of the life-time values of the company's customers. Hence, measuring CLV can also use for market segmentation and for allocation of marketing resources for acquisition, retention, and cross-selling [3].

Marketing is the art of attracting and keeping profitable customers. It seems logical to choose your customers properly for both acquisition and retention. One way of doing so can be through Customer Lifetime Value (CLV). A profitable customer is defined as 'a person, household, or company whose revenues over time exceed, by an acceptable amount, the company cost of attracting, selling, and servicing that customer. The excess is called CLV which is also known by other names such as customer profitability, lifetime value (LTV), customer equity and so on [1]. According to Rodpysh,, Aghai and Majdi [4], data of CLV could help companies to decide better about strategies development and resource allocation to each customer.

Data mining is a powerful technology with great potential to help insurance firms focus on the most important information in the data they have collected about the behaviour of their existing and potential customers. Data mining assists insurance sector in predicting fraudulent claims, medical coverage and predicting customers buying pattern. Data mining is applied in the insurance industry and tremendous competitive advantages accrue to those companies who have implemented it successfully. Some of the areas data mining can be applied to insurance industry are in identifying risk factors that predict profits, claims and losses, customer level analysis, marketing and sales analysis, developing new product lines, reinsurance, financial analysis ,estimating outstanding claims provision and detecting fraud etc.[5].

The aim of this study is to assess the applicability of data mining techniques in insurance industry to build models that can segment customers based on value.

1.2 Back ground of the organization

Ethiopian Insurance Corporation (EIC) was established on 1st January 1976 under Proclamation No. 68/1975. Since then, the corporation expanded 61 distribution channels throughout the country. 38 years of experience backed by 1,147 qualified and experienced manpower. EIC provides Life, Property and Liability Insurance covers. And it market share is regarded as lion market share because of its strong and reliable financial standing. It built longstanding and strong affiliation with many international insurance organizations and associations.

It maintained a comprehensive range of outward reinsurance contract, and accepting inward reinsurance (including Co-insurance) business on selective basis. It has been also engaged in different investment areas.

EIC has set a strategic goal to increase the number of policy issued by the corporation. In fulfilling its purpose of establishment, to provide insurance service to the public at large, and also spread the risk by increasing the number of policyholders, and hence become more profitable.

In July 1, 2011 EIC has fully implemented the insurance software packages that are the first in kind for the insurance industry in the Country. The insurance packages (Life-INSIS and INSIS

Non-life/ GENERAL-INSIS) have been fully operational. Moreover, those who have IT background and well acquainted with the system have been selected and assigned for certain period in each district offices found in Addis Ababa and out of Addis Ababa to assist the staffs in using the system.

The marketing department of the EIC is responsible for CRM implementation in order to create, maintain and expand customer relationships. For instance, to acquire new customers, the sales forces or marketing agents of EIC have been trained to increase the accessibility of insurance service.

1.3 Statement of the Problem

According to Du [6], the recent customer relation analyses have some serious drawbacks. The most important one is that based on those analyses company usually consider the customer as an isolated object and having value only when he/she deals with this company. Neglect the network value of each customer and the value from potential purchase probability.

The benefits of CRM have been enjoyed by different organizations for many years, so effective customer relationship management in the companies should have a mechanism to increase the value of their customers. However, some organizations are still unable to succeed in this domain because of various reasons including the recent and the prominent issues which are not followed in their marketing strategies. These issues could consist of problems related to customers' satisfaction, retention, acquisition and so on. Accordingly, it is important for organizations to pave solutions for the existing problems in order to retain and maximize the profitability of their existing customers, rather than simply acquiring new ones. For instance, insurance companies can evaluate their customers' values in order to enhance their decision making process in dealing with different aspects of the company.

Insurance companies have vital roles in providing insurances-life, health and non-life, which meet the requirements of the customers. According to Umamaheswari and Janakiraman [5], insurance is the equitable transfer of the risk of a loss, from one entity to another in exchange for payment. It is against the risk of a contingent uncertain loss. In doing so, growing interest

towards insurance among people, innovative products and distribution channels sustain the growth of the insurance sector.

In our country's context, many researches have been done in the application of data mining for various purposes. But few researches have been conducted in the insurance domain. For instance, Hintsay [7], conducted a research on the use of data mining technology as to support in business decision-making is growing fast. He attempted to design predictive modelling by using data mining techniques in support of insurance risk assessment. He aimed at designing a computer system which warns early for motor insurance. Additionally, Fikre [8] developed predictive model for policy renewal of personal accident policies. He said that, the absence powerful models formed a problem to decide which customer could get insurance coverage. Thus they can accept or reject the contract with such customers using DM technology such as decision tree classification technique.

Based on the aforementioned issues, data mining technology can help the insurance firms in taking crucial business decisions. With the help of data mining techniques, the customers' data can be handled effectively. With regard to this, Tsiptsis and Chorianopoulos [9] stated that data mining helps the market specialists for decision making process. Therefore, discovering the hidden knowledge and patterns from the database can support the marketing managers to make better decision and allocate resource. Besides, the information from the result serves to increase the competitive advantage of the company, and the prediction from data mining techniques increases customers' satisfaction and ensure the financial security of the insurance company. For better success of insurance policy of the organization a mechanism to explore the value of customers which reside in customer relationship management is required. Thus, it is essential to have enough information about the current customers and their behaviour.

Currently, EIC operationalized customer database for both life and non-life insurance namely; LIFE INSIS and NON-LIFE INSIS insurance packages, to handle customer information and to support the tasks of customer service delivery. The databases have been used by domain experts and managers to administer for underwriting, claim and marketing agents data. The historical data of existing customer demographic and transactional data are mostly found in LIFE INSIS

data base, which handle policy holders of life insurance information. Mainly, the life insurance policies are established by considering, term, whole, and endowment life insurance plans to share unexpected and unpredicted with only economic losses caused by death, disability and old age. However human life is measured by monetary value, life insurance offers a guaranteed sum to the insured or its dependants for insurable risks, which are measured by the payment made the insured or the earning policy holders. Yet, information related to the insured the risk covered, the time the period covered by the policy, the payment made by the insured, the expense incurred to attract the insured ones, the demographic and behavioural data could be used to assess the profitability of the existing customer and to underlying patterns categorize and predict underlying customers based on their value. However, the absence using the past data for understanding and classifying the customers data based on their value, is one of the challenge that EIC marketing managers facing. Moreover, inadequate system that resolves the difficulty of customer profitability analysis is the other problem noticed in the corporation. Due to this, a model that could help to segment life insurance customers of EIC based up on the customer value data is required.

Presently, the EIC life insurance's benefits and the cost are measured and determined by the corporation's underwriting (risk assessment) aspects and premium system, which involves lengthy process. The corporation's risk assessment system assists to identify policyholders involving high or low risks. However, establishing customer segmentation scheme based on their current and potential value of the existing policyholders who contribute high or low value needs to be conducted in order to capitalize its economic value and maintain strong customer relationship strategies (customer retention and acquisition) across policyholders' lifetime. Regarding these issues data mining technologies could help to handle the customer segmentation problem using the customers' value. Based on the value of customers, the DM clustering techniques could help to develop descriptive models that allow visualizing similar characteristics that customers exhibit within the group and the distinguishing characteristics from the other group. Moreover the DM classification technique could also support to build predictive models by utilizing the result of clustering models to assign each cluster records to target variable for predicting the value of customers'.

In this study, an attempt was made to investigate the applications of data mining techniques that would support to build meaningful segments made up of personal, demographic, policy and transactional data and to categorize or profile customers of EIC's life insurance based on customer value.

The study was intended to pave ways which could help marketing managers to identify valuable customer groups that would support their decision making practice in CRM. This could in turn promote effective utilization of their resources in order to maximize the corporation's profit and strengthen relationships with the customers by considering their value. By taking the above issues in to account, there is a need for system to handle the problems concerning customers' value. Therefore, the research was aimed to pave answers for the following research questions.

- ✓ What are the most significant variables that help to segment life insurance customers?
- ✓ What are the basic elements of customer lifetime value which enable to segment customer in the data base of in EIC?
- ✓ How data mining techniques and algorithms can be used for dealing with customers' segmentation based on their value?

1.4 Objective of the Study

1.4.1 General Objective

The general objective the study is to apply data mining technique for customer segmentation based on the customers' value in Ethiopian Insurance Corporation.

1.4.2 Specific Objectives

To achieve the general objective stated above, this study formulated the following specific objectives:

- ❖ To understand the business and identify the sources of data stored at the corporation databases
- ❖ To collect and to understand the customer data based on the problem identified
- ❖ To apply preprocessing tasks to the selected data and prepare dataset for segmentation
- ❖ To explore and select appropriate data mining tool, technique and algorithm that supports customer segmentation

- To build descriptive and predictive models for customer segmentation based on customers value
- To evaluate the results of the models with meaningful justification.

1.5 Significance of the study

The aim of this study is to assess the applicability of data mining techniques in the insurance industry to build models that can segment customers based on the value they contribute. Based on these, the subsequent benefits can be gained from the finding of this study.

Primarily, the researcher gained an experience of conducting a research as this study was conducted for academic purpose; hence, the finding of this study, could motivate other researchers to conduct further researches in the area.

Secondly, the result of the study could help EIC to manage customer data and gain business advantage. It may also improve the insurance business process. Thus, the insurer, in this study EIC, can be a beneficiary by visualizing the characteristics of customer groups and classifying valuable customers. Besides, the findings of the study can be used by other organizations dealing with insurance businesses.

Thirdly, the findings of the study can be used by other organizations dealing with insurance businesses. The main objective of life insurance was to protect the insured by sharing of risks against unpredicted losses such as old age, disability and death. The models built under this study can help to understand the most valuable customers and the factors (attributes) that influence their life time value.

In the other way, the result of the study can also contribute to the study area to build customer segmentation models of data mining application for insurance companies in order to maximize the value of customers in their marketing strategies.

For the university, it may serve as documentation for further investigation.

The finding of this study can be used by insurers to increase the quality of service given to its policy holders in order to maintain the standard or the life quality of the insured ones. In other words, the customers can be beneficiaries of the quality service provision.

Moreover, the study findings can provide insight for further researches to apply data mining technologies to advance insurance industry.

1.6 Scope and Limitations of the Study

In this study, data mining techniques were examined to develop interesting segments and to extract meaningful patterns based on their demographic, personal, policy and transactional data, which were gathered from LIFE INSIS data base of EIC life insurance at Life Addis District (LAD). Besides, the insurance types of individual policy holders' were selected along with their insurance and cover types. For this study, a one year (from August, 2011 to August, 2012) historical data were extracted to build descriptive and predictive value-based customer segmentation model using DM techniques. Due to the unavailability of historical data, the research was conducted using life insurance.

During the study, CRISP-DM was adopted to undertake the data mining process. The models of customer segmentation were developed by; descriptive models using k-means clustering algorithm and predictive models using J48 decision tree classification algorithm.

In conducting the study, there were some limiting factors such as schedule and budget in accordance with the objective and aim set by the researcher. Besides, attributes such as insurance premium value were not included in the dataset because it was sensitive attribute that involve protection against customers' privacy and business secrecy. Thus, this study is limited to accessing the premium payments made by customers, which useful to measure the value of customers. This in turn affects the result of this thesis work.

In addition, due to the lack of accessing historical data of customers, the data collected for this study was extracted from the customer data base located at EIC LAD. As a result, it restricts the study to incorporate data from the different data base of the corporation branches. In line with this, for this the study, 12 month historical data was used for modelling purpose. This is not sufficient to evaluate give details the value customers. Further studies need to consider these issues.

1.7 Organization of the study

This thesis is organized into six chapters. The first chapter is introduces a brief background about subject of study and the problem identified to conduct this thesis. It also includes the objective, significance, scope and limitation of the study. Chapter 2 covers review of relevant literatures on data mining definition, processes, tasks, techniques and algorithm .It also gives brief description of customer segmentation and customer value concepts along with the DM techniques applications in the insurance domain. In addition, related works in the area of CRM and the problem domain of this study are reviewed. Under Chapter 3, the methodology research is going to be discussed with brief explanation. Chapter 4 consists of three stages CRISP-DM, which defined in the study context namely as; business understanding, data understanding and data preparation. In chapter 5, the model building and evaluation processes are going to be discussed more in detail. The customer segmentation models developed by K-Means clustering and J48 decision tree classification algorithms are presented with different parameters and evaluated using appropriate metrics. The final chapter, chapter 6 presents the summary of the finding, the conclusions drawn from the study and the recommendations for further researches.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Data Mining

"Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions [11]."

According to Berry and Linoff ,[12] "data mining is an emergent and rising area of research and development, both in academic world as well as in business, connecting interdisciplinary studies and development adjacent to diverse domains". They stated that data mining can be seen from two perspectives. From a broader perspective data mining can enhance a better decision making in the business activity through the process of understanding the environment in which that the business operates and to gain better economic advantage over their competitors. from a narrow perspective , as the business industry providing more customer targeted products and service data mining technologies (tools ,techniques and algorithms) can maintain and evaluate continuous relationship between companies and their customers [12].

Rygielski, Wang and Yen, [13] also defined "data mining is a sophisticated data search capability that uses statistical algorithms to discover patterns and correlations in data". Data mining can find out and extract useful from information which are hidden in large databases to reveal the unseen patterns and relationships. They stated that data mining is one knowledge discovery process steps. According to them there are six basic data mining models that can deal with business problem. These models are called classification, regression, time series, clustering, association analysis, and sequence discovery. These models help to make prediction, to forecast and to resolve business difficulties based the information provided [13].

2.2 Data Mining and OLAP

Different technologies have been used to analyze data stored in the database or data warehouse to facilitate decision making process. The following discussion is an overview of OLAP with respect to data mining technology.

According to Han and Kamber, [14] OLAP (Online Analytical Processing) provides an analysis modeling to view an aggregate or summarized data which are found either in data base or in data

ware house . So, they stated that the ability of OLAP to offer careful extraction of data from various view of point established a ground a data mining discipline. Therefore the summarized data simplifies data mining process.

As data mining goal is to extract the hidden knowledge from huge amounts of data for a better decision making, OLAP also goal is to supply an aggregated data in order to look from different directions or dimensions. To strengthen relationship between data mining and OLAP Han and Kamber, [14] stated that "Prior to acting on the pattern uncovered by data mining, an analyst may use OLAP in order to determine the implications of using the discovered pattern in governing a decision". Therefore, data mining and OLAP can be seen technologies that are interlinked [14].

They also they stated that the other factor for both integration is the invention of OLAM (Online Analytical Mining) also referred to as OLAP mining. "OLAM systems are particularly important because most data mining tools need to work on integrated, consistent, and cleaned data, which again, requires costly data cleaning, data transformation and data integration as pre-processing steps [7]".

	Basic Function	**Characteristics**
Data Mining	To reveal hidden information	Data driven or business -driven
Data Warehouse	Decision making	Subject-oriented, integrated, time-variant and nonvolatile collection of data
OLAP	Data summarization /aggregation tool	Extraction of data from various view of point

Table 1 Data Mining, Data Warehouse and OLAP

2.3 Data Mining and Data Warehousing

Han and Kamber [14] stated that data mining should be applicable to any kind of data repository. Data warehouse is a repository of information, which are collected from many sources and stored in a unified schema to facilitate decision making process. The following discussion presents the overview of data warehouse technology.

"A data warehouse (DW) is a database that is maintained separately from the organization's operational database for the purpose of decision support" [14]. The process of constructing and using data warehouse is called Data warehousing." A DW provides integrated, enterprise-wide, historical data and focuses on providing support for decision makers with respect to data modeling and analysis". The main characteristics of DW are; it is a subject-oriented, integrated, time-variant and nonvolatile collection of data in order to maintain decision-making process. A DW is usually established through an incremental process. DW is not static rather it is dynamic process .The establishment of data warehouse can be described as an incremental or evolutionary process. These evolutionary steps are performed through creation of virtual data warehouse, the data mart, and the enterprise warehouse. They are described as follows [14]:

I. **Virtual Data Warehouse**.

This step involves creation of a view over operational data warehouse. However it is simple to build, it needs excess computational capacity of the operational database system. Middleware tools can allow users to access the data.

II. **Data Mart**.

The data mart contains a particular data which is valuable to targeted groups. For instance, the data can be human resource or customer relationship management data. So that in organization data mart only handles data which only concerns the specific group.

III. **Enterprise Warehouse**.

Unlike the data mart, enterprise warehouse contains data or information that flows in all over the entire organization. To design and to build enterprise warehouse is not an easy work because the to design the integration of system from that range from operational level to top level system , from internal system to the external system may takes several years . So it is a difficult task.

2.4 Data Mining Process

2.4.1 Data Mining and Knowledge Discovery Process (KDP)

There is confusion between the concepts of Data Mining and Knowledge Discovery Process. Many researchers use DM as a synonym for knowledge discovery; DM is also just one step of the KDP [4].

According to Cios, Pedrycz, Swiniarski and Kurgan, [15] "the knowledge discovery process (KDP), also called knowledge discovery in databases seeks for new knowledge in some application domain." So that KDD involves some activities to expose hidden knowledge in databases in order to make a better decision making and gain a business advantage. They also stated that the process of KDD shall meet four basic criteria. Those are: identifying valid, novel, useful, and ultimately understandable patterns in data. Thus, the process of KDD attempted to cover the following basic tasks [15].

- Understand Data storage and access process
- Utilize efficient algorithms for analysis of huge amount of data.
- Constitute interpretation and visualization
- Select appropriate model that an enhance a better way of doing things
- Facilitate learning in the domain area

However data mining is considered as one of the KDD process Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, [16] defined it as "the process of using DM methods to extract what is deemed knowledge according to the specification of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformation of the database." Based on the aforementioned definition, they classify the stages of data mining in to five stages. These are stated below [16].

- Selection – This stage consists on creating a target data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.
- Preprocessing – This stage consists on the target data cleaning and preprocessing in order to obtain consistent data.
- Transformation – This stage consists on the transformation of the data using dimensionality reduction or transformation methods.
- Data Mining – This stage consists on the searching for patterns of interest in a particular representational form, depending on the data mining objective (usually, prediction)
- Interpretation/Evaluation – This stage consists on the interpretation and evaluation of the mined patterns.

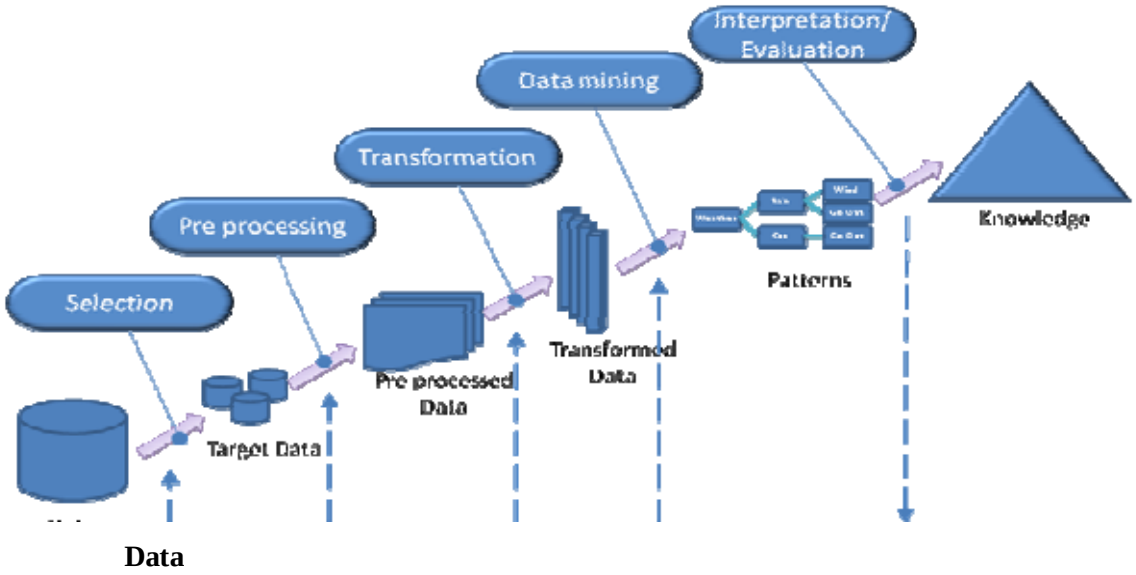


Figure 1 KDD process cycle (adapted from Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, [16])

2.4.2 The CRISP-DM process

According to [10], the CRISP-DM process was developed by the means of the effort of a consortium initially composed with Daimler Chryrler, SPSS and NCR. CRISP-DM stands for CRoss-Industry Standard Process for Data Mining. It consists on a cycle that comprises six stages:

- Business understanding – In this phase, one has to understand the objective of the domain area and figure out the available of requirements from business point of view. After assessing the current situation, the next task involves changing the business problem in to data mining problem and goal. Finally preliminary plan needs to conduct in order to meet the objective that has been set.

- Data understanding – This phase involves tasks such as collecting initial datasets and understanding the relationship between the attributes of datasets .So that one can get more insight with it. And also the description of data and the quality of data should be clearly justified.
- Data preparation – The major goal of data preparation phase is to create final data set in order to be used by modeling tools. In other words it provides final datasets from primary

data for modeling phase for further course of action. Among major tasks of this phase data cleansing is one of them. It enables to select appropriate e attributes, records based on the constituted goal of data mining. So that activities such as missing values, outliers, noisy data and inconsistent data should be done. Then after data integration and data transformation tasks are followed. Finally the data set should be in an acceptable format in which the modeling tools can recognize it .this doesn't mean it change the data meaning rather its forms suitable set up for the next phase.

- Modeling – In modeling phase one should select modeling techniques from the available ones .for instance neural network, decision tree are popular. The dataset should be classified into training and test sets for modeling purpose. Among the modeling technique. In this phase one thing should be noted is the selection modeling technique should fit with the goal of data mining.
- Evaluation – During the evaluation phase, the performance and effectiveness of the models created in the previous phase are going to be assessed based on particular metrics. Beside this, measurement and analysis will be made on the model constructed with regard to the goal of the business. If the model doesn't meet the objective of the business, as the below picture depicts, new iteration will be established. So that the initial phase is again going to be restudied. Otherwise the next phase called deployment is going to be performed.
- Deployment – Finally, after the above steps, the extracted knowledge from the enterprise data base should be managed and provided to target or end users of the system. In other words, the model constructed may then be operated in the existing business activity. In addition documentation is also required for further studies. More of the explanation of data mining process will be explained in the next chapters.

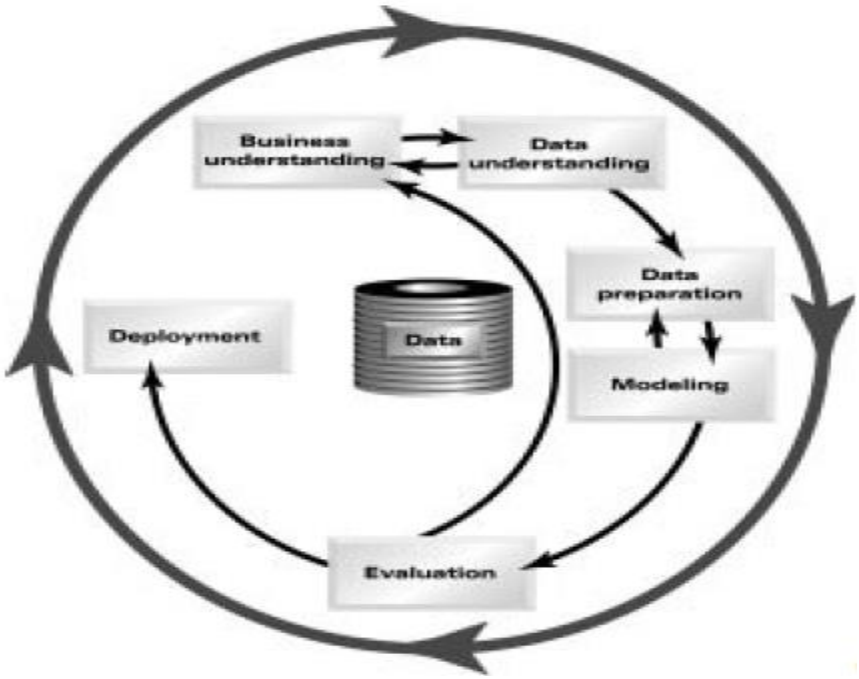


Figure 2 CRISP-DM process (adapted from Chapman, Clinton, Kerber, Khabaza, Reinart, Shearer, and Wirth, [10])

The following table shows phases of KDD and CRISP-DM

KDD process	**CRISP-DM process**
1. Selection	1. Business understanding
2. Pre processing	2. Data understanding
3. Transformation	3. Data preparation
4. Data mining	4. Modeling
5. Interpretation	5. Evaluation
6. Post KDD	6. Deployment

Table 2 Summary of phases in KDD and CRISP-DM

2.4.3 Hybrid Models

According to Cios, Pedrycz, Swiniarski and Kurgan [15], the hybrid models are enhanced the knowledge discovery process by combining the academic and industrial models in data mining projects. The development hybrid models was adopted from the CRISP-DM model as its can be used for academic research. Thus, these models are research-oriented, which present data mining step than the modeling step. The six steps of hybrid models allow a number of feedback mechanisms. Moreover, the knowledge discovered in the final step for a specific domain may be applied in other domains. The following descriptions present the six steps of hybrid models.

- Understanding of the problem domain- The initial step involves task such as the problem definition and project goal determination, identification of key people and grasping the current solution to the problem through close consultation with the domain experts. then the project goals are transformed to DM goal and the preliminary selection of DM tools to be used in the study is conducted
- Understanding of the data- This step involves tasks such as the collection of data and choosing the size and format of the datasets. Furthermore to the quality of the data are assessed by checking the completeness, redundancy, missing values, plausibility of attribute values, etc. lastly, the usefulness of the data are verified with respect to the DM goals.
- Preparation of the data-In this step, the data going to be used are prepared to apply the DM methods. It consist of tasks such as sampling, testing the correlation and significance of the data, cleaning the data, checking the completeness of the tuples, handling noisy and missing values. Then, the dimensionality of the data is reduced by feature selection and extraction algorithms. This step also comprises, the derivation new attributes, summarization of the data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purpose
- Data mining- This step involves DM methods are applied on the preprocessed data to discover knowledge.
- Evaluation of the discovered knowledge- In this step the results of DM models are evaluated whether the discovered knowledge is novel and interesting and the results of the models are interpreted with respect to domain experts' knowledge. In addition, the approved models are taken and the whole process is revised to pinpoint an alternative

solution, in order to improve the results achieved. Finally, the errors arisen in the process are listed and arranged.

- Use of the discovered knowledge- the final step comprises planning the regarding the usage of the discovered knowledge. The knowledge discovered in the current domain may be applied in other domains. Also, a plan is created concerning the implementation of the knowledge discovered and the documentation of the whole project. Lastly, the deployment of the model takes place.

2.4.4 SEMMA

According to Olson and Delen [17], SEMMA is data mining process steps, which is developed by SAS institute to give solution to data mining projects. The name itself stands for the process known as Sample, Explore, Modify, Model, and Assess. It provides statistical and visualization techniques, to undertake data mining process. It consists of five stages along with iterative experimentation cycle [17].

- Sample – During this stage, representative sample data are extracted from the portion of a large data, which is large enough to contain important data and yet small enough to manipulate quickly.
- Explore – This stage helps the user for better understanding of the data set through exploration of the data by searching for unanticipated trends and anomalies. It enhance to visualize the data for discovery process
- Modify – The aim of this stages is to undertake necessary adjustments to the data through creating, selecting, and transforming the variables for model construction purpose.
- Model –This stage involves construction models using appropriate modeling techniques that can explain patterns in the data.
- Assess – Finally, the usefulness and reliability of the models needs to be assessed and evaluated. It helps to estimate the performance of the models.

2.5 Data Mining Tasks

Data mining tasks are used by considering the domain of the problem and the structure of the patterns expected to be extracted. Using a single method of DM for all problems is very difficult.

Thus, the selection of data mining method depends on the particular problem. Most common data mining tasks are classification, clustering, association rules and sequential pattern

2.5.1. Classification

Among different data mining tasks, Classification model is one of them. It is also known as supervised learning. This supervised method of data mining technique can predicts the continuous numeric and nominal attributes values. The attributes are further divided in two namely; the input and output fields. Thus, the inputs are used to predict the outcome of the output fields [9].

Classification encompasses two levels: classifier construction and the usage of the classifier constructed. The former is concerned with the building of a classification model by describing a set of predetermined classes from a training set as a result of learning from that dataset. Each sample in the training set is assumed to belong to a predefined class, as determined by the class attribute label. The model is represented as classification rules, decision trees, or mathematical formula. The later involves the use of a classifier built to predict or classify unknown objects based on the patterns observed in the training set [18].

Classification maps data into predefined group or classes. Because the classes are determined before examining the data, classification is often considered as supervised learning. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data which are already known to belong to the classes [6].

2.5.2 Clustering

Clustering is a data mining tasks that segments a heterogeneous population that have common characteristics in to subgroups. What makes clustering different from classification is that it doesn't depend on predefined classes [12]. In another words, Han and Kamber [14] said that "the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. That

is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived"

2.5.3 Association rule

Association rule is unsupervised data mining task that do not involve a single output field to be predicted. Rather it detects the associations between discrete events, products, or attributes. In addition, it supports to analyze past co-occurrences of events, purchases, or attributes and discover associations. Mostly it is used for market basket analysis. According to Guo [19], given a collection of items and a set of records containing some of these items, association discovery techniques discover the rules to identify affinities among the collection of items as reflected in the examined records.

2.5.4 Sequential patterns

In fact discovering sequential patterns is similar to the mining of association rules. However, in sequential patterns discovery time element is taken into consideration. Tsiptsis and A. Chorianopoulos [9] said that "they take into account the order of events and detect sequential associations that lead to specific outcomes." They show the link or the correlation between transactions while association rules represent intra transaction relationships. Mostly, it is applied to control inventory in retail stores, to predict certain kind of disease based on patients' history and to discover user access patterns for the web sites.

2.6 Data Mining Techniques for Customer Segmentation

Currently, CRM and marketing departments are tied up with customer segmentation approaches. The first question to be raised is, why firms need to segment their customers?. What is this customer segmentation? According to Tsiptsis and Chorianopoulos [9], customer segmentation is the process of dividing customers to different groups and similar groups based on the various attributes (variables) and characteristics that can yield meaningful and valuable information. Based on this beneficiaries (by characterizing their customers'), firms can use customer segmentation as market differentiation instrument to establish distinguished strategies that helps them for better understanding of their customers' and making decisions. In these days, the

approach of customer segmentation transformed from developing product and service and dominating the marketing scheme to building customer relationship scheme.

In these days, firms use this approach to increase their profitability, to raise their competitive advantage in dynamic markets, through assessing, identifying, concentrating and analysing their customers' needs, wants and preferences. Hence, they can build more understandable and meaningful segments that simplify their marketing activities such as targeting, promotion and positioning [9].

The process of customers' segmentation should contemplate organizational objective of the firm. There different types of customer segmentation that been used for various companies depending on their business context and attributes used for segmentation. For instances customers can be segmented based on their life time value, socio-demographic and life-stage information, and their behavioural, need/attitudinal, and loyalty characteristics.

According to Berry and Linoff [12], customer segmentation is a popular application of data mining with established customers. Data mining techniques can offer customer insight, which is vital for establishing an effective CRM strategies and resource allocation. As a result, DM technology can help to extract useful patterns and knowledge from customers' database to analyse, visualize and predict customers' characteristics. As many studies revealed, DM techniques such as classification and clustering are mostly used for customer segmentation and profiling [12, 9].

2.6.1 Customer Segmentation and Clustering Techniques

Berry and Linoff, [12] defined clustering as "task of segmenting a heterogeneous population into a number of more homogeneous subgroups or cluster." In customer segmentation, clustering techniques can identify meaningful natural groupings of customers' records and they can help to group customers into distinct segments that have similar characters'. Particularly, cluster models are developed when there is need to combine attributes in large dataset. What makes clustering techniques different from other DM techniques is, there is no predefined class label rather the

records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters [12, 9].

Tsiptsis and Chorianopoulos, [9] also said that clustering techniques examine the homogeneity of records or customers with respect to the clustering fields and assign them to the revealed clusters accordingly. It is aimed at detecting groups with internal homogeneity and interclass heterogeneity. It is a DM technique has been commonly used for market researches to develop customer segmentation based attributes on their life time value (LTV), socio-demographic and life-stage information, and their behavioural, need/attitudinal, and loyalty characteristics. As it was it was mentioned above, this technique is quite capable to manage a large number of attribute to create data driven segments (as there is no predefined classes, which are not known in advance). Hence, clusters are induced by the observed data patterns and, built properly to integrate the results with real business meaning and value [9].

There are different clustering algorithms that support to extract homogenous or similar groups from customers' records. But they differ from one another in approach and for the purpose they were built. Therefore, it requires careful assessments before they are deployed to solve the business problems. For instance, agglomerative methods, partitioning methods, density-based methods grid based methods and constraint –based method and so on.

For instance, distance-based clustering technique such as k-means is the most efficient and the fastest clustering algorithm that can handle both long (many records) and wide datasets (many data dimensions and input fields). It is a distance-based clustering technique and, unlike the hierarchical algorithm, it does not need to calculate the distances between all pairs of records. Thus its speed and scalability make the algorithm to be used widely. An also unlike hierarchical it doesn't take many times to build the model. The following descriptions presented some of clustering algorithm, which help to develop clustering models.

2.6.1.1 K-Means Clustering Technique

The K means clustering algorithm is one of the unsupervised learning methods which cluster the subsets of the dataset to the nearest mean or to the centroid. According to Tsiptsis and Chorianopoulos [9], "K-Means is one of the most popular clustering algorithms. It starts with an

initial cluster solution which is updated and adjusted until no further refinement is possible (or until the iterations exceed a specified number). And each of iteration refines solutions by reducing the within-cluster variation."

The advantage of using k-means is because: produces may produce tighter clusters than hierarchical clustering, especially if the clusters are globular [20]. Unlike the other algorithms, this algorithm required to determine the k value in advance. So during the model development experts and analysts are required to make trials. The basic steps followed in k- means clustering algorithm are as follows. These are [12, 20]

Step-1 choose the k initial records as cluster centre's (initial seeds) randomly and assigns each record to it's the closest cluster

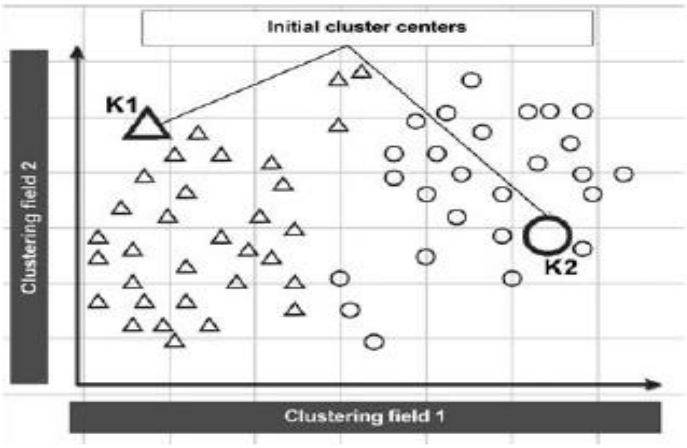


Figure 3 Initial cluster centers (adapted from Tsiptsis and Chorianopoulo, [9])

Initially, it randomly selects the k (the number cluster centers /centroids), one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

Step-2. As new records are added to the clusters, the cluster centres are recalculated to reflect their new members.

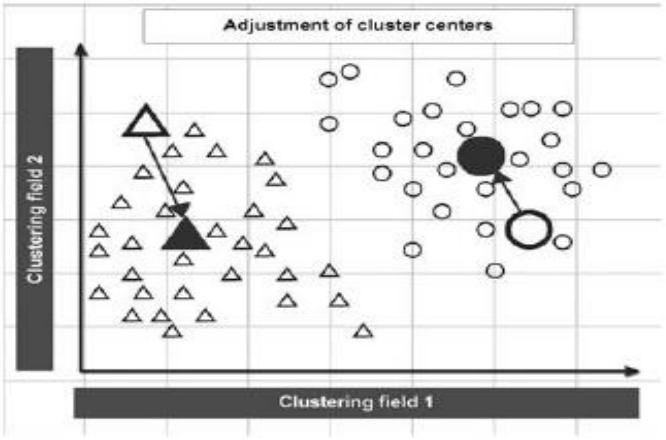


Figure 4 Adjustment of cluster centres (adapted from Tsiptsis and Chorianopoulo, [9])

The second step involves taking each point belonging to a given data set and associating it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. Step 3 update the centroid or reassigned to the adjusted clusters.

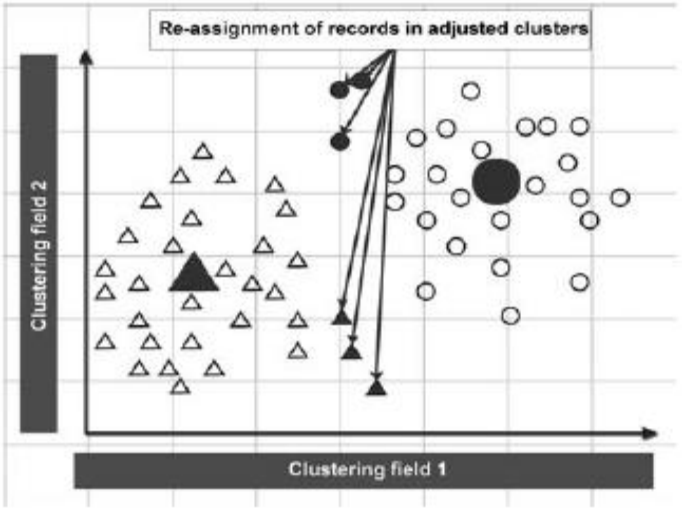


Figure 5 Re assignment of records in adjusted clusters adapted from Tsiptsis and Chorianopoulo, [9])

At this point we need to recalculate the k new centroids as bar centers of the clusters resulting from the previous step

Step 4 -Step back to go back to Step 2, stop when there are no more new assignments.

The iteration continues until it converges and the migration of records between clusters no longer refines the solution.

2.6.1.2 Agglomerative or hierarchical

Among the clustering algorithms, hierarchical algorithm is the oldest one. It finds new clusters using previously found ones. As its name indicates, each record starts as separate cluster and then gradually merged or grouped in to super cluster. The distance between all pairs of records is calculated in order to find most similar clusters.

Generally, the hierarchical algorithm works as follows:

1. First , it find the two closest objects and merge them into a cluster
2. Then, it finds and merges the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains, return to step 2

The drawback of hierarchical algorithm is that it cannot handle more than a few thousand cases effectively. Thus, sampling the cluster population is required. This task is time consuming and is not an ideal to sample cluster population. Therefore, it is challenging to apply it for business clustering tasks. Yet, other clustering algorithm such as k-means can handle millions of records without sampling.

2.6.1.3 Kohonen network/Self-Organizing Map (SOM)

Another clustering algorithm is Kohonen networks or Self-Organizing Map (SOM) which works based on neural networks that usually used to produces a two-dimensional grid or map of the clusters. There two layers in the network namely; the input layer and output layer. The input layer consist all clustering fields (input neurons or units) while the output layers contains the output neurons which will form the derived clusters. These two are connected to each other with weights, which are initially set at random and later refined as the model is trained. It uses Euclidean distance to assign records. Figure 6 presents the SOM outline graphically.

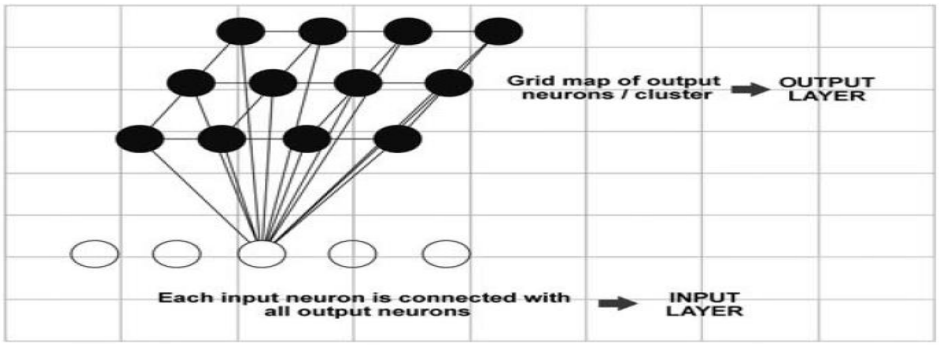


Figure 6 Self-Organizing Map (adopted from Tsiptsis and Chorianopoulo, [9])

Although it provides different view on clustering, it takes longer time to train than the K-means algorithms.

Tsiptsis and Chorianopoulo, [9] suggested that that the results of cluster models could also be applied for proper supervised models such as decision tree to more understandable and simple rules that can reveal significant patterns and differentiate the best charactering attribute in each cluster. Therefore, the results of cluster models can be used to profile the groups/segments e revealed. Because profiling is necessary for understanding and labelling the segments based on the common characteristics of the members. The similarity of decision tree rules and business rules increases its usability and communicability. In another words, DT rules are easier to be communicated and interpreted to specific business objective. Moreover, it can be used to classify new record based up on the discovered clusters. The results of clusters can be represented using classification models

2.6.1.4 Evaluation of Cluster Models

Even though there are various clustering algorithms, cluster models needs to be evaluated in terms of the number and relative size of the clusters, their cohesion, and separation. Besides a good clustering solution comprise highly separated (minimum inter-cluster distance) and compactly cohesive (maximum intra-cluster) clusters [9]. The subsequent topics clarify how clustering results are measured based up on the aforementioned issues.

I. The Number of Clusters and the Size of Each Cluster

Primarily, cluster solutions are examined by viewing the number of the clusters revealed. The numbers of cluster needs to manageable and useful for the specified problem domain. The number of records assigned to each cluster should to be considered. It means if majority of the record are assigned to a particular cluster, it point out that group requires more assessment. Contrarily, the cluster that contains small number of record indicates that the group requires special consideration. But if the analyst observed that the cluster solution as outlier it needs to be set apart for further investigation [9].

II. Cohesion of the clusters

The concept clusters cohesion is to indicate that a good clustering solution is expected to be composed of dense concentrations of records around their centroids. When the values shows high variance it shows that the clusters are non-homogeneous, which suggest further partitioning of

the dataset. Moreover, there are different statistical measures that help to summarize the degree of concentrations and the level of cohesion of the revealed clusters. Standard deviations and pooled standard deviations, and maximum (Euclidian) distance are the most common ones. When we evaluate the standard deviation of the clustering field for each cluster, low values, which point out a small degree of dispersion, are anticipated. On the other hand, the pooled standard deviations measures the weighted (according to each cluster's size) average of the individual standard deviations for all clusters based on each cluster size. Here also, low dispersion and low values are looked-for.

The level concentration of each cluster can be measured by expecting of maximum distance from the cluster centre using the Euclidean distance measure. It denotes the range of each cluster to show how far apart the remotest member of the cluster lies.

Commonly, the cluster cohesion measured using sum of squares error (SSE) measurement standard to indicate the average distance of each members of the cluster and their centroid. This measurement is based on the (squared Euclidean) distances between the data points and their centroid. Thus, it is the sum of squares difference of each data points and the centroid. The cluster models are evaluated by SSE method as follows [9]:

$$SSE = \frac{1}{N}\sum_{i \in c}\sum_{x \in c_i} dist(c_i, x)^2$$

- where c_i is the centroid of cluster i , x a data point or record of cluster i and N the total cases. A solution with smaller SSE is preferred.

Equation 1 Evaluation of SSE method

III. Separation of the clusters

To evaluate the separation of cluster is another concept in the evaluation of cluster model. Here, a proximity matrix is constructed with the distances between the cluster centroids.

SSB is criterion used to measure separation of clusters is based on the (squared Euclidean) distances of each cluster's centroid to the overall centroid of the whole population. The computation of SSB works as follows [9].

$$SSB = \frac{I}{N}\sum_{i \in c} N_i * dist(c_i, c)^2$$

- where C_i is the centroid of cluster i, c the overall centroid, N the total cases, and N_i is the number of cases in cluster i.

Equation 2 Evaluation of SSB method

2.6.2 Customer Segmentation and Decision Tree Classification

Decision tree is widely techniques for customer segmentation. It generates a tree by partitioning the dataset iteratively. In every split of the DT tree model, the best predictive attributes, which can split the datasets in to meaningful, are automatically selected to produce the DT model. The datasets are divided by best predictive attributes into sub-segments that exhibits similar characteristics based the predefined class labels. DT results are easy to interpret and convertible according to firms business objectives to understand their customers and their offerings. That's what makes DT tree models transparent. Besides, the tree models can be converted to IF_THEN rules to by combining the conditions in the model, starting the top of the tree to one of the leaves. Nowadays, a decision tree algorithm provides multiple tasks. Their scalability and speed makes them to be used in customer segmentation [9].

Decision trees are part of the Induction class of DM techniques. An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modeling, the decision is simply the predicted value. The decision tree DM technique enables to [19]:

- Classify observations based on the values of nominal, binary, or ordinal targets
- Predict outcomes for interval targets

- Predict the appropriate decision when you specify decision alternatives

The following figure shows a simple decision tree which determines a cars mileage from its size, transmission size type, weight. The leaf nodes are in square boxes which represents the three classes of mileages. From the decision tree, the conclusion can be made medium size, automatic will have medium mileage.

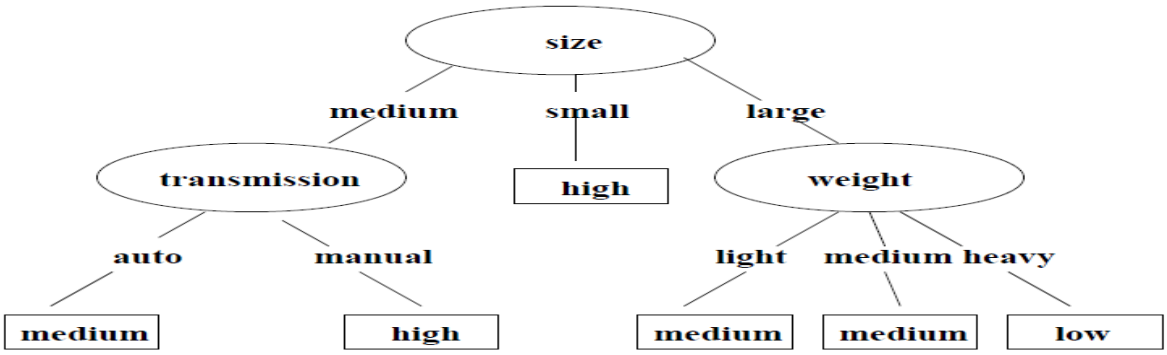


Figure 7 Simple decision tree model (adopted from Fu, [30**])**

Decision trees are used for classification and prediction purposes. There are two ways that they can be applied. Primarily, they can be used as a part of selection criteria. As well as, they also be used to support the use and selection of specific data within the overall structure.. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer. Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output [17].

Decision trees operate by recursively splitting the initial population. For each split they automatically select the most significant predictor, the predictor that yields the best separation with respect to the target field .Through successive partitions ,their goal is to produce ‘‘pure’’sub—segments with homogeneous behavior in terms of the output. They are perhaps the

most popular classification technique .Part of their popularity is because they produce transparent results that are easily interpretable, offering an insight into the event under study [9].

Decision trees are powerful and popular tools for classification. A decision tree is a tree-like structure, which starts from root attributes, and ends with leaf nodes. Generally, a decision tree has several branches consisting of different attributes, the leaf node on each branch representing a class or a kind of class distribution. Decision tree algorithms describe the relationship among attributes, and the relative importance of attributes. The advantages of decision trees are that they represent rules which could easily be understood and interpreted by users, do not require complex data preparation, and perform well for numerical and categorical variables.
Moreover, decision trees are good technique for classification and prediction of customer segmentation revealed in the clusters models because [9]:

- They are transparency and the intuitive form of their results, decision trees are commonly applied for an understanding of the structure of the revealed clusters.
- Additionally, decision trees can also be used as a scoring model for allocating new records to established clusters.
- Decision trees can translate the differentiating characteristics of each cluster into a set of simple and understandable rules which can subsequently be applied for classifying new records in the revealed clusters. It is also a more transparent approach for cluster updating.
- It is based on understandable, model-driven rules, similar to common business rules, which can more easily be examined and communicated.
- Additionally, business users can more easily intervene and, if required, modify these rules and fine tune them according to their business expertise.

2.6.2.1 The Pruning Method: Error Reduction in Decision Tree Model

Scholars suggested that decision tree models, less is more and the simplicity of the generated rules is also a factor to consider besides predictive ability. They advised analyst to eliminate models that over-fit the training datasets. Unless patterns and knowledge's extracted from the DT model may prevent to provide generalizable predictions and tend to reduce its acceptability [9].

In addition, Han and Kamber [14] said that DT tree models can reveal some errors due the noisy datasets. To solve this problem, it involves need to select an appropriate method, which enables to remove useless tuples in the training set and reduce anomalies in reflected in decision tree model.

Pruning method for DT models is panacea to overcome the over fitting in the dataset. Therefore, the inconsistency appeared because of noises and outliers in the training set can be solved through pruning the tree model. The pruning DT can increase the degree of relevance of the model and its applicability It uses statistical techniques to remove irrelevant branches of the tree model. One can choose the pruning of trees before (pre-pruning) or after the model is built. The former works by altering the parameter of the run information to stop splitting the training subset at a given node while the later (post running) eliminate the branches of sub trees from full grown tree. Hence, the branches are replaced and labelled with the leaf of most frequent class sub-tree being replaced. Post pruning is mostly used method than the pre pruning to build DT model. The pruned tree of DT model produces more coherent trees; perhaps they are faster and better at correctly classifying the test data. Figure 6 illustrates unpruned and pruned DT [14].

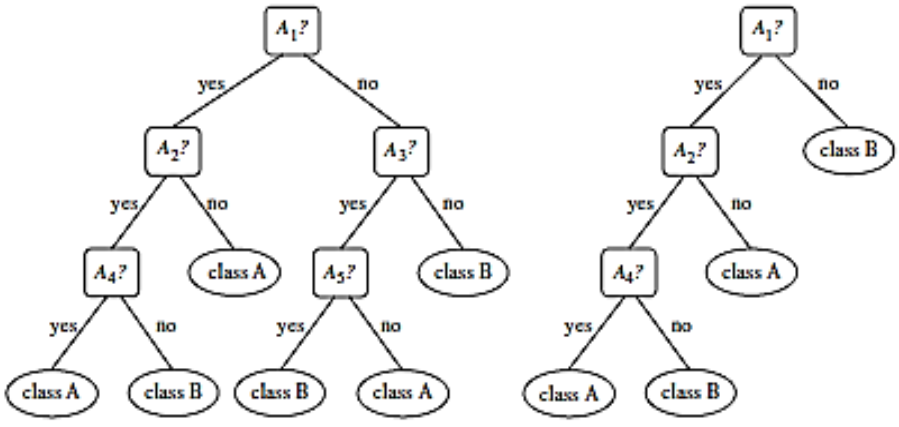


Figure 8: Unpruned and Pruned DT (From Right to Left: adopted from Han and Kamber, [14])

To sum up, DT algorithms have a capacity to reveal the relationship among attributes and instances with respect to their class .in addition they can be integrated with clustering models to classify and predict according their cluster. Their applicability in customer segmentations have vital role. In addition, it is advised to use the tree pruning method of decision tree algorithms to build a fine model.

2.6.2.2 Decision tree algorithms and evaluation methods

Different kinds DT algorithms are available with different tree growth methods. Even though they differ from one another by the measurement criteria they use to choose best split, all of them are aimed at maximizing the total purity by identifying sub-segments dominated by a particular outcome. The most common DT tree algorithms are C&RT, C5.0 and CHAID [9]. The first two are explained below.

2.6.2.2.1 Classification and Regression Trees (C&RT) algorithm

C&RT is decision tree algorithm that produces binary splits by splitting two splits of child nodes. For this algorithm typically, the Gini impurity measure is used. Moreover, the Gini coefficient is used to measure the dispersion that depends on the distribution of the outcome categories. It ranges from 0 to 1 and has a maximum value (worst case) in the case of balanced distributions of the outcome categories and a minimum value (best case) when all records of a node are concentrated in a single category. For a particular split it is the weighted average of the resulting child nodes. At each branch, all predictors are evaluated at each branch. Moreover, for the split, the predictor that results in the maximum impurity reduction or equivalently the greatest purity improvement is selected. The Gini impurity measure is computed as follows [9]:

$$Gini = 1 - \sum_i P(t_i)^2$$

Equation 3 Gini impurity measure

where $P(t_i)$ is the proportion of cases in node t that are in output category i.

2.6.2.2.2 The C5.0 algorithm

The C5.0 is later version of the widely used C4.5 algorithm DT algorithm, which can produce more than two sub-groups at each split by offering non-binary splits. In other words, it measures how well an attribute separates the training data according to the target class [9]. The algorithm generates a classification-decision tree for the given data-set by recursive partitioning of data. The decision is grown using Depth-first strategy. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is

considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes [27]. The possible splits are evaluated based on the information gain that rooted is concepts in information theory known as Entropy. Entropy measure is computed as follows:

$$Entropy(S) = p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

where $p_{(+)}$ is the proportion of positive examples in S and $p_{(-)}$ is the proportion of negative examples

Equation 4 Entropy measure

The information depends on the probabilities (proportions) of the outcome classes and it can be expressed in bits which can be considered as the simple Yes/No questions that are needed to determine the outcome category. It chooses the best predictor of the split, which scores maximum information gain ratio. One quality of using this algorithm is the results are revealed in normalized format along with their ratio. This is in turn solves the bias in the previous version of the algorithm (C4.0) toward large and bushy trees. The information gain works as follows.

$$Gain(S, A) = Entropy(S) - \sum_{ve\ Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where $Values(A)$ is the set of all possible values for an attribute A and S_v is the subset of examples in S which have the value v for attribute A

Equation 5 Information gain measure

However, both algorithms may produce bushy tree, tree pruning is often integrated to produce smaller trees with similar predictive performance.

2.6.3 Customer Relation Management and Customer Life Time Value

2.6.3.1 Customer Relation Management

Customer Relationship Management (CRM) has become a leading business strategy in highly competitive business environment. CRM can be viewed as 'Managerial efforts to manage business interactions with customers by combining business processes and technologies that seek to understand a company's customers. Companies are becoming increasingly aware of the many potential benefits provided by CRM. Some potential benefits of CRM are as follows. First, it enables to increase customer retention and loyalty. Second, CRM enhances higher customer profitability. Third, it contributes to the creation of value for the customer. Forth, it enables customization of products and services. Finally, CRM lower process, higher quality products and services [21].

2.6.3.2 Customer life time value

2.5.2.1 Customer value

According to Two Crows Corporation [11], customers are arguably the most valuable asset of a firm - customers drive profits. Hence, maximizing customer value is one of the key objectives of customer relationship management - from acquiring and retaining profitable customers through targeted marketing to increasing their value over time through cross- and upselling campaigns. In particular when marketing resources are tight, it is often necessary to identify the most "valuable" customers up front so as to allocate these limited resources.

According to Jung, Suh and Hwang [22], customer value are classified into three categories; current value, potential value and loyalty. The value creation process consists of three key elements: the value customer receives; the value organization receives; and, by successfully managing this value exchange, maximizing the lifetime value of desirable customer segments. By considering that how much customer's valuable in firms, to measure their customers value is another issue. How does one measure customer value? Past customer profitability is insufficient when trying to predict the future value: a client who has just subscribed to five new magazines at a promotional rate is not guaranteed to renew all of them once the special rate ends. Furthermore, customers are free to "leave", at the latest when their current contract expires. Customer Lifetime Value (CLV) is a metric that encompasses both the past and the future value of the client while

reflecting the uncertainty associated with the latter. CLV of a given client consists of the profit generated by the customer currently and the present value of all expected future profits associated with this client. Both the revenues derived from the client and the costs associated with maintaining a customer relationship with her are usually incorporated into CLV [13].

2.7 Application of Data Mining Techniques in Insurance Industry

Under this section, the researcher reviewed and presented the role of DM techniques in insurance industry. Some of the applicable areas are stated below,

I. Risk assessment

Fikre [8] said that estimation of more important factors in risk assessment were difficult for underwriters to do it manually On the other hand , data mining technology can improve existing business models by detecting important variables and by revealing between risk factors. Thus data mining techniques can help to mine useful patterns and new knowledge to support their decision making process in product development, marketing, claim distribution analysis, asset-liability management and solvency analysis [8].

According to Guo [19], DM techniques such decision tress and neural networks can help to build predictive models to predict the risk class. Thus, insurers can set more accurate insurances rates that can enable them to modify their pricing scheme and increase their competitive advantage.

II. Claim Provision

The settlement of claims is often subject to delay, so an estimate of the claim severity is often used until the actual value of the settled claim is available. The estimate can depend on the following [19]:

- Severity of the claim.
- Likely amount of time before settlement.
- Effects of financial variables such as inflation and interest rates.
- Effects of changing social mores.

The estimate of the claims provision generated from a predictive model is based on the assumption that the future will be much like the past. If the model is not updated, then over time, the assumption becomes that the future will be much like the distant past. However, as more data

become available, the predictive DM model can be updated, and the assumption becomes that the future will be much like the recent past. Data mining technology enables insurance analysts to compare old and new models and to assess them based on their performance.

III. Reinsurance:

Based claim historical data of insured, DM techniques can help insurers to develop predictive models to identify policy holders and policies can be reinsured [5].

In another words, group of paid claims can be used to develop predictive model of the expected claims experience of another group of policies. These predictive models can be identified suitable policies for reinsurance based on the loss experience of similar policies in past [19].

IV. Customer Level Analysis

According to Guo [19], to create a successful customer retention strategy, one should analyse the customer's level revealed in the data rather than evaluating them as whole. For this reason, DM techniques such as association rules can enhance insurances firms to select which policies and services to be offered to which customers. In addition, applying the association rule DM technique for customer level analysis can help insurances firms:

1. To Segment the customer database to create customer profiles
2. To analyze customer segments for multiple products using group processing and multiple target variables
3. To analyze claim and rate a single customer segment for a single product
4. To discover sequential patterns and analyses customer segments given market basket

V. Developing New Product Lines

DM techniques can advance Insurance firms to identify customers segments which contribute high profit. Accordingly, companies can prioritize marketing campaigns with respect to the policy offerings; hence, can utilize all of their available information to better develop new products [5].

VI. Fraud detections

Fraud detection is another applicable area in data mining. According to Umamaheswari and Janakiraman [5], DM technologies can support Insurance firms to discover fraudulent claims and related factors that lead to fraud waste and abuse. Therefore, they can identify which transactions are most likely to be fraudulent.

VII. Customer Acquisition

Insurance firms can acquire new customers to increase their profitability by reducing cost of acquisition in order to expand their market share. Conventionally, this requires high effort of insurance brokers and sales person of the insurances. But modern DM technologies can be strength the effort of acquiring new customers. The DM techniques such as clustering can be used to for this purpose to build segmentation models using the existing customers' data. Therefore, cluster models can augment to identify the segments among already insured customers through which uninsured customers could be targeted. This technique helps to partition the customer database into groups that might or might not exhibit similar characteristics. Most of the time, customers with similar characteristics are clustered into groups. Although, the numbers of clusters can be predefined by algorithms such as k- means, but the clusters are not predefined. Thus, the decipherment of clusters model can't be accomplished without experts' knowledge [23].

VIII. Customer Retention

Customer retention is one the main objective CRM which assists marketing department of insurances firms to get sustainable profits through development of strong relationship with their customers. DM techniques can support to build customer segmentation models that help to identify valuable customers who are having high tended to leave by allowing with increased likelihood to leave, allowing time for targeted retention campaigns. Thus, the most valuable customer groups of customers are identified; to increase the customer satisfaction and firms' profitability. Moreover DM models can support customer development by matching products with customers and better targeting of product promotion campaigns. It can also help to reveal distinct customer segments, facilitating the development of customized new products and

product offerings which better address the specific preferences and priorities of the customers [9].

2.8 Related Works

For the past decade the research in the area became more popular. Among them some of them which are related to customer relationship management are related below.

One of the local researches was conducted by Woubishet [2] on the application of data mining techniques in order to support customer relationship management at Ethiopian airlines. He selected clustering techniques to segment customer information stored on Frequent Flyer Programs (FFP's) database. Ethiopian Airlines has FFP called Sheba miles which is aimed at rewarding loyal customers. This means that, customers, are rewarded if they travel above the usual base miles. The reward is additional miles to travel in to another time.

The methodology selected by the researcher was CRISP- DM. K-means algorithm was selected to segment customers data into five clusters depending on the similarity of the behaviour they exhibit .And also, decision tree classification is used to allocate assign the new customers data to the segments created. The tool selected for model building is Knowledge Studio. The result from the cluster model, among five of five clusters, three of the clusters contained 21% of customer records that generated the highest revenue, and differed in the total frequency of trips and tenure of customers. The cluster containing medium and low revenue generating customers contained 27% and 52% of the customer records respectively. He concluded that applying data mining techniques could definitely support CRM activities at Ethiopian Airlines. Then he recommended that, further study is needed to do on customer segmentation using demographic information and utilizing other clustering algorithms and association rule algorithms could give better results [2].

Hintsay [7] conducted a study to build a predictive model using data mining techniques to support risk assessment in insurance industry. He undertakes the study on Nyala Insurance Share Company (NISCO). Specifically the aim of the research was to apply the data mining technology to address the risk in underwriting activities.

Among the various issues or subjects considered in underwriting management, the selected class for the research was the motor class. The motor policy renewal problems are also mentioned and suggested.

According to Hintsay [7], classifying customers based the risk they involve in the specified insurance company was one the major problems noticed over the year. So the researcher built a classification models using decision tree and neural network. The predefined classes of risks in the study were: low medium and high risk classes. The tools selected in the study were See5 and the BrainMaker Software. Those tools can support decision tree classification and neural network respectively. He pointed see5 software is best for selecting attributes for decision tree classification and the BrainMaker software was selected to train and to build neural network model software .A decision tree classifier, is mainly used in this study for attribute selection that could be used as an input for the neural network. He said that both tools have the facility to partition the dataset randomly into training and testing sets.

From the four branches of NISCO, 1332 records were collected. The total variables of the records are 25. Because of the missing values four attributes were totally discarded and other six attributes were extracted from the existing ones. For the study the data sets selected were 1160. By considering the classes for low risk 629, for medium risk 305 and for high risk 226 were resulted. This record classification where made based on the assessment report made on strong and weak points of the policy's. But the data mining techniques can improve the result by models such as decision tree and neural networks .Then the 1160 dataset were "divided into two: 90% (1044) for model building and testing set and the remaining 10% (116) for validation set". He said out of 1044 records which were selected for model building and testing set " 940 (90%) of the 1044 facts were used for training and the remaining 104 (10%) were used for testing purpose". Finally the results for decision tree classification , according to the three predefined classes(low, medium and high risk policies) the classification accuracy was 98.15%, 94.12%, and 92.86% respectively and the validation test was correctly classified 95.69% of the validation. Next, the neural network model correctly classified 92.24 % of the validation set. He said that high-risk groups are correctly classified and the remaining, low and medium-risk groups, the classification accuracy were of 98.15% and 76.47% respectively [7].

Another study which was entitled as “predictive data mining technique in the case of Ethiopian insurance of corporation” was conducted by Fikre [8]. According to the researcher identified the problem in the corporation is related to life insurance .Specifically, Fikre [8] focused on the policy renewal of personal accident policies. He said that, the absence powerful tool like data mining tool formed a problem to decide which customer could get insurance coverage and which is not by considering their prior health problem. Thus they can accept or reject the contract with such customers. He selected classification techniques in order to classifying the risk categories of the accident whether it is big or small risk claim lasses. The tool he used for the study was KNOWLEDGE studio. He selected he tool because of its availably and its functionality on the data mining techniques such as decision tree, clustering and neural network. And these techniques were selected for the study. The model was constructed using knowledge SEEKER algorithm and data partitioning.

As a result, the model accuracy of 96.61%.and the classification error for small risk claim is 29.5% and the for big claim risk it is 27%. Finally he concluded the frequency occurrence of the values of an attribute has an impact on the accuracy of the model. In addition to he stated that this data mining technology can gain a visible advantage to insurance underwriters by informing them which customers can cause big claim or small claim risk to identify and to measure client’s profitability.

The researcher recommended that insurance company can be beneficiary of data mining technology if they integrate with their customer data bases in to build predictive models such as his work. Because, it helps insurance firms in order to make better decisions. Also he suggested that balanced data portioning is more preferable than sampling approach .Finally as the study used knowledge SEEKER algorithm for modelling. He pin pointed to apply other algorithms such as heat SEEKER, which also found in KNOWLEDGE studio tool, can also give better result [8].

Wodajo [24] conducted a research by applying data mining for combating corruption using corrupt activity data in federal ethics and anticorruption commission of Ethiopia. The researcher target is to reveal hidden useful knowledge from the existing database of the FEACCE through

assistance of data mining tools and techniques. According to the researcher the organization doesn't have mechanism to identify corruption activity of offenders with their personal characteristics that are vulnerable to commit corruption. For the study she selected CRISP –DM process. Data mining techniques such as classification, association rules and clustering were used for modelling purposes. During the experiment, for Association rule, a priori algorithm was selected while k means was used to create clusters. And to build classification model, decision tree J48 algorithm was used in the study .From the data base of FEACCE, 22 attributes and 1000 records were extracted. Then after the remaining 3189 collected from manually. 11 attributes were selected. Attributes which a have ordinal values has been transformed to nominal forms. WEKA was the selected tool for modelling. The result from the study showed, if clustering models are an input for classification purpose, a good classifier models can be generated.
According to the researcher, data mining techniques can extract hidden knowledge in the organization data base in order to take preventive action on the corruption activities. Finally she recommended using data mining techniques such as time series artificial neural network and document summarization. And also suggests instead of using categorical data, continuous data can improve the result of the model [24].

Another research entitled as "Application of Data Mining for Customer Segmentation: The Case of Buusaa Gonofa Microfinance Institution" was conducted by Reganie [25]. The researcher said due to lack of appropriate tool to segment customers before looking potential customers, the intuition couldn't identify value customers and measure their profits. The goal of the study is to build a model that can help to classify customers for of Buusaa Gonofa microfinance institution. The methodology selected by the researcher was CRISP- DM. The researcher used clustering techniques to segment customers in to appropriate number of clusters. K means clustering algorithm was used to cluster customers which exhibit similar characteristics. Then after, he built a predictive model to predict potential customers. For that reason or classification purpose, J48 algorithm was used.

Finally, the predictive model achieved 99.5% accuracy. The researcher concluded that, the result of predictive modelling is encouraging. Perhaps, data mining technology can be applied in micro finance industries in order to extract interesting patterns and knowledge. He recommended the institution should consider developing an integrated warehouse to apply data mining techniques.

He also suggested further study may also require using other data mining technique and algorithms for a better performance of CRM in the institution [25].

The aforementioned related works helped the researcher to grasp on different aspects. The researcher obtains a good insight regarding on the application of data mining techniques in CRM. The reviewed works helped the researcher to gain more understandings and to look at multiple views regarding on the selection and application of DM techniques, tools, and algorithms used in each of the study. In addition, the researcher appreciates that how DM techniques can be utilized to develop descriptive and predictive models that can be used to segment customers. Besides, the review of these studies helped the researcher to realize how to conduct a research in data mining area. Moreover, it supported the researcher to have know-how regarding the role of data mining in insurance domain and how to use the technology for marketing function such CRM in order to segment customers for particular purpose. In general, the researcher is encouraged by reviewing related works to conduct this stud

CHAPTER THREE

METHODOLOGY

In order to meet the goal of the study appropriate data mining methodology has to be selected. Therefore, the selection criteria were based on the nature of the problem identified and on the extensive effort required to review related literatures from previous studies in the study area.

3.1 Research Purpose

The goal of the study was to apply data mining technique for customer segmentation on customers' records in the LIFE INSIS database of EIC life insurance at LAD. Based on the information gained regarding the business practice and the service given by the Corporation to its customers and information obtained from the customers' database, the researcher designed methodology to assess the problems and to find solutions for the problems in the aforementioned insurance corporation. The aim of the study was to build a meaningful model that segments life insurance customers of EIC according to their value through exploring and applying appropriate data mining techniques. The result of this study could help EIC marketing managers to make better decision and to provide a better service to their customers; hence it could also help to assess the profitability of their existing customers.

3.2 Research Design

This study is an experimental research in which quantitative and qualitative approaches are implemented for data collection and data analysis. The research design was selected based on the idea suggested by Singh [31] that intended to evaluate something new in order to contribute some knowledge on already existing one; it enables as to improve the conditions under which we observe to arrive at more precise results. The design and the approaches are selected because it is appropriate to examine the application data mining techniques to build customer segmentation model based on customer value. For this purpose, 12 months (from August, 2011 to august, 2012) the historical data pertaining to the existing policy holders of the EIC life insurance has been collected from LIFE INSIS database, which located at LIFE Addis District (LAD) LAD. The data base was chosen as it justifies the historical data related to policy holders customer demographic, transaction policy and personal information, which helps to build the mentioned

customer segmentation models. Due to limitation of accessing the historical data of life insurance customers, the dataset of the study was constructed from LIFE INSIS data base located purposively. The selections of attributes are based on the consultation domain experiments.

3.3 Research framework

For this study, the CRISPDM (Cross-Industry Standard Process for Data Mining) was adopted as research framework. This methodology consist of 6 (six) steps with some feedback loop. It is a cyclic and combined methodology that especially is designed to give a general framework for data mining analysis. CRISP-DM is preferred in data mining researches because it is non-propriety, freely available and application-neutral standard for data mining models, and it is widely used by researchers in the field for more than a decade [10]. The other reason for the selection of CRISP-DM in this research is because it allows the aspects of industrial and business models.

The CRISP-DM methodology is described in terms of a hierarchical process model (see figure 9), comprising four levels of abstraction (from general to specific): phases, generic tasks, specialized tasks, and process instances [10]:
At the top level, the data mining process is organized into a small number of phases. Each phase consists of several second-level generic tasks.

The second level is called generic, because it is intended to be general enough to cover all possible data mining situations. The generic tasks are designed to be as complete and stable as possible. Complete means to cover both the whole process of data mining and all possible data mining applications. Stable means that we want the model be valid for yet unforeseen developments like new modeling techniques.

The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in specific situations. For example, at the second level there is a generic task called build model. At the third level, we might have a task called build response model which contains activities specific to the problem and to the data mining tool chosen.

The description of phases and tasks as discrete steps performed in a specific order represents an idealized sequence of events. In practice, many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions. The CRISP-DM process model does not attempt to capture all of these possible routes through the data mining process because this would require an overly complex process model and the expected benefits would be very low.

The fourth level, the process instance level, is a record of actions, decisions, and results of an actual data mining engagement. A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

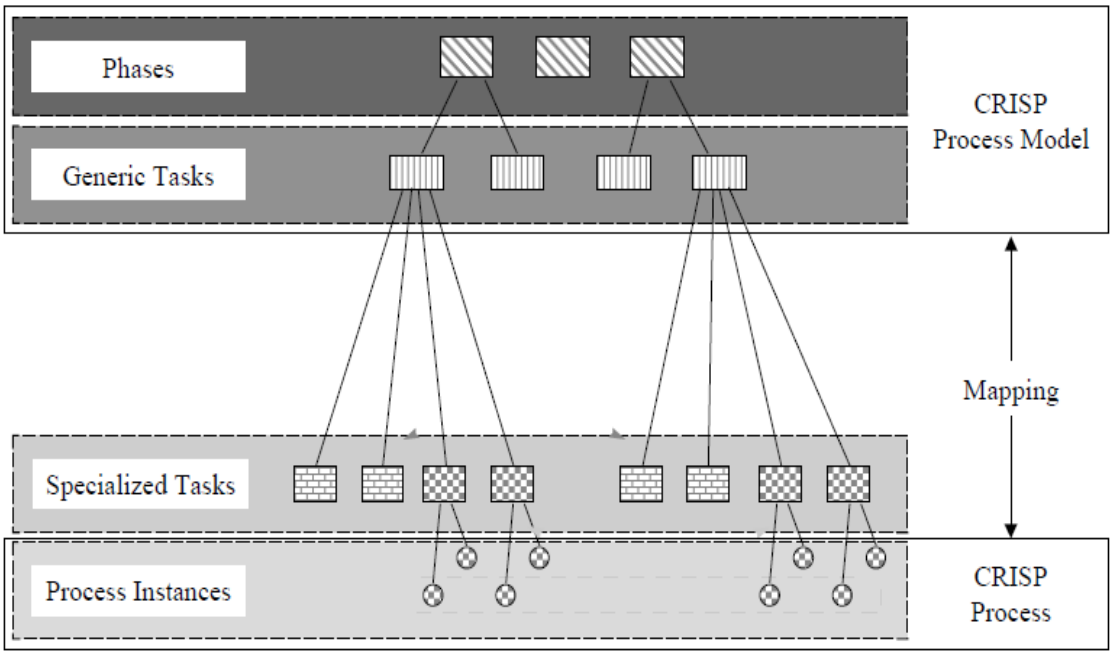


Figure 9 CRISP-DM- phases, generic tasks, specialized tasks, and process instances (adopted from adopted from Chapman, Clinton, Kerber, Khabaza, Reinart, Shearer, and Wirth, [10])

Figure 9 depicts that CRISP-DM breaks down the life cycle of a data mining project into six phases. Those are business understanding, data understanding, data-preparation, modeling, evaluation and deployment.

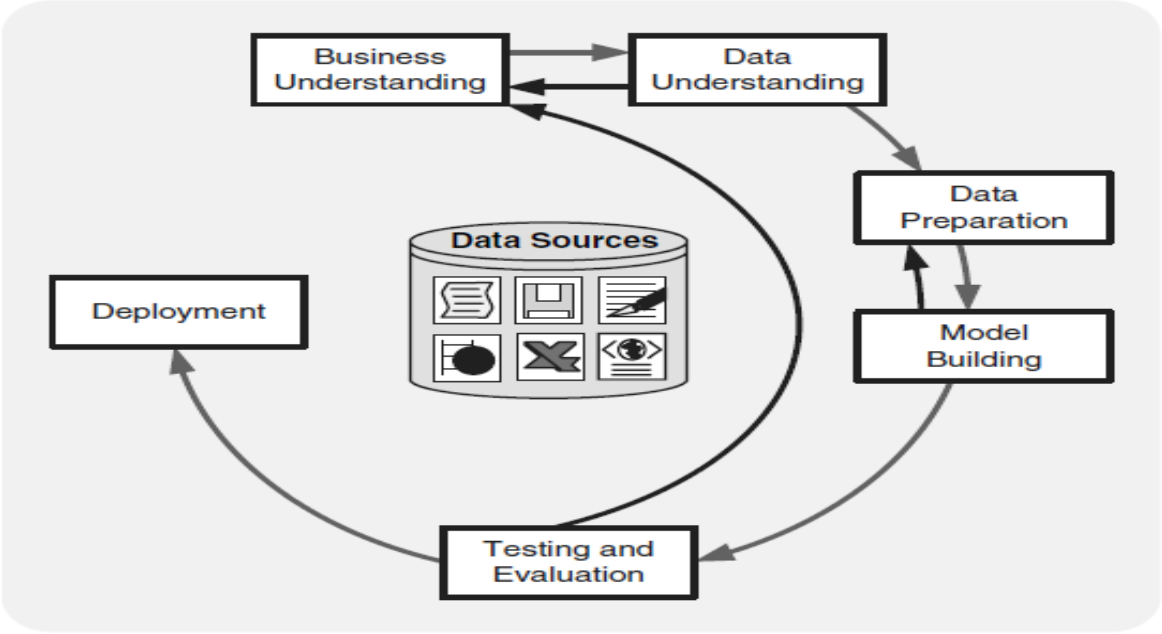


Figure 10 Crisp-DM process (adopted from Olson and Delen, [17])

As it has been indicated in **Figure 10**, the six steps of CRISP-DM processes are cyclical that can be implemented iteratively. And, the subsequent explanations are the descriptions of the methodology within the study context.

3.4 Business understanding phase

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives [10].

During the Business Understanding Phase a literature review is performed in order to gain insight with data mining studies which related are to customer segmentation, customer life time values and customer relationship management studies which have been solved by the application of data mining techniques and methods in previous research projects. The study domain area is insurance organization.

The researcher made an effort to study the objective of Ethiopian Insurance Corporation and the situation of the business it is involved through document analysis and interviews.

Interviews was undertaken to understand the business .Both formal and informal interviews were conducted with the manager of marketing, the manager of life insurance at LAD, Deputy Manager of ICTM department and specific employees who are engaged in operational tasks on the existing system of the corporation. This was done in order to get insight with the specific problems at Ethiopian Insurance Corporation which has not yet been solved to segment their customers' value and the measurement taken to segment their value. Currently service sector business organizations are more customers centric. So that it is are very important for the performance and for more effective and efficient to get insight with customer data. Some information was gathered informally with employees and administrative staff (IT experts and managers). Based on the outcomes of the performed research, the project goal and objectives, and the main research questions are formulated.

At EIC, they to store customer and transactional data in excel file and paper forms. In 2011 the corporation implemented the system called INSIS. The system designed is divided into two specific tasks. Those are called Life INSIS and General INSIS. The system handles operational and transactional data, policy data, accounting data, claim processing and agent commission processing. The system stores the data of clients and transactions in oracle database. The qualified employees can make enquiries and changes to the data stored. Per the objective of this research, the Life INSIS database, which handles the customer transactional, policy, personal and demographic information is nominated. Based up on these, there is a need to build descriptive and predictive model to segment life insurance customers of EIC based on their value to support EIC marketing managers to identify valuable policy holders and make better decision in CRM. Moreover, the study conducted on main branch of Ethiopian insurance company because the newly opened branches are limited to have historical data required for the study.

3.5 Data understanding phase

During this phase the investigation starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

To achieve the objective of this study, the researcher attempted to understand the data reside in the specified data base (LIFE INSIS) that can help to segment the customers of EIC based on

their value contribution. Considering the problem described and the business task, data pertaining to policyholders 12 months (from August, 2011 to August, 2012) the historical data is extracted from LIFE INSIS data base of EIC life insurance into MS-excel. To achieve the data mining goal formed from the business problem, relevant of the information were selected from four tables. The names of these tables are customer data or **p. people** table, **insured type** (**insurance value**) table, **Cover type** table (risks to be covered) and **occupation table.** With consultation of domain experts, data gathered are grouped into two major classes namely; personal and demographic and life insurance policy and transactional attributes to categorize the data into corresponding shared characteristics. MS-Excel is used for understanding the nature of the data such as the relationship of attributes and their records from the selected tables. The redundancy of data is revealed among the attributes of the four tables. To avoid this, the selected tables need to be merged. Then, DataPrepartor-1.7 is used to visualize the names of attributes and their data types since it's provides summarizing preview of the data in table format with user-friendly graphical interface. The collected datasets are composed of the nominal, numeric and date data types. Yet, the data that explain the policy holders' premium payments, which supports to compute the customer value, are missing due confidentiality and sensitivity information. However, the discussion with the domain experts confirmed that sum assured, duration, and the agent commission data are adequate to compute the existing policy holders' value using the MS-excel. Finally, from the initial 16 attributes and two derived attributes along with 27845 are selected for data preprocessing task.

3.6 Data preparation phase

Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data cleaning, transformation and preparation were conducted in this phase. Data exploration at a greater depth was applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding [17].

The data preparation phase covers all activities needed to construct the final dataset. Thus, the final dataset can be fed into the modeling tool. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

In this study, data preprocessing involved tasks such as data cleaning, data transformation, data reduction and attribute selection.

DataPrepartor-1.7 is used for preprocessing task because it provides variety of statistical views for data visualization and it has different methods to handle missing values and outliers. For instance, using the Z -Score Method, one can handle outliers in numeric attributes. Attributes that registered high missing values removed from the dataset. Modal value (most frequent value) is used to replace the missing values of nominal attributes while mean value (average value attributes) is used to replace the missing values of numeric attributes that have the same class label (on the labelled datasets). These methods are chosen since they are the most common measure of variables central tendency that in turn can be used for imputation of Missing Values.

WEKA data mining tool is selected for preprocessing task such as discretization, normalization and attributes selection. With WEKA, numeric attributes are discretized in order to replace the labels attribute by intervals, which is easy to interpret and consistent to apply different DM techniques. In addition, numeric attributes are also normalized to prevent bias when attributes have very different ranges. Besides, the dimensionality of the data was evaluated using information gain evaluation method of WEKA. MS-Excel is also used for data preparation; pre-processing and analysis by using it functions such as sort, formulas (to compute CLV), filter, find and replace and so on. Besides, it is used for documentation purpose. The data sets in MS-Excel should be converted CSV (Comma Separated Value) and ARFF (Attribute-Relation File Format for processing). The input of the data into data mining applications proved to be simple with the conversion of an Excel spread sheet datasets into a CSV file format and then an ARFF file format for modeling purposes.

Finally, with help domain experts' most important attributes, which helps to segment policy holders of EIC life insurance based on their value are selected final datasets which were prepared for modeling purpose.

3.7 Modeling

In this phase, DM modeling techniques that can solve the problem identified are selected and applied, and their parameters are calibrated to optimal values. For this reason, combinations of clustering and classification models are used in this study for building the customer segmentation model.

The clustering technique is selected as it is capable to find groups that have similarity in the data. In addition¸ it is a technique that has been applied to explore homogeneity of records of customers with respect to the clustering fields and assign them to the revealed clusters. Furthermore, using clustering technique can enable to discover groups with internal homogeneity and interclass heterogeneity. It is DM technique has been commonly used for market researches to develop customer segmentation based attributes on their life time value (LTV), socio-demographic and life-stage information, and their behavioural, need/attitudinal, and loyalty characteristics [9]. Thus, clustering technique is applied this study in order to develop customer segmentation model based on the customer value to discover homogenous characteristics that policy holders exhibit each group.

The other DM technique selected for this study is decision tree classification technique. DT is classification is supervised learning method used for classification and prediction purposes. For this study, DT classification is selected based the on the recommendation of Tsiptsis and Chorianopoulo, [9], which said that the results of cluster models could also be applied for proper supervised models such as decision tree to more understandable and simple rules that can reveal significant patterns and the most differentiating attribute in each cluster. Considering the goal of the study, DT models supports to understand the revealed clusters and to predict the factors influence customer value. Besides, DT can be applied to profile the segments revealed and to classify new record based up on the discovered clusters. Also usability and communicability makes a DT to be used for this study.

In this study, in order to build DM models that can solve the identified problem, the WEKA (Waikato Environment for Knowledge Analysis.) was selected. WEKA is open source software which can implement most of the technical aspects of the CRISP-DM standard data mining

methodology [26]. Besides, it was selected because of its accessibility and familiarity of the researcher with the tool.

Weka was selected because [20] of the following reasons:

- It provides many different algorithms for data mining and machine learning;
- It is open source and freely available;
- It is platform-independent;
- It is easily usable by people who are not data mining specialists;
- It provides flexible facilities for scripting experiments;
- It is kept up-to-date with new algorithms being added as they appear in the literature review.

WEKA software has implementations of numerous classification, clustering and prediction algorithms. WEKA is a tool for data analysis and includes implementations of data pre-processing, classification, regression, clustering, association rules, and visualization by different algorithms.

In this study, the unsupervised learning, k –means clustering algorithm is applied in order to group or segment customers with similar or homogenous characteristics. This algorithm is selected because it was identified as the most efficient and the fastest clustering algorithm that can handle both long (many records) and wide dataset. Unlike the SOM, it does not take longer time to train the cluster. Besides, users can predetermine the number of clusters to be formed is in advance. K-means clustering algorithm was applied to reveal underlying segment (into k numbers of clusters) and to analyse similar characteristics that policy holders exhibit. Then, the developed models k- means clustering models were used to identify groups of life insurance policy holders who contribute high or low value to the corporation.

In this study, decision tree classification models are developed in order to understand the segments and profile customers according to their value, which is revealed in the cluster model. . The DT tree J48 algorithm is used to classify and predict customers of life insurance based on their value. J48 is based on the C4.5 decision tree algorithm in Weka data mining tool, which builds decision trees from a set of training data using the concept of information entropy. Furthermore, the J48 is capable to select best predictive attributes that can split the datasets according to the target class by measuring the information gain of the attributes. This in turn is

very important to the study in order to select most important that can segment life insurance policy holders of EIC based on their value. The output of the chosen clustering model is used as an input for the decision tree classification model. Because, the similar characteristics of customers revealed in each cluster can be used to classify and predict customers according to their value. DT models supports to realize and to label the segments based on the common characteristics of the members

3.8 Evaluation

We need to thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached. Thus the models outcome should be evaluated from business objectives where it was formulated and stated and in the first phase. This in turn requires investigating the data mining goals too. Gaining business understanding is a cyclic procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations [10].

In this study the researcher examined different models and algorithms which gives or yields a better result. The results of the models developed using classification and clustering were evaluated.

The descriptive models built by k-means algorithm are evaluated by measurements such as Sum of Squared Error (SSE), time taken to build the model and number of iterations. However more weight is given to SSE as it was identified as the most common measurement to evaluate k-means clusters. As result, the cluster model which shows minimum SSE, less time to build the model and minimum number of iterations is selected to build the predictive model. The DT models developed by J48 algorithm are evaluated using average accuracy rate revealed in the correctly classified instances and the confusion matrix.

3.9 Deployment

Data mining can be used to both verify previously held hypotheses, or for knowledge discovery (identification of unexpected and useful relationships). Through the knowledge discovered in the earlier phases of the CRISP-DM process, sound models can be obtained that may then be applied to business operations for many purposes including prediction or classification and clustering depending on the of key situations of the corporation. . These models need to be monitored for changes in operating conditions, because what might be true today may not be true a year from now. If significant changes do occur, the model should be redone. It's also wise to record the results of data mining study. So the documented evidence is available for future studies.

Creation of a model is generally not the end of the study. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained needs to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise [10]. In this study, the model built is not deployed. So that the outcome of model selected and built should be arranged and to be used as users and clients who operates and get benefit the system in EIC.

CHAPTER FOUR

BUSINESS UNDERSTANDING, DATA UNDERSTANDING AND DATA PREPROCESSING

In order to reach the goal of the study, first, understanding business and then gathering the data and preparing the data are essential steps. Thus, the purpose of this chapter is to understand the business, to identify the sources of data, to clarify data mining goal and to prepare data set for experimentation.

4.1 Business understanding

Business understanding is the initial stage of the DM process. It is where the study objectives and requirements are described from business viewpoint in order transforming it to DM problem definition. Thus, the researcher spent some time to understand and to get familiar with the life insurance concepts and terms , forms , policies, work processes and insured (customers life insurance) information stored at LIFE INSIS database. This required reading and reviewing different literatures on life insurance and customer relationship management. Observing the work process and customers' database (LIFE INSIS) of the corporation at LIFE ADDIS district was another task. Besides that, to realize the mission, vision and objective of Ethiopian Insurance Corporation at Life Addis district, the researcher collected information from magazines, bulletins and internal records that are documented in EIC office and website (see 4.1.1). Moreover, interviews with the department of ICTM and Marketing supported the researcher to acquire inordinate information to the know-how of the study area and the business

4.1.1 Life insurance in Ethiopian Insurance Corporation (EIC)

4.1.1.1 PURPOSE OF INSURANCE

Information gained from EIC describes that "the fundamental purpose of insurance, whether of people or of property, is protection against possible economic loss, economic loss being simply defined as the unintentional and permanent loss of something which has monetary value [32]."

Common examples of economic loss are: - Theft of one's household goods, damage or destruction of a car in an accident, total or partial destruction of a factory by fire arises from a neglected cigarette, an airplane crash, etc.

In each of the above examples, the loss has two characteristic features. Firstly, it is unintentional and unexpected. Secondly, it can be measured in terms of money. Therefore each of the above losses can be a proper subject of insurance. There two main types of insurances offered by EIC. These are life and non-life insurance. Due to the aforementioned objective of the study and the availability of historical data regarding customers 'demographic, policy and transactional information inside the corporation customers database, the researcher selected life insurance area for the study. The target of following descriptions was to discuss the basic terminologies life insurance, the insurance policies (risk covered) offered by the corporation and related activities held at EIC life insurance.

4.1.1.2 INSURANCE OF PERSONS

Unlike the value of property, the value of a human being cannot be measured in terms of money. But as far as insurance is concerned, the economic value of a person is basically represented by his income. The total or partial, temporary or permanent loss of this income represents an economic loss to all those who are dependent on that income for their livelihood. The loss of such income can come about through a variety of causes; but not all such causes are necessarily insurable since as indicated above an insurable risk must (among other things) involve a loss that is unexpected and, so far as the individual insured is concerned, Unpredictable. Therefore, Life Insurance (in its widest sense) is concerned only with those economic losses caused by death, disability and old age, all of which substantially meet the qualification of an insurable risk. The following description will have detail information [32].

4.1.1.3 DEATH

The continued life of an income earner may have economic significance to many people. Certainly it is most important to any dependents he may have. These dependents have to bear the expenses of last illness and burial. Their greatest economic loss, however, is the loss of the future earnings on which their livelihood depends (apart from the emotional trauma suffered as the result of the death of a loved one).

The continued life of an income earner is also important to his employer and to his creditors, if any. If the individual is in business for himself or as a partner, his life may be of economic significance to his employees, to his partners and to other business associates as well.

Death is, therefore, always a cause of economic loss. But life insurance offers a way of reducing the impact of such losses no matter how large they may be or when death may occur.

Although as stated above the fundamental purpose of life insurance is protection against economic loss arising from death, it can also be used as means of saving as it will be explained later. .

4.1.1.4 DISABILITY

The economic loss suffered as the result of disabling accident or sickness is of two main types. Firstly, income is lost during the period of prolonged disability and, secondly, expenses are incurred for the necessary medical care and treatment (and at a time when such expense can least afforded). Such losses are covered by accident and health insurance contracts which help reduce the impact of the losses by replacing lost income in whole but more usually in part and/or reimbursing a substantial portion of medical expenses necessarily incurred.

4.1.1.5 OLD AGE

The financial problems of old age and the possibility that even a relatively well-to-do person may outlive his financial resources constitute the third area of possible economic loss with which life insurance is concerned. Protection against this possibility is provided by the retirement annuity (or pension) policy, which pays an income at regular intervals to the insured annuitant for as long as he lives.

Generally, life insurance covers only economic losses caused by death, disability and old age based on the qualification that insurer stated as insurable risk.

4.1.1.3 BASIC FORMS OF LIFE INSURANCE

In EIC Life insurance has developed along three different classes of insurance; i.e. ordinary, industrial, and group insurance. Based on those three classes of Life insurance, under each of them, there are also three main forms of or plans. These plans are term, endowment and whole life. Life insurance is not a contract of indemnity but a valued contract. The various life

insurance plans are developed to meet the different needs and circumstances of persons and in general they can be used appropriately as follows:

4.1.3.1.1 TERM

In spite of the long-term nature of many life insurance needs, there are many needs that are continued to somewhat limited periods of one's life. This is true for example, of certain family needs. In almost every family the period of greatest financial need coincides with the years in which children are growing up and being educated. Term life insurance on the life of the father can be used to provide protection during these years Term polices provide a great deal more life insurance cover for a given premium outlay than either whole life or endowment polices.

Another most common temporary need for life insurance arises out of personal indebtedness. When a person borrows a sum of money, he usually expects to repay it out of his earnings. His death before the loan is paid thus constitutes a clear possibility of economic loss to the lender (creditor). Often there is no estate from which the creditor can collect the unpaid debt; and even when collection is possible some delay is unavoidable. The creditor therefore, has a definite need for insurance on the life of his debtor for the period of the loan.

The debtor, too, has a need for insurance in this instance. If he should die before repaying the loan in full, his family's share of any property he leaves will be reduced by the amount of the indebtedness remaining at his death. For this reason, many borrowers insure their life for an amount sufficient to repay their loan in the event they die before payment has been completed. Term life insurance is clearly appropriate for either creditor or debtor.

Term life insurance is especially appropriate in connection with a mortgage loan for the acquisition of a home. The policy owner borrows a relatively large sum of money, usually to be repaid in monthly installments over a period. His death during this period could easily mean the loss of the family home. Term life insurance, often called mortgage protection insurance, is available for protection in this situation. The policy is issued to cover the identical period of the mortgage loan and for an amount that decreases from year to year at the same rate as the unpaid balance of the loan, in other word; the policy is a decreasing term life insurance policy.

Decreasing term life insurance policies or riders known as "family income benefit" contracts are also used to provide an insured's family with monthly income from the date of death of the insured to the end of an agreed period. Convertible term life insurance plans are especially appropriate during their period of low income faced by many young people who are just getting started in a career. Their economic prospects are favorable. Eventually, they will no doubt reach significantly higher income brackets, and whole life insurance will be well within their means. However, their present incomes are quite limited.

Finally, the continuing inflation of the past several years has prompted many life insurance companies to offer term life insurance on an increasing basis. This increasing term converge is usually provided in a special provision or a rider that may be attached to a basic life insurance contract. The effect is that during the term of the rider, the proceeds payable upon the death of the insured are increased by stated amounts each year. Such rate of increase is often designed to coincide roughly with an estimated increase in the cost of living.

4.1.3.1.2 ENDOWMENT

Endowment life insurance policies are appropriate for any situation in which a fund needs to be accumulated by the end of a specified period. Such a fund could be used to purchase or supplement retirement pension, to finance children's university education, to start a small business and a host of other purposes .In the meantime, the policy afford cover for dependents.
Thus, in endowment insurance the primary need is that of saving, insurance protection being only incidental. Before maturity, such saving can be drawn out as a loan to meet financial emergencies, or the policy used as a security for a loan from commercial bank, etc…

4.1.3.1.3 WHOLE LIFE

The basic life insurance need is to provide funds for the support of dependents after one's death. The person in modest circumstances relies heavily on whole life insurance to meet those needs, since it is the only plan that will meet them on a guaranteed basis; regardless of the date of the insured's death. Limited payment whole life policies are generally appropriate in a situation

where the insured wishes to restrict the premium payment period to the years prior to his retirement age so as to eliminate the payment of premiums from reduced income.

4.1.1.4 SUPPLEMENTARY CONTRACTS

Due to accident or sickness persons are exposed for disability In EIC there are two particular type of supplementary insurance contract, which are designed to cover such risks. The two supplementary contracts (or riders) are the Supplementary Accident Insurance Contract (SAI), and the disability wavier of premium (WP) contract. They are called "supplementary contracts" or "riders" because they can be issued only in conjunction with a life insurance policy and not separately alone. A specified additional premium is payable for each of them.

4.1.1.4.1 SUPPLEMENTARY ACCIDENT INSURANCE (SAI)

The rider covers death or dismemberment or loss of time by accidental means as defined therein. The risk covered is defined in the contract said that if the insured, before reaching the age of sixty years and while this contract is in force, shall sustain bodily injury effect directly and independently of all other causes through external, violent and accidental means of which, except in the case of drowning or of intentional injury revealed by an autopsy, there is evidence of visible contusion or wound on the exterior of the body, the corporation on receipt and approval of proofs will pay indemnity The aggregate of such indemnity benefits in no case exceeding the sum assured stated on the face of the policy or on an endorsement therefore.

The sum insured of the rider usually identical to that of the basic life policy, it cannot be more. For example, if the sum assured of the basic policy is Birr 10,000 the sum insured of the rider is also limited to Birr 10,000

Generally, policyholder's beneficiaries, if

- the losses arise solely from the "bodily injury" referred to in the above definition of the risk covered; and
- the losses occur within 90 days from the date of accident.

As it is stated above, the rider covers death or dismemberment or loss of time by accidental means. In the event of the death of the insured as above, the sum insured is paid to his beneficiaries together with the sum assured of the basic life policy. The total amount therefore, will be double the sum assured of the basic policy. But if the injury does not result in death rather

in dismemberment, benefits are paid for the loss of sight, limbs and thumb and index finger. The loss of one limb, sight in one eye and thumb and index finger of one hand , half of the sum insured is paid however there is the provision that the maximum benefit payable for any combination of losses is the sum insured. If the injury does not result in death or dismemberment but alone shall cause:

a) Temporary, Total and Continuous disability which prevent the insured from the date of accident from performing any and every duty pertaining to his occupation, a weekly indemnity of 5 per mille of the sum insured is paid; or.
b) Temporary Partial disability which prevents the insured from the date of accident or immediately following total disability under a) above from performing one or more duties pertaining to his occupation, a benefit of ¼ of the weekly indemnity is paid during the period of such partial disability, provided that the period for which weekly benefits are paid under a) and b) above shall not exceed 52 weeks in respect of any one accident, and further provided that disability of less than one week duration, benefits are shall not be paid .
c) Permanent total and continuous disability which prevents the insured from engaging in any occupation or employment for wage or profit or from giving attention to any business whatever and provided that indemnity has been paid under a) and b) above for 52 weeks, a monthly benefit equal to 0.88 of the sum insured is paid for the duration of such disability in no case beyond the point where aggregate payments under a), b) and c) equal the sum insured.

The definition of disability ought to be noted. For the first 52 weeks, it is the inability of the insured to perform his occupation but beyond the first 52 weeks, it is inability to perform any occupation.

4.1.1.4.2 DISABILITY WAIVER OF PREMIUM CONTRACT (WP)

The WP contract provides protection against the possibility that prolonged disability of the insured due to accident or sickness might prevent him from continuing to make his premium payments. If the insured is totally and permanently disabled and if such disability continues for a

specified minimum waiting period (such as 6 months), the insurer will "waive" the payment of premiums (including those of attached riders) falling due under the policy for as long as the disability continues or until all premiums are waived, whichever is shorter. The effect is the same as if the policy owner had actually paid the premiums himself. This, therefore, "insures the insurance" since it prevents the policy form lapsing for nonpayment of premiums due to disability of the insured.

4.1.1.4.3 TOTAL AND PERMANENT DISABILITY

The current EIC WP contract defines this as "Disability resulting from bodily injury or disease which prevents the insured from engaging in any and every business or occupation and from performing any work for compensation or profit and which disability has continued uninterrupted for a period of at least six months or the entire and irrecoverable loss of sight of both eyes, or the loss by severance of both hands above the wrists, or the loss by severance of both feet above the ankle, or the loss by severance of one hand above the wrist and one foot above the ankle." This can be disability due to either accident or disease; but the insured must be unable to engage him-self in any occupation as the result of the disability. The waiver of premium benefit is given only if the disability commences:

I. While the policy and the rider itself are in force on a premium paying status (i.e. not under ETI or paid-up); and
II. Prior to the anniversary date of the policy nearest the 60^{th} birthday of the insured.

4.1.1.4.4 COMPREHENSIVE ACCIDENT INSRUANCE (CAI)

This rider covers death, disability or loss of time. Loss must however be effected directly and independently of all other causes through external, violent, and accidental means of which there is evidence of visible contusion or wound on the exterior of the body.

This is newly developed product. The accidental cover provided only to dismemberment and of major human body parts is now the cause of dissatisfaction among the customers life insurance. Hence to avert such dissatisfaction, they intended to put in place full accidental

benefits stated under workmen's compensation or personal accident polices attached to the main life assurance policy. Disability as herein appears is disability which is the result of an accident and must be Total and Permanent and such that there is neither then nor at any time thereafter any work, occupation, or profession that the life assured can ever sufficiently do or follow to earn or obtain any wages, compensation or profit.

Accidental injuries which independently of all other causes and within 90 days from the happening of such accident, result in Loss of sight or the amputation of a limb shall be deemed to constitute such disability

4.1.1.4.5 SUPPLEMENTARY PRE-NEED FUNERAL EXPENSES RIDER

This is a rider attached with a permanent policy. On death of the policy holder the amount (sum assured under the rider) is payable. This amount is usually equal to the average cost of funeral expanses the dependents of the deceased policy holder would incur during the funeral process.

4.1.1.4.6 SUPPLEMENTARY TERMINAL ILLNESS /DREAD DISEASE RIDER

This is rider attached to permanent assurances. The benefit is payable on diagnosis of a terminal illness or dreaded disease as defined in the policy document. The benefit is 50% of the basic sum assured up to a maximum of Birr 40.000 if the life assured suffers from certain specified diseases or conditions during the term of the policy. The specified diseases and conditions covered under this rider include: cancer, stroke, coronary heart, disease, major organ transplant, total renal failure, paraplegia, blindness, major burns and coma.

4.1.1.4.7 MEDICAL EXPENSES INSURANCE

As pointed out previously disability due to accident or sickness can bring about significant economic loss to anyone. The losses were also identified i.e. loss of treatment necessarily incurred. Health insurance policies can provide cover for such losses by providing reasonably adequate disability income and by reimbursing a substantial portion of the medical expenses

incurred. The EIC health insurance portfolio is composed entirely of policies which provides only for the reimbursement of medical expense

4.1.1.4.8 GROUP LIFE INSURANCE

The discussion of life insurance in the previous topics was related to life insurance issued on the life of one person. However it can also be arranged whereby one policy covers the lives of a group of persons.. Just as in individual life insurance, the form of group life insurance can be term, whole life or endowment. Any collection of individual which presents itself as a group is not necessarily acceptable for group life insurance purposes. To be accepted as group, it needs to satisfy all of the following requirements.

- The group must have a minimum of 20 members
- The group must have been formed for a purpose other than that of obtaining insurance
- The group must be of such a nature that it is composed of persons all of whom are actively at work for wages, salaries, profit or other compensation
- It is desirable that the group be formed for an indefinite period and that there and that there will be a regular entry of new and young members into the group.
- The group must have facilities for central administration of the group scheme
- The group organization must have legal capacity to enter into contract

4.1.1.5 CLAIMS ADMINSTRATION

In majority of cases in which a claim is resisted (rejected) it would likely be so resisted by all insurers. Claim administration is a decision making process which involves interpretation of the contract language. Companies may differ in their interpretation and in their practice.

The different claim decisions are due to different responses by the insurers to various forces, both internal and external, which influence claim administration. Some of them are stated below.

I. Insurance Regulators

Insurance is a regulated business, having its own statutes governing that business. Insurance statutes and regulations have an important influence on claim administration because they carry the force of law.

II. The Judiciary

With insurance, as with any type of contract disputes may arise over interpretation of contract language for a given set of facts. It is overly simplistic to believe that policy language, clearly stated, controls every situation. Numerous rulings could be cited in which courts have held that certain rights of claimants overrode policy provisions. For e.g. in one case the insured died after paying the initial premium but before the application was approved. The company contended that an ambiguous language postponed coverage until the application was approved. The court however ruled for the claimant, holding that an average person would suppose the coverage began upon completion of the application requirements. Thus what the policy says must be interpreted in light of court rulings that are both relevant and recent. If there is any ambiguity, it most likely will be resolved against the insurer. What is more important, courts are giving increasing attention to the **"reasonable expectation"** of applicants, insured's and beneficiaries. The old doctrine of **caveat emptor, let the buyer beware,** has been weakened considerably.

III. The General Public

The general public, inevitably, has influence on claim decisions. This is only natural, as insurance is provided as a service to the public. This influence is felt in many ways and for many reasons. The public is the potential market; the public has become increasingly powerful through consumer activities in influencing governmental actions **(legislative, executive and judicial)**. The public looks upon the payment of the claim as the fulfillment of obligation undertaken by the insurer when the policy was sold.

IV.The Policy owner

Most claims are paid from the premium receipts. If an insurer follows a "liberal" claim practice compared to another company, it may find itself forced to increase premium rates more sharply as a result. This, in turn could result in poorer persistency as policyholders turn

to other insurers for coverage. The influence of premium rates is also pronounced in group insurance, both life and health. The interests of policy owner also require equitable claim treatment without unfair discrimination.

V. The Company (Insurer)

In essence, the claim department delivers the promises that the agent or broker has sold to the policy owner. When the need arises for filling a claim, it is often the agent who is first involved receiving the notice of loss and then latter explaining the claim decision. If both the agent and client are disenchanted with the claim performance, both may look elsewhere for employment and placing their business. Again the claim function is under a typical company pressure of cost effectiveness.

The aforementioned information's were very helpful to understand the business practice of life insurance held at EIC.

4.2 Data understanding

In this phase, the initial data collection, data description and data mining goal are going to be discussed. During this phase the investigation starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information

During the interviews, the Deputy Manager of ICTM department explained that the INSIS system database handles more than two thousand tables. There are 26 employees who are hired to operate and support the system. . From those 20 of them work on application of the system. And again from the twenty of employees 6 of them work on the hardware part of the system. The data stored in the EIC INSIS data base divided in to three general domains: marketing, non-life and life tasks. Marketing data are such as agent name, agent Id, agent commission, name the company of they attract to the insurance company and others. Both Non-life and life data have underwriting and claim data separately. From two thousand tables of data the researcher was allowed to use some of the data for the confidentiality and security of EIC customers' data.

Especially information such as premium and its related were restricted to use for their security and privacy issues. Based on the relevance criteria the data, four tables were selected from the database namely; p. people table, insured type (insurance value table), Cover type table (risks to be covered) and occupation table. The selection was made on based on the policyholder's personal, demographic, transactional, policy and agent commission information.

4.2.1 Dataset collection

With the help of experts, data regarding to life insurance was extracted from the LIFE INSIS database and exported to MS –Excel. For the purpose of the study customer information are selected from four tables. Every insured person has an Insured ID, which is unique to the individual insured. The insured individual's data are recorded under different kinds of sectors such as private individuals, government organizations and non-government organizations. Data on policies, terms and personal information are recorded individually for every insured individual's. For this stud, historical data of 12 months or one year was taken from August, 2011 to August, 2012. Using historical data based on time frame can avoid the volatility segments revealed in the study result. The raw data collected has 16 attributes and 27845 records. By considering the confidentiality of the data, the null values and the relevance the attributes to the study objective, attributes selected from the four tables are stated below.
From risk covered (COVER_TYPE) table INSR_TYPE, INSURED_VALUE, INSR_BEGIN, INSR_END, DATE, and COVERED_TYPE are selected.

From P_PEOPLE table attributes namely; BIRTH_DATE SEX, AGE, COMPANY, OCCUPATION_ID, MARITAL-STATUS and FORIEGNER are selected.

From INSURED_TYPE table attributes namely; AGE, INSURED_STATE INSURED_VALUE, INSURED_BEGIN, INSURED_END, DURATION and DATE_COVERED were selected.

And finally from OCCUPATION table OCCUPTION ID (which describes the policy holders' occupation type) are selected. With the help of domain experts in, customer data records were exported to Microsoft Excel. Generally, 16 attributes selected are selected by avoiding the redundancy and by considering to degree of their relevancy to the study area. They are presented as follows:

INSURED_ID, INSR- TYPE, INSURED _VALUE, SECTOR, COMPANY-TYPE, BIRTH – DATE, SEX, AGE,OCCU PATION_ID, MARITAL _STATUS, FORIEGNER, INSR_BEGIN, INSR_END, COVER_TYPE, RISK _STATE and DURATION

Additionally two attributes were derived from the computation of CLV (Customers Lifetime Value) namely; current value and value-seg.

4.2.2 Data Description and Data Mining Goal

Useful patterns and knowledge were extracted from customers' information using clustering and classification. Descriptive data mining models such as clustering methods are used to describe the segments of and the relationship among attributes and classes. Predictive data mining model such as decision tree and regression are more powerful to forecast the future value of customers segments categorically or numerically. Generally data mining techniques and their interpretation models results can enhance and provide EIC life insurance managers and marketing managers' meaningful information regarding their customers. For this reason an attempt was made:

a) to collect and organize customers datasets from the EIC life insurance database;
b) to take all necessary data preparation steps to produce dataset;
c) to compute the customers life time value elements based the customers data;
d) to find an appropriate segment based on the historical data of EIC customers and to describe the relationship with in each other;
e) to reveal which customer segment is more valuable for the organization or which groups customers contribute more revenue to the organization; and
f) to describe the characteristics of those groups and the relationship among them.

4.2.3 CLV computation

The CLV computation has two elements known as current value and future value [28]. The computation of lifetime value is described as follows:

$$\text{LTV} = \text{current value} + \text{potential value} \quad \text{..... (a)}$$

The current value is computed based on the amount of payment the insured are asked for the service that the insurer provide (sum assured) minus the cost that the insurer incur to attract the customer divided by the duration or year of their total policy will expired.

$$CV = \frac{\text{Average amount} - \text{Cumulative amount}}{N} \quad \cdots\cdots (b)$$

- average amount is the average amount that customer asked to pay for the service
- cumulative amount is the average charge customers are asked to pay for the service and
- N is total period of service

The potential value of calculation is for customer lifetime value

$$PV = \sum_{j=1}^{n} prob_{ij} * profit_{ij} \quad \cdots\cdots (c)$$

- $prob_{ij}$ is the probability that customer i will use the service j among n-optional services.
- $profit_{ij}$ means the profit that a company can receive from the customer i who uses the optional service j.

In addition to that two derived attributes were contained within the dataset based on the computation of customer life time value (CLV) component called current value. They are called curr-value and value segmentation (value seg). The former one is the numerical representation of the computation while the latter implicates categorical values of current value computation. In this study, the categorical computation of current value is classified in to two classes or segments known as “high value” and “low value”. The segmentation is computed by using the median split technique. This was done by the help of experts and literatures.

Table 3 presents the initial collection along with their description. It consists of attributes name, descriptions, and their types.

No	Attribute Name	Description	Data types
1.	INSURED_ID	The unique identification of Insured person	Numeric
2.	INSR- TYPE	The type of insurance policy purchased	Nominal
3.	INSURED _VALUE	The payment made to insurer from insured person	Numeric
4.	SECTOR	The business type of the insured person or organization	Nominal
5.	COMPANY-TYPE	The type work that the insured is engaged	Nominal
6.	BIRTH –DATE	The birth data of the insured person	Date
7.	SEX	The sex of the insured person	Nominal
8.	AGE	The age of the insured person	Numeric
9.	OCCUPATION_ID	The unique identification of the work of the insured person	Numeric
10.	MARITAL _STATUS	The marital status of the insured person	Nominal
11.	FORIEGNER	The nationality of the insured person	Nominal
12.	INSR_BEGIN	The date , month , year and time of that the insurance policy starts	Date
13.	INSR_END	The date , month , year and time of that the policy insurance ends	Date
14.	COVER_TYPE	The type of insurance that the insurer	Nominal
15.	RISK _STATE	The payment status of insured person	Numeric(0.1 1.12)
16.	DURATION	The duration that the insurance covers	Numeric

Table 3 Initial Attributes Collected from EIC Life Insurance and Their Description

Based on the above information those attributes or variables can be classified in to personal and demographic attributes and life insurance policy and transactional attributes.

Personal and Demographic Attributes	**Life Insurance Policy And Transactional Attributes**
1. AGE	1. COVER_TYPE
2. BIRTH –DATE	2. DURATION
3. COMPANY_TYPE	3. INSR_BEGIN
4. FOREIGNER	4. INSR_END

5. MARITAL_ STATUS	5. INSR_TYPE
6. OCCUPATION_ID	6. INSURED_ID
7. SECTOR	7. INSURED_VALUE
8. SEX	8. RISK_STATE

Table 4 Classification of customer information

Insurance types attribute in LIFE INSIS database is represented by unique numeric numbers and it describes the policy types that customers purchased. **Table 5** describes unique values of insurance policy types and their descriptions.

0	Insurance type (insr-type)	Description
1	5001	Ordinary endowment
2	5002	Anticipated endowment
3	5003	Educational endowment
4	5004	Annuity endowment
5	5005	Joint endowment
6	5006	Annuity joint endowment
7	5007	Group endowment life
8	5101	Whole life
9	5201	Individual term insurance
10	5202	Individual mortgage protection insurance
11	5203	Group term insurance with yearly renewable
12	5204	Modified large group term insurance
13	5205	Group mortgage protection insurance
14	5206	Joint mortgage
15	5207	Pre needed funeral expense insurance
16	5208	Group Pre needed funeral expense insurance
17	5301	Medical individual
18	5302	Medical group

Table 5 Lists of insurance types and their description

Cover _type attribute has different categories, which indicate the risk covered by the policy. **Table 6** presents descriptions of cover types for the policy customers purchased.

No	Cover type	Description
1	SAI	Supplementary accident insurance – death and permanent disability
2	CAI	Complementary accident insurance – death, total , partial and permanent disability medical
3	PNFE	Pre- needed funeral expense insurance
4	WP	Waiver premium
5	HIHOSPSICK	Hospitalization and sickness
6	HIMAT	Maternity benefit
7	HIPREGN	Pregnancy check-up
8	HIEYE	Eye glass benefit
9	HIDENTAL	Dental check-up
10	DEATH	Death
11	DEATENDOW	Death and endowment

Table 6 List of cover types and risks covered by the policies

Table 7 depicts categories of payment status or risk state of policyholders with their descriptions

No	Risk state	Description
1	0	Initial payment was made
2	11	Half premium payment
3	12	Full premium payment

Table 7 Lists of Payment Status of Customers and Their Description

Table 8 illustrated the marital status of policy holder's and their labels tp provide whether the person is married, divorced or single

No	Marital status	Description
1	M	Married
2	D	Divorced
3	S	Single

Table 8 Lists of marital status labels and their description

The original data set contains 16 attributes and one derived attribute. The aim of the next phase of data pre-processing was to produce datasets that contains only those relevant attributes and insurance needed for analysis among customers demographic information along with basic policy information are taken for consideration. Finally, the selection of the dataset is performed by the assistance of literatures and domain experts.

4.3 Data preprocessing

During this phase multiple steps are to be followed to prepare data set for data mining tool for modelling purpose. Tasks such as data cleaning, data selection, data transformation, data integration, and organization of format of the data were going to be performed.

4.10.1 Data Cleaning

Data cleaning involves different techniques in order to fill in missing values, to handle outliers and to smooth noisy data and to detect inconsistencies and correct the data set [14]. In order to get quality dataset for experiment data cleansing is required. Tools such as DataPreparator and Weka are used for cleansing.

I. <u>Missing value</u>

Missing value prediction is a data cleaning activity in order to increase the quality of data and to get better results. There are various reasons why missing values occurs. For instance, they may have empty values, incomplete data and non-existing data. Therefore, values were identified and corrective measures were done on the raw data [17]. In general, missing values occurred in (6) six attributes namely; SECTOR, MARITAL_STATUS, OCCUPATION _ID, COVER_TYPE, RISK _STATE and DURATION. For instance, Sector had 20683 missing values while OCCUPATION _ID had 20640 and MARITAL STATUS of 20673 missing values from the total number 27845 instances. These attributes are removed from the data set because the uniformity and incompleteness of their instance values. But for Cover _Type, Risk _State and Duration

attributes only one missing value was identified. Modal values were used to replace the missing values for nominal attributes (Cover _Type, Risk _State) while the mean values were used to replace the missing values of the numeric attribute (Duration).

Attributes Containing Missing Values

Row	Attribute Name	Type	#Labels	Num Missing	% Missing	Select
0	SECTOR	nominal	2	20,684	74.309	☑
2	MARITAL_STATUS	nominal	5	20,674	74.273	☑
1	OCCUPATION_ID	numeric	0	20,641	74.155	☑
3	COVER_TYPE	nominal	12	1	0.004	☑
4	RISK_STATE	nominal	4	1	0.004	☑
5	DURATION	numeric	0	1	0.004	☑

Figure 11 missing valuues from EIC dataset with DataPreparator-1.7

Options

Numeric: Mean | Update

Nominal: Mode | Update

Numeric Attributes Containing Missing Values

Row	Attribute Name	Min	Max	Mean	St Dev	Num Mis...	% Missing	Impute	Select
0	OCCUPATION_ID	9,001	90,014	10,172.174	3,770.36	20,641	74.155	Mean	☐
1	DURATION	1	25	9.115	2.738	1	0.004	Mean	☑

Num Selected: 0 | Deselect All | Select All

Nominal Attributes Containing Missing Values

Row	Attribute Name	# Labels	Mode	Modal Freq	Num Missing	% Missing	Impute	Select
0	SECTOR	2	Private Indiv...	7151	20,684	74.309	Mode	☐
1	MARITAL_STATUS	5	S	6246	20,674	74.273	Mode	☐
2	COVER_TYPE	12	DEATH	9453	1	0.004	Mode	☑
3	RISK_STATE	4	11	25058	1	0.004	Mode	☑

Num Selected: 2 | Deselect All | Select All

Figure 12 Replacing Missing Values for numeric and nominal attributes with DataPreparator-1.7 software

II. Handling outliers

During data cleansing some attributes showed outlier data. For instances "Occupation ID" attribute has 9 outliers. Those outliers were identified because there were discrepancies between the Occupation ID stated in the dataset and the description unique values of insured occupational information. In other words, the outliers of Occupation ID attribute, do not resemble the occupation type of the customers (have no meaning). Moreover, they were not mentioned in the descriptions of the customer occupational information gained. Therefore, those records were removed. On the other hand, to handle outliers in numeric attributes Z-Score method was applied. For instance, outlier values of the **insured value** attribute were corrected using the Z-Score method.

The Z-Score method uses the Z-Score statistic defined as:

$$\text{Z-Score} = \frac{(\text{value} - \text{mean})}{\text{standard deviation}}$$

It gives the number of standard deviations a value is above or below the mean. An outlier is a value that has Z-Score above a specified upper limit or below a specified lower limit. In this study, Winsorize option is used to replace outliers with the values corresponding to the specified Z-Score limit in DataPreparator 1.7 tool. For this study; first, the minimum and maximum z-scores are calculated and adjusted in the option. Then, Winsorize option is selected to handle on the numeric attributes to replace outliers registered. This method was applied on insured value and current value attributes.

III. <u>Noisy data</u>

Among COVER_TYPE attribute labels HPREGNANCY and HMAT labels were the indication of Maternity benefit and Pregnancy check-up respectively for female sex. But some of them are recorded as cover type attributes of the mentioned labels were recorded for male instances. Due to inconsistency with respect to the sex -10 instances were removed

4.10.2 Data integration

Data integration was is crucial step as it reduces problematic issues due to the redundancy of customers' data, the format of the data and levels of data. 27845 records and 16 attributes from

four (4) tables were selected and exported to MS –Excel from the LIFE INSIS database. The beginning and end date of the policy purchased were cross checked with the duration of insurance type or policy it covers. In addition, the formats of the date are adjusted to avoid discrepancy. In the case of *Insured Age*' the *Birth Date* of the insured was used as a reference to confirm and calculate the actual age. Due to wrong spelling, some inconsistencies were avoided among the work description of the insured with their Occupation-Id. Finally, to meet the goal of data mining the above four tables are merged to get integrated data of life insurance customers of EIC.

4.10.3 Data transformation

The attribute named as "Duration" instances were transformed from continuous to categorical with the help of experts. Pre-transformation the numeric values of the attributes were labelled 1, 4 5, 9, 10, 12, 15, 16, 17, 20, 21, 22 23, 25. The new label names given to the instances of attribute "duration "are divided into three categories namely as short term (from 1-5 year) , medium term (from 6-10) and long-term (from 11-25 year).

'Risk state' attributes instances labelled as 0, 11, and 12 depicts the three payments made by insurer namely: initial, half and full payment to insurer are changed to their original name to add meaning full information to the output. Therefore, the new instances name for 0 is "initial payment", for 11 is "half payment" and for 12 as "full payment". And also the name of the attributes changed from "risk state" to "payment status".

Insurance types 5001, 5201, 5202, and 5301 are replaced by A, B, C & D respectively

In the initial data set the labels of instances "sex" attributes are represented in numeric numbers such as: 1 for male and 2 for female. So to minimize the ambiguity that may occur during the model building (when rules are generated), they are changed to M and F labels to represent sex attribute instances.

4.10.4 Discretize Numeric Attributes

Discretization transforms numeric (continuous) attributes to nominal (categorical or discrete) attributes. The range of a numeric attribute is divided into intervals and each interval is given a label. Attribute values are replaced by the labels of the intervals into which they fall. Using discretization method can give generalized information which is easier and meaningful to

interpret data mining results conducted on different data mining tasks. As a result, the experiment conducted on different data mining techniques and algorithms will have consistent representation of dataset. Generally, using reduced number dataset that prepared through discretization (interval labels) over large dataset (ungeneralized dataset) advances the mining results more efficient, consistent, simplified and easy to interpret and represent. [2]. Here, insured value and the current value attributes are transformed into categorical or discreet values. In this the study, to discretize the aforementioned attributes, Equal width discretization was used to divide the ranges of a numeric attribute into a specified number of intervals of equal width. This method considers the class information

AGE is divided in to three (3) intervals 0- 14, 14- 28, 28-41 for lower age, middle age and adults respectively by using *Weka .filters. Unsupervised. attribute. Discretize* option, which uses equal-width method. The number of bins is set to 3.

No	Age	Numbers of tuples
1	0 -14.333333]'	67
2	('(14.333333-27.666667]	9527
3	'(27.666667-41)'	12028

Table 9 Unsupervised Equal Width Discretization

Then by using a text editor, the categories of ages numbers are assigned to their approximate values (0- 14, 14- 28, 28-41)

Another discretization technique applied is on insured values in by using WEKA Supervised discretization under *filter* option. This was selected because the supervised discretization tries to generate intervals by considering insured value of customers' data to make consistent class distribution.

No	Insured value	Numbers of tuples
1	(min value -67719]'	10161
2	(67719-77366.5]	11441
3	'(77366.5-max value)'	20

Table 10 Supervised Equal Width Discretization

4.10.5 Scaling or normalization of Numeric Attributes

Scaling is required for data mining algorithms that accept only attribute values within certain ranges. It is also required to prevent bias when attributes have very different ranges. In this study, to normalize numeric attributes of the data set, Z-Score normalization was used to transform attribute values using the following formula.

$$V_B = \frac{V_A - m_A}{S_A}$$

m_A = the mean of attribute A

S_A = the standard deviation of attribute A

V_A = a value of attribute A

V_B = the value obtained by transforming V_A

The normalized value of insured value attributes after discretization

No	Normalized insured value	Numbers of instances
a.	'(min value -0.115407]'	10161
b.	(0.115407-0.132349]'	11441
c.	'(0.132349- max value)'	20
Total	3	21622

Table 11 Normalization of "Insured Value" Attribute

4.10.6 Data Reduction

4.10.7 Dimensionality reduction

The dimensionality of the datasets is reduced by attribute selection using the information method in order to choose best attributes before the modeling task is implemented.

4.10.7.1 Information gain

Information gain evaluates the worth of an attribute by measuring the information gain with respect to the class. The evaluation formula of information gain is described below.

Info Gain (Class, Attribute) = H(Class) - H(Class | Attribute).

The result of information gain value of attributes is shown below.

Ranking	Attribute	Gain value(%)
1.	INSURED_VALUE	0.984938
2.	AGE	0.846832
3.	SEX	0.032486
4.	COVER_TYPE	0.004398
5.	PAYMENT_STATUS	0.002556
6.	Duration	0.000293

Table 12 Information Gain Ranking Filter

Table 13 presents the lists of attributes selected for the final dataset and their types

No	Selected attributes	Data Types
1.	Age	Nominal
2.	COVER_TYPE	Nominal
3.	Duration	Nominal
4.	INSURED_VALUE	Nominal
5.	PAYMENT_STATUS	Nominal
6.	SEX	Nominal
7.	value-seg	Nominal

Table 13 Attributes Selected For Final Dataset and Their Description

Finally, after attribute selection and data preparation process (data cleaning. reduction transformation and integration) takes place, the amount records of to be used in the dataset must be specified for modelling purpose.

4.10.8 Instance Selection

In this study, in addition to attributes selection, removal of irrelevant instances was conducted. For instance among instances of attribute of insurance types, only those policies which indicate individual insured person was selected. So that polices bought by group of insured was removed. After all necessary steps and appropriate tasks of data preparation process, total numbers instances removed are 6223. So when deduct from the dataset before pre-processing task of 27845 datasets, it gives the total amount of 21622.

After the calculation of customer lifetime value on the current value component, insured customers are segmented into high current value and low current value. Experts' advices were considered with the results of median split value to segment life insurance customers of EIC. Customers which fall above an average of 5880 current value are grouped as those have "high value" while those customers which fall below the average of current value labelled as low. After the computation the derived attribute named as "curr- value" is removed and replaced by value segmentation attribute called "value –segm" from the data set.

Summary

In this chapter an attempt was made to understand the business practice of EIC at LAD, to understand the data that are relevant to the study domain and to prepare relevant attributes for modeling purpose. In business understanding phase the business practice in EIC life insurance is conducted through interviews with domain experts and document analysis.

During data understanding phase, 12 month policy holders' data were collected from LIFE INSIS database from four tables, namely; P. PEOPLE, INSURED_TYPE, COVER_TYPE and OCCUPATION tables. The demographic, personal, policy and transactional information of policyholders' were chosen.The initial dataset is made16 attributes and 27845 records and the customer value is computed using CLV model that the supports to categorise the customer value. Current value and value-seg were derived from the computation

The data preprocessing phase is lengthy and time consuming task in relation to the other phases it involves consists of major tasks such as data cleaning, data transformation, data reduction

Six attribute namely; SECTOR, MARITAL_STATUS, OCCUPATION _ID, COVER_TYPE, RISK _STATE and DURATION registered missing values. Three attributes namely; SECTOR, MARITAL_STATUS, OCCUPATION _ID registered high missing values (>74%) and they removed from the dataset. Nominal attributes (COVER_TYPE, RISK _STATE), which registered one missing value each, modal values were used to replace the missing values. Numeric attribute such as DURATION, which registered one missing value, is replaced by mean value. The duration of policy the purchased is transformed categorize into long-term medium term and short term. The age of policy holder is transformed into 0-14, 14-28 and 28-41 categories. INSURED_VALUE was discretized to replace the labels by intervals and it was

normalized using WEKA Supervised discretization option. The dimensionality of the data was evaluated using information gain evaluation method. Totally, 6223 instances were removed are in pre-processing task. Finally after preprocessing task attributes from initial data set such as INSURED_VALUE, AGE, SEX, COVER_TYPE, PAYMENT_STATUS, DURATION and one derived attributes called “value seg” were selected . Generally 7 attributes and 21622 records were included in the final datasets, for modeling purpose

CHAPTER FIVE

MODEL BUILDING AND EVALUATION

In the previous chapter, data understanding and data preparation phases were discussed along with their tasks. It was indicated that data preparation was conducted based on the researcher's understanding of the data, and the final output of the data preparation stage was made to produce the dataset for modelling purpose. In this chapter, the selection and application of suitable data mining techniques, algorithms and tools as well as the evaluation of the model built are presented. The choice of DM tools, techniques and algorithms should consider the data mining objectives as it was stated in chapter three under business understanding phase. In addition, the nature of the data influences the model to be used for the study.

In this study, the Weka Data Mining Software was used to build the model and evaluate the result of the model built. Therefore, based on the parameters and quality required, the researcher used Weka 3.7.9 software to implement the model building and evaluation processes.

5.1 Experimental Design

To identify the appropriate method for a given modelling purpose, one should use various types of viable models along with a well-defined experimentation and assessment strategies because there is no universally known best method or algorithm for a data mining task. Accordingly, the model-building step of this study encompasses the assessment and comparative analysis of the various models built so far.

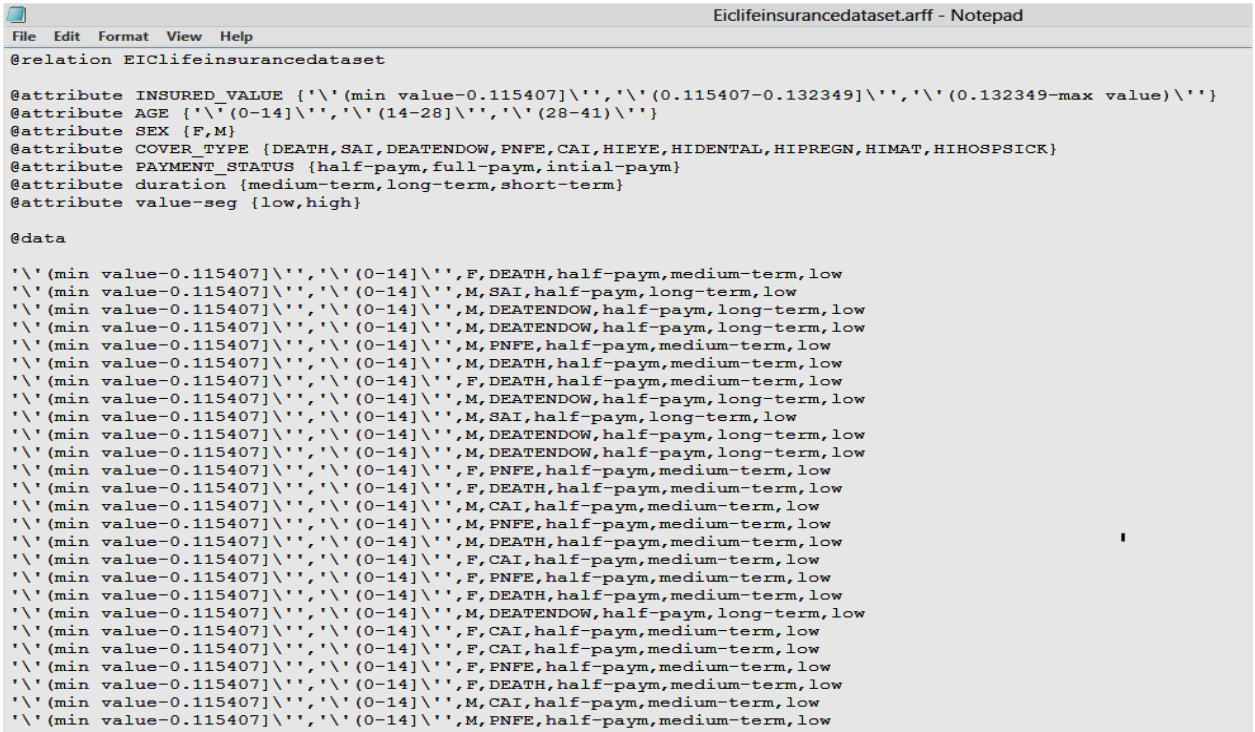
The goal of data mining is to find the hidden patterns to uncover the unknown knowledge/new knowledge or relationship or variation among the data found in specific database or dataset.

Data mining tasks can be categorized as supervised and unsupervised learning methods, and both of the learning methods embrace many techniques and algorithms. The goal of this study is to apply data mining techniques to segment customers based on the customer lifetime value for betterment in the customer relationship management (CRM) of the organization. Based on these facts, both supervised and unsupervised learning methods were experimented within this study. Accordingly, the J48 decision tree algorithm was selected from the supervised learning methods

whereas k-means clustering algorithm was selected from the unsupervised ones. The reasons for the selection of these algorithms are explained in the subsequent sections.

5.1.1 Format of the Dataset

Before the model building task was implemented, the dataset was first saved in Comma Separated Value (CSV) format and then was converted to Attribute-Relation File Format (ARFF) and saved using a text editor. In other words, the columns and the rows in CSV format (EIClifeinsurancedataset.csv) are converted into lists of attributes and data in ARFF file format (EIClifeinsurancedataset.arff). The ARFF has two different parts; namely, the Header and the Data. The Header part contains the declared name of the relation that can express the dataset (@RELATION <relation-name>) and the declared lists of attributes with respect to their data types (@ATTRIBUTE <attribute-name> <data type>). The data types of an attributes can be nominal, string or numeric. The data part of ARFF contains collections of records in relation to their attributes. The data-sets of Ethiopian Insurance Corporation life insurance at Life Addis District in ARFF format is indicated below in figure 13.


Eiclifeinsurancedataset.arff - Notepad

File Edit Format View Help

```
@relation EIClifeinsurancedataset

@attribute INSURED_VALUE {'\'(min value-0.115407]\'','\'(0.115407-0.132349]\'','\'(0.132349-max value)\''}
@attribute AGE {'\'(0-14]\'','\'(14-28]\'','\'(28-41)\''}
@attribute SEX {F,M}
@attribute COVER_TYPE {DEATH,SAI,DEATENDOW,PNFE,CAI,HIEYE,HIDENTAL,HIPREGN,HIMAT,HIHOSPSICK}
@attribute PAYMENT_STATUS {half-paym,full-paym,intial-paym}
@attribute duration {medium-term,long-term,short-term}
@attribute value-seg {low,high}

@data

'\'(min value-0.115407]\'','\'(0-14]\'',F,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,SAI,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,PNFE,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,SAI,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,PNFE,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,CAI,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,PNFE,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,CAI,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,PNFE,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,DEATENDOW,half-paym,long-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,CAI,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,CAI,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,PNFE,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',F,DEATH,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,CAI,half-paym,medium-term,low
'\'(min value-0.115407]\'','\'(0-14]\'',M,PNFE,half-paym,medium-term,low
```

Figure 13 ARFF format of EIC life insurance dataset

5.2 K-Means Clustering: Model Building and Analysis

Clustering is one data mining techniques used to make essential segments or groups that have common characteristics. This technique produces data-driven segments which are not known before. In another words, they are applied for unsupervised segmentation to induce natural groupings of records/customers with similar characteristics. The revealed clusters are directed by the data and not by subjective and predefined business opinions [9].

In this study, k-means clustering algorithm was applied for customer segmentation task. The K means clustering algorithm is one of the unsupervised learning methods which clusters' the subsets of the dataset to the nearest mean or to the centroid. The advantage of using k-means algorithm for clustering purpose is because of its speed and scalability. K- Means was selected because it is one of the fastest clustering algorithms, which can efficiently handle long and wide datasets with many records and many input clustering fields. [9, 20]

In the k means algorithm, we:

1. Place K points into the space represented by the objects that are being clustered .These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated

For this study, the total number of instances used in the dataset of EIC is 21622.During the model building phase of clustering using Simple K-Means algorithm, the parameters of the algorithm are tuned with the following options .

Parameters Used For Tuning	Selected Option
1. Don't replace missing values – This option Replaces Missing Values Globally with Mean/Mode.	True
2. Display standard deviations – this option is used to display standard d deviations of numeric attributes and counts of nominal attributes in the experiment result.	True

3. Number of clusters – this option is used to set number of clusters	2 (Default Value)
4. Seed – This potion is used the random number seed to be used.	10 (Default Value)
5. Maximum iterations – this option is used to set maximum number of iterations	500 (Default Value)
6. Distance function -- the distance function to use for instances comparison	Euclidean Distance (Default)

Table 14 Parameters of k-mean cluster model

Four sets of experiments were conducted on EIC life insurance dataset that consists of 21622 instances and seven attributes (AGE, COVER_TYPE, Duration INSURED_VALUE, PAYMENT_STATUS, SEX, and value-seg) using the k means algorithm.

Experiment #1

The first experiment is conducted by setting the k value to 2 and default seed value to 10

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     EIClifeinsurancedataset
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
Ignored:
              value-seg
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 16743.0

Initial starting points (random):

Cluster 0: '\'(0.115407-0.132349]\'','\'(28-41)\'',F,PNFE,half-paym,medium-term
Cluster 1: '\'(min value-0.115407]\'','\'(14-28]\'',F,PNFE,half-paym,medium-term

Missing values globally replaced with mean/mode
```

Figure 14 Run information of experiment#1

The k means clustering results with standard deviation in respect to their attribute values are presented in Table 15**.**

Cluster index	***INSURED_VALUE***	***AGE***	***SEX***	***COVER_TYPE***	***PAYMENT_STATUS***	***Duration***
Cluster 1 12032 (56%)	min value-0.115407]'=4% 0.115407-0.132349=95% 0.132349-max value= 0%	0-14=0% 14-28=0% 28-41=99 %	F=98% M= 1%	DEATH =33% SAI=0% DEATENDOW=0 % PNFE=33% CAI =33% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0 %	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=99% Long-term=0% Short-term =0%
Cluster 2 9590 (44%)	min value-0.115407]'=99 % 0.115407-0.132349=0% 0.132349-max value= 0%	0-14=0% 14-28=99% 28-41=0%	F= 88% M=11 %	DEATH =33% SAI=0% DEATENDOW=0 % PNFE=32% CAI =32% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0 %	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=98% Long-term=0% Short-term =0%

Table 15 K value =2 and seed value =10

From the above description of the clustering model (cluster#1), 12032 instances (56% of dataset) were assigned to first cluster (cluster#1).It is slightly greater in number of instances than the second cluster. The clustering result shows that among insured value categories, customer who paid from less than 67719 of the total sum assured value makeup 4% from the group , customers whose sum assured value that lies between 0.115407-0.132349 (67719-77366.5) dominates the group by making up 95% of the group . And also policy holders who paid sum assured value greater than 77366.5 accounts 0%.

From the categories of ***AGE*** attribute, customers greater 28 years dominated the age groups by accounting for 99%. From the sex group female customers are accounts 98% while male customers' accounts for 1%. From the categories cover type attributes, DEATH consists of 33%, Pre- needed funeral expense insurance (PNFE) accounts for 33% and Complementary accident

insurance – death, total, partial and permanent disability medical (CAI) accounts for 33% within the group. Among categories of **payments status** attribute customer who made half payment dominated the group for accounting 99%. Customer who purchase insurance policy for medium-term accounts 99%.

The second cluster consists of 9590 instances which accounts for 44% of the EIC dataset. Based on the results of **Cluster #2**, among the group of the insured value attribute, policyholders who paid sum of insured value less than 67719 dominate the group by accounting for 99%. Customers whose age was between 14-28 accounts form 99%. From these group, female customer accounts for 88% while male customers form 11% of the group.

From cover type attribute categories, DEATH consists of 33% Pre- needed funeral expense insurance (PNFE) consists of 32% and Complementary accident insurance – which consists of death, total, partial and permanent disability medical (CAI) accounts for 32%. Among **payments status** attribute categories customer who made half payment dominated the group for accounting 99%. Customer who purchase insurance policy for medium-term accounts 98%

In the first experiment among six attributes, cluster one and cluster two share three attribute values in the cover type, payment status and duration. So in order to increase the intra class similarity and to decrease inter class similarity the researcher conducts the following experiment (**Experiment#2**)

Experiment #2

The second experiment is conducted by setting k to 2 and seed value to 100. **Figure 3** shows run information of **Experiment#2** and changes made k-means parameter in relation to the first experiment.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 100
Relation:     EIClifeinsurancedataset
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
Ignored:
              value-seg
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 28686.0

Initial starting points (random):

Cluster 0: '\'(0.115407-0.132349]\'',\'(28-41)\'',F,DEATH,half-paym,medium-term
Cluster 1: '\'(0.115407-0.132349]\'',\'(28-41)\'',F,CAI,half-paym,medium-term

Missing values globally replaced with mean/mode
```

Figure 15 Run information of Experiment#2

Table 16 summarizes the results of **Experiment#2** clustering model furnished with standard deviations to shows the values of the whole attributes for better understanding.

Cluster	***INSURED_VALUE***	***AGE***	***SEX***	***COVER_TYPE***	***PAYMENT_STATUS***	***Duration***
Cluster 1 14456 (67%)	Min-value-0.115407]'=47% 0.115407-0.132349=52% 0.132349-max value= 0%	0-14=0% 14-28=44% 28-41=55%	F=93% M=6%	DEATH =49% SAI=0% DEATENDOW=0% PNFE=49% CAI =0% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0%	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=98% Long-term=0% Short-term =0%
Cluster 2 7166 (33%)	min value-0.115407]'=46% 0.115407-0.132349=53% 0.132349-max value= 0%	0-14=0% 14-28=43% 28-41=55%	F=94% M=5%	DEATH =0% SAI=0% DEATENDOW=0% PNFE=0% CAI =100% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0%	half-paym=100% full-paym =0% intial-paym=0%	Medium-term=99% Long-term=0% Short-term =0%

Table 16 K value =2 and seed value =100

As it has been presented in **Table 16**, the clustering model built in experiment#2 shows that 14456 instances (67% of the data set) were assigned to cluster#1 and 7166 instances (33% of the dataset) were assigned to cluster#2 . The experiment shows that, cluster#1 is slightly greater in number of instances than cluster#2.

Under cluster#1, policyholders who paid the sum of assured value (***INSURED_VALUE)*** less than 67719 Birr registered 47% from the group and policyholders whose ***INSURED_VALUE*** lies in the range of 67719-77366.5 Birr registered 52% from the group.

From the categories age attribute, customers whose age is between 14-28 consists of 44% instances and customers aged greater 28 years accounts for 55%.

From sex group, female customers are accounting 93% while male customers' account for 6%.

From categories of cover type attribute, a customer who's ***COVER- TYPE*** (protections) covers for DEATH and PNFE (Pre- needed funeral expense insurance) each accounts for 49%. Among the categories of ***PAYMENT_STATUS*** attribute, customer who made half payment dominated the group accounting for 99%. Customers who purchased insurance policy for medium-term account for 98%.

Based on the result of **Cluster #2**, the following pattern was revealed.

From the categories of ***INSURED_VALUE*** attribute, customers who paid sum insured value less than 67719 Birr account for 46% while customers whose ***INSURED_VALUE*** is between 67719-77366.5 Birr makeup 53% of the group.

From the categories age attribute, customers whose age is between 14-28 consists of 43% instances and customers aged greater 28 years account for 55% .Female customers account for 94% while a male customers makeup 5 % of the group.

From cover type attribute category, customers whose policy covers the risks of ***CAI*** **(**Complementary accident insurance -consists of death, total, partial and permanent disability medical insurance) dominate the group by accounting 100%. Among **payments status** attribute categories customer who made half payment dominated the group accounting for 99%. Customers who purchased insurance policy for medium-term account 99% of the total.

In the second experiment, to improve the quality of the cluster (to increase the intra-similarity), yet the k value was 2 but the number of seeds was increased from 10 to 100. As a result, the sum squared error is increased from **16743.0 to 28686.0**. In addition to that, ***experiment#2*** disclosed

that ***clusters #1*** and clusters#2 share ***INSURED_VALUE, AGE, SEX PAYMENT_STATUS and duration*** attribute values**.** In other words, both cluster shares five attributes values. Comparing to the first result relatively higher SSE was registered. Given two clusters, the one which registered the smallest SSE is preferred. Hence, the quality of the cluster model decreases as the intra -cluster similarity decreases. Therefore, the researcher made the third experiment to increase the intra-cluster similarity and to reduce inter –class similarity of the clustering model.

Experiment #3

This experiment done by setting k value to 2 and the seed value to 200

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 200
Relation:     EIClifeinsurancedataset
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
Ignored:
              value-seg
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 16566.0

Initial starting points (random):

Cluster 0: '\'(0.115407-0.132349]\'','\'(28-41)\'',F,PNFE,half-paym,medium-term
Cluster 1: '\'(min value-0.115407]\'','\'(14-28]\'',F,DEATH,half-paym,medium-term

Missing values globally replaced with mean/mode
```

Figure 16 Run information of experiment#3

Table 17 shows all attributes and their categories with respect to their values.

Cluster index	INSURED_VALUE	AGE	SEX	COVER_TYPE	PAYMENT_STATUS	Duration
Cluster 1 11852 (5%)	min value-0.115407]'=3% 0.115407-0.132349=93% 0.132349-max value= 0%	0-14=0% 14-28=0% 28-41=99%	F= 98% M=1%	DEATH =32% SAI=0% DEATENDOW=0% PNFE=33% CAI =33% HIEYE =0% HIDENTAL=0% HIPREGN =0%	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=99% Long-term=0% Short-term =0%

				HIMAT=0% HIHOSPSICK=0 %		
Cluster 2 9770 (5%)	min value- 0.115407]'=99% 0.115407- 0.132349=0% 0.132349-max value= 0%	0-14=0% 14- 28=97% 28- 41=1%	F=88% M=11 %	DEATH =34% SAI=0% DEATENDOW=0 % PNFE=32% CAI =32% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0 %	half-paym=99% full-paym =0% intial-paym=0%	Medium- term=98% Long- term=0% Short-term =0%

Table 17 K value = 2 and seed value=200

Table 17 revealed the clustering model built under ***experiment#3***. The results show that, of 21622 instances (100%), 11852 instances (55% of the dataset) were assigned to the first cluster (cluster #1) and 9770 instances (45% of dataset) were assigned to the second cluster (cluster #2). The first cluster (cluster #1) is slightly greater in number of instances than the second cluster (cluster #2). Cluster#1 revealed that, policyholders who paid sum of assured value less than 67719 Birr registered 3% while policyholders who paid sum of assured value between 67719-77366.5 Birr registered 93% of the group. From the age attribute categories, customers' aged greater than 28 years dominated by accounting for 99%. Female customers accounted for 98% while male customers accounted for 1% in the cluster. From categories of ***COVER_ TYPE*** attribute, customers whose policies covers for DEATH registered 32%, PNFE (Pre- needed funeral expense insurance) registered 33% and ***CAI*** (Complementar***y*** Accident Insurance - Consists Of Death, Total, Partial And Permanent Disability Medical Insurance) registered 33% of the instances assigned to cluster#1.Among the categories of ***PAYMENT_STATUS*** attribute, customer who made half payment dominated the group for accounting 99%. Customers who purchased insurance policy for medium-term account for 99%.

Depending on the results of **Cluster #2**, among group of the insured value attribute, policyholders who paid sum of insured value less than 67719 Birr dominate the group by accounting for 99% of instances assigned to the cluster (cluster#2). Customers whose age is between 14-28 account for 99%. Female customer accounts for 88% while male customers consist of 11% of the group. From cover type attribute categories, DEATH consists of 34% Pre-needed funeral expense insurance (PNFE) consists of 32% and Complementary accident insurance – which consist of death, total, partial and permanent disability medical (CAI) accounts for 32%. Among **payments status** attribute categories, customers who made half payment dominated the group accounting for 99%. Customer who purchase insurance policy for medium-term accounts 98%.

During the third experiment (**Experiment #3**) the problem sharing same attribute value in the second attribute was solved by decreasing from five to three.

Experiment #4

This experiment is conducted by setting k=2 and seed value =300.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -V -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 300
Relation:     EIClifeinsurancedataset
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
Ignored:
              value-seg
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 16743.0

Initial starting points (random):

Cluster 0: '\'(min value-0.115407]\'','\'(28-41)\'',F,DEATH,half-paym,medium-term
Cluster 1: '\'(0.115407-0.132349]\'','\'(28-41)\'',F,DEATH,half-paym,medium-term

Missing values globally replaced with mean/mode
```

Figure 17 Run information of experiment#4

Table 18 presented the k means clustering results with standard deviation in respect to their attribute values after the parameter of the seed value is changed to 300.

Cluster	INSURED_VALUE	AGE	SEX	COVER_TYPE	PAYMENT_STATUS	Duration
Cluster 1 10169 (47%)	Min-value-0.115407]'=99% 0.115407-0.132349=0% 0.132349-max value= 0%	0-14=0% 14-28=93% 28-41=5%	F=88% M=11%	DEATH=33% SAI=0% DEATENDOW=0% PNFE=32% CAI =32% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0%	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=98% Long-term=0% Short-term =1%
Cluster 2 11453 (3%)	Min-value-0.115407]'=0% 0.115407-0.132349=99% 0.132349-max value= 0%	0-14=0% 14-28=0% 28-41=100%	F=98% M=1%	DEATH =33% SAI=0% DEATENDOW=0% PNFE=33% CAI =33% HIEYE =0% HIDENTAL=0% HIPREGN =0% HIMAT=0% HIHOSPSICK=0%	half-paym=99% full-paym =0% intial-paym=0%	Medium-term=99% Long-term=0% Short-term =0%

Table 18 K value = 2 and seed value=300

As it was revealed in **Table 18**, the first cluster, cluster #1, consists of 10169 instances which account for 47% of the data set. It entails that customer who paid from less than 67719 of the total sum of insured value, comprised of 99%. So it dominated the group .

From the age attribute categories, customers whose age lies between 14-28 consists of 93% instances and customers aged greater 28 account for 5%.

From sex group, female customers account for 88% while male customers account for 11%.

From cover type attribute category, customers whose cover type fall under ***DEATH*** consists of 33%, ***PNFE*** account for 32% and ***CAI*** consist 32% of the instances.

Among the categories of ***PAYMENT_STATUS*** attribute, customer who made half payment dominated the group for accounting 99%. Customer who purchased insurance policy for

medium-term account for 98% while customers who purchased insurance policy for short term account for 1 %.

The second cluster (**Cluster #2**) consists of 11453 instances which accounts for 53% of the EIC dataset. this cluster relatively it greater by the number instances it contains than the first cluster .Based on the result of **Cluster #2**, from the group of ***INSURED_VALUE*** attribute ,policyholders whose sum of assured value lies between 0.115407-0.132349 means that (67719-77366.5] consist of 99% of the group. From the categories age attribute customers whose age are greater 28 year accounts for 99%. From ***SEX*** attributes, female customer accounts for 98% while a male customer consists of 1 % of the group. From cover type attribute is categories, customers that their cover type fall under DEATH consists of 33%, PNFE (Pre- needed funeral expense insurance) each accounts for 33% and CAI (Complementary Accident Insurance - Consists of Death, Total, Partial and Permanent Disability Medical Insurance) consists 33% of the instance. Among **payments status** attribute categories customer who made half payment dominated the group accounting for 99%. Customer who purchased insurance policy for medium-term accounts 99%.

5.3 Evaluation and Model Selection of Clustering models

Previously four experiments were conducted to select cluster models are used or implemented to cluster the EIC life insurance dataset. To evaluate k-means clusters, the most common measurement criterion is Sum of Squared Error (SSE). During the computation for SSE, for each point, the error is the distance to the nearest cluster. In order to compute SSE, we square these errors and sum them. Therefore the result of SSE shows the relation of objects within the cluster. The equation of SSE is given as: [29]

$$\mathbf{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} \boldsymbol{dist}^{2}(\boldsymbol{m_i}, \boldsymbol{x})$$

- x is a data point in cluster C_i and m_i is the representative point for cluster C_i
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, (the number of clusters)

Equation 6 Sum of squared error computation

In addition to SSE, in this study the selection criteria of k means clusters models also considers other measurements such as time taken to build the model and number of iterations. Therefore the cluster model which shows minimum SSE, less time to build the model and minimum number of iterations was selected. But more weight (70%) is given to SSE and the rest 30 is divided equally for time to build the model and minimum number of iterations measurements

Among the four experiments conducted experiment#3 registered a better performance than the other experiments of k-means clustering models. The selection of the clustering model was undertaken in consultations with domain experts to make the segmentation of policyholders more meaningful and useful to the corporation activities. **Table 19** shows the summary of performance clustering model experiments

Experiment-No	Time taken to build the model	SSE	Number of iteration
Experiment#1	**0.61**	**16743.0**	**2**
Experiment#2	**0.11**	**28686.0**	**2**
Experiment#3	**0.08**	**16566.0**	**2**
Experiment#4	**0.11**	**16743.0**	**3**

Table 19 Experiments of Clustering Model

Based on the parameters, results and selection criteria ***Experiment#3*** was selected as the best cluster model to segment life insurance policyholders of Ethiopian Insurance Corporation (EIC). The description of the clustering model conducted under experiment#3 is described as follows.

I. Cluster#1 revealed that policy holders, most likely female, whose age are greater than 28 and who paid sum of insured value between 67719-77366.5 Birr and their policy covers for DEATH, CAI **(**Complementar***y*** Accident Insurance -Consists Of Death, Total, Partial And Permanent Disability Medical Insurance) and PNFE (Pre- needed funeral expense insurance prefers to pay the half of the premiums payment and tend to purchase policies from 6-10 years. Of 21622 instances of the dataset of, 11852 instances (55% of the dataset) were assigned to this cluster. This group of policyholder's contributes **more or high value** for the corporation.

II. Cluster #2 was made-up of youth and adult customers which female policy holder's accounts for (88%) and male policy holders of accounts 11% .The sum of insured value of this group is less than 67719 Birr. And their policy covers for DEATH 33% Pre-needed funeral expense and Complementary accident insurance for 34%, 33%and 33% respectively. The characteristics of customers having the above the criteria are most likely tended to purchase policies for medium-terms, and pay sum of insured value less than 67719 Birr and made the half payment of the premiums. Of 21622 instances, 9770 instances (45% of dataset) were assigned to cluster #2. This group of policyholder's contributes **less/low value** for the corporation.

The selected cluster model will be used as input for classification model building. Weka can recognize clustering data as input to predict the actual classes of the cluster data. In another words, classification model can be built from the output cluster models. The following steps present how to use the k-means clustering models data for decision tree classification.

Step 1. After running the cluster model in Weka, go to result list option and right click on the selected model or result.

Step 2. Then, select "Visualize cluster assignments" option

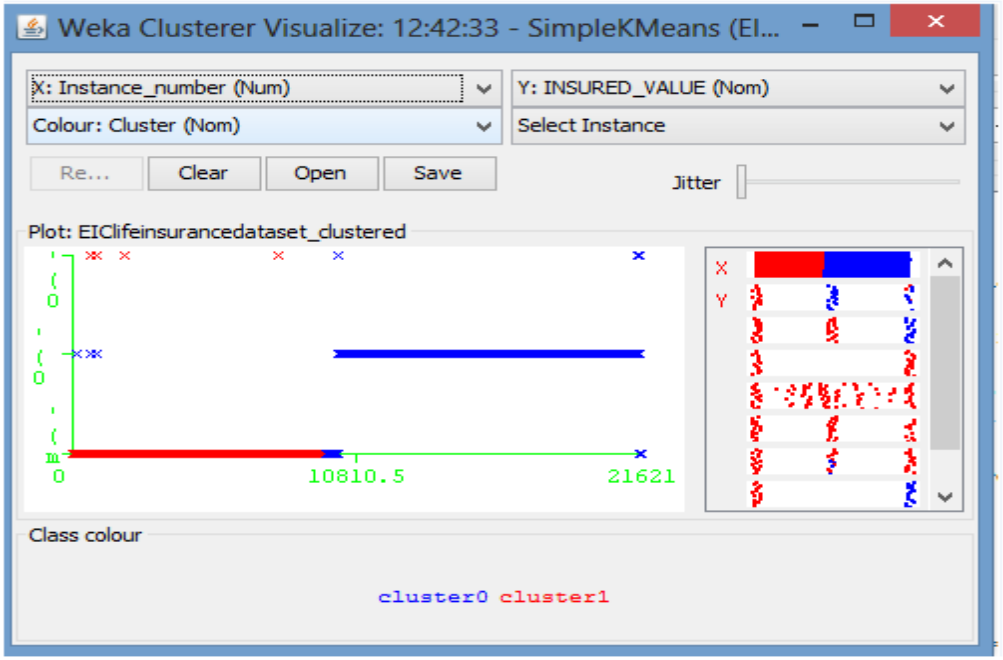


Figure 18 Visualizing and saving clusters in Weka

Step 3. Finally, select “Save” option to locate the visible instances to file in ARFF

In this study the result of k-means algorithm of clustering analysis were used to build decision tree J48 classification model predict class of customer lifetime value. Therefore, it will be used to categorize the actual class of CLV and forecast potential CLV of EIC life insurance customers.

5.4 DECISION TREE CLASSIFICATION MODEL

As stated earlier, the output of cluster model is used for building decision tree classification models. The reason of using decision trees as classifier model is to decipher the distinguishing or distinctive nature of each cluster into set of valuable rules that can be humanly understandable and easily to interpretable to business knowledge. These rules of the decision tree model can be used to classify and predict the new labels. Even though using decision tree for classifying cluster causes problems, but it is more transparent approach for cluster updating and it can jointly assess many attributes and reveal those which best characterize each cluster [9]. Moreover, rules generated by DT models can help insurance experts and business participants to get more insight and to realize the patterns of clusters model make necessary adjustments (fine -tune) to the results according to their business objective [9].

The decision tree model is the supervised classification model that was used to predict the actual class to distinguish whether the instances belongs to specified class. Particularly, the J48 algorithm was selected, which is WEKA’s implementation of the C4.5 decision tree learner. The algorithm uses a greedy technique to induce decision trees for classification and uses reduced-error pruning.

During the model building of decision tree model, test options such as percentage split and cross-validation were used. With percentage split, the data set is divided into a training set and a test set. For the training set 66% (14271) of the instances in the data set was used and for the testing set the remaining part (7351) was used. Cross-validation test option is selected because it can speed up the time to training process using random sampling from large data sets. Hence, it enables differential misclassification costs. The standard for this is 10-fold cross-validation method. The data is divided randomly into 10 parts in which the classes are represented approximately the same proportions as in the full dataset (stratification) [9]. Each part is held out

in turn and the algorithm is trained on the nine remaining parts; then its error rate is calculated on the holdout set. Finally, the 10 error estimates are averaged to yield an overall error estimates.

The following experiments will explain the selected parameter, the test option used and results of the experiments conducted in this study.

Experiment#5

During the first experiment the J48 algorithm was equipped with the default setting in **weka.gui.Genericobjecteditor** . The setting will be described as follows:

- The minimum number of instances per leaf minNumObj =2
- Seed -- The seed used for randomizing the data when reduced-error pruning is=1
- numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree. The default value this parameter is 3

The test option used for this is experiment cross validation. **Figure 18** presents the classifier run information of ***Experiment#5***

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     EIClifeinsurancedataset_clustered-weka.filters.unsupervised.attribute.Remove-R1,8
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
              Cluster
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
```

Figure 19 Run Information of Decision Tree Model of experiment#5

The tree model generated under this model is presented in figure 19.

```
J48 pruned tree
------------------

AGE = '(0-14]'
|   COVER_TYPE = DEATH: cluster1 (15.0)
|   COVER_TYPE = SAI: cluster1 (2.0)
|   COVER_TYPE = DEATENDOW: cluster1 (7.0)
|   COVER_TYPE = PNFE: cluster0 (13.0)
|   COVER_TYPE = CAI: cluster1 (13.0)
|   COVER_TYPE = HIEYE: cluster1 (4.0)
|   COVER_TYPE = HIDENTAL: cluster1 (4.0)
|   COVER_TYPE = HIPREGN: cluster1 (2.0)
|   COVER_TYPE = HIMAT: cluster1 (2.0)
|   COVER_TYPE = HIHOSPSICK: cluster1 (5.0)
AGE = '(14-28]'
|   INSURED_VALUE = '(min value-0.115407]': cluster1 (9519.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (4.0/1.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster1 (4.0)
AGE = '(28-41)'
|   INSURED_VALUE = '(min value-0.115407]'
|   |   COVER_TYPE = DEATH: cluster1 (192.0)
|   |   COVER_TYPE = SAI: cluster0 (0.0)
|   |   COVER_TYPE = DEATENDOW: cluster0 (0.0)
|   |   COVER_TYPE = PNFE: cluster0 (191.0)
|   |   COVER_TYPE = CAI: cluster0 (191.0)
|   |   COVER_TYPE = HIEYE: cluster0 (0.0)
|   |   COVER_TYPE = HIDENTAL: cluster0 (0.0)
|   |   COVER_TYPE = HIPREGN: cluster0 (0.0)
|   |   COVER_TYPE = HIMAT: cluster0 (0.0)
|   |   COVER_TYPE = HIHOSPSICK: cluster0 (1.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (11437.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster0 (16.0)

Number of Leaves  :     25

Size of the tree :      30
```

Figure 20 Decision Tree Model of experiment#5

The decision tree Classifier model is a pruned decision tree in graph form. As it was specified earlier, cross validation test option was used in experiment#5. The first split of the decision tree was displayed on ***"Age"*** attribute and secondly, it was revealed on ***Insured-Value*** attribute. In the tree structure, a colon represents the class label that has been assigned to a particular leaf, followed by the number of instances that reach that leaf. The above the tree structure has the number of leaves is 25 and the size of the nodes in tree is 30.

The result of J48 decision tree can be more understood when it is transformed to IF_THEN rule. The numbers in parentheses after the leaf nodes indicate the number of records assigned to that node while the numbers after slash (/) sign represents the numbers of incorrectly classified instances.

The following interpretation was mined from the **experiment#5**. First the J48 decision tree was converted to IF_THEN rule. Then, the rules and knowledge discovered from the classifier model was transformed into textual form. For instance, some of extracted rules are explained as follows.

Rule#1 *IF* AGE = '(0-14]' AND COVER_TYPE = DEATH *THEN* cluster1 (15.0)

This rule shows that policyholders whose age is between 0 to 14 and their insurance policy covers for death only are classified under cluster1. Fifteen instances fulfil these criteria and all of them are correctly classified

Rule#2 *IF* AGE = '(0-14]' AND COVER_TYPE = DEATENDOW *THEN* cluster1 (7.0)

Under this rule, seven instances are recorded to cluster1 class label to indicate customers whose insurance policy covers death and endowment life insurance policies and their age group is found between 0 and fourteen. The rule indicates that those life insurance policies are purchased most likely for children. The funds made to purchase Endowment life insurance policies are mostly accumulated to the end of a specified period. Such a fund could be used to purchase or supplement retirement pension to finance children's university education, to start a small business and a host of other purposes. In general, such type of life insurance is primarily required for saving and the insurance protection being only incidental. The experiment shows that all instances labelled under this class are classified correctly.

Rule#3 *IF* AGE = '(14-28]' AND INSURED_VALUE = '(min value-0.115407]' *THEN* cluster1 (9519.0)

Middle age policyholders of age between 14 -28 and the sum insurance value is less than 67719 Birr are labelled under cluster1 .Of 9519 instances assigned to this class label , all instances are correctly classified.

Rule#4 *IF* AGE = '(14-28]' AND INSURED_VALUE = '(0.115407-0.132349]' *THEN* cluster0 (4.0/1.0)

Insurers whose age is between 14-28 and the sum of insured value fall between 67719 to 77366.5 Birr were classified under clusters1. Of four instances assigned to this class, one instance incorrectly classified

Rule#5 *IF* AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = PNFE *THEN* cluster0 (191.0)

If the age of insured is greater than 28 and less than 65 and insured value of the customers below 67719 Birr and the risk covered by their policy implies for pre needed funeral expense, then class label fall under cluster0. Based on the above rule one hundred ninety one policy owners assigned this class and all of the instances are correctly classified.

Rule#6 *IF* AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = CAI *THEN* cluster0 (191.0)

This rules revealed that Policy holders whose age is greater than 28 years and the total insured value is below 67719 Birr and their policy covers for Comprehensive Accident Insurance (**CAI**) are classified under as cluster0 label**.** According to experts consulted; this rider covers death, disability or loss of time. Loss must however be effected directly and independently of all other causes through external, violent, and accidental means of which there is evidence of visible contusion or wound on the exterior of the body accident insurance. Of 191 instances assigned to this class label, all of them predicted correctly. The evaluation of classifier model is described in confusion matrix as follows.

Actual Class	PREDICTED CLASS		
	Class=cluster0	Class= cluster1	Total
Class=cluster0	A =(TP) 11849	B=(FP) 3	11852
Class=cluster1	C=(FN) 1	D=(TN) 9769	9770
Total	11850	9772	21622

Table 20 Confusion matrix of Experiment#5

Note*: Cluster0 represents customers whose value is classified as high while cluster1 represents for customers their value is classified as low. The current values of policyholders were represented in the actual class and their potential value is revealed in the predicted class of confusion matrix.*

The Decision tree presented by using confusion matrix (**Table 20**) represents the count of instances into four elements. A= (TP) shows the number of true positive instances correctly classified. B = (FP) shows the number of instances misclassified as negative, but belong to TP

class. C= (FN) shows the number of false positive instances incorrectly, but belong to true negative classes. And finally D= (TN) shows the number of true negative instances correctly classified to the class. Each element in the matrix is a count of instances. Rows represent the true classes and columns represent the predicted classes.

As a result, among 21622(100%) instances 21618(99.9815) instances were correctly classified and 4 instances were incorrectly classified which accounts for 0.0185% the result. 11849 instances, which are actual class of cluster0 instances were predicted as cluster0 . But three instances, which belong to cluster0 was predicted under cluster1. And one instance, which true class of cluster1 is predicted under class of cluster0. 9769 instances, which are true class of cluster1, were predicted under class of cluster0.

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP}{TP+FP} * 100$$

Equation 7 Accuracy Rate Computation

Accuracy rate of the classifier model is: $\frac{11849+9769}{11849+3+1+9769} = 21618/21622 = 99.9815\ \%$

Experiment#6

In the second experiment, percentage split test option is used to split the dataset of EIC into training and test data set. The default setting of percentage split was applied. Therefore, 66% of the dataset was used for training while the remaining was used for testing. The parameter for percentage split was:

- minNumObj -The minimum number of instances per leaf minNumObj =2
- Seed – default seed value is used = 1 the seed used for randomizing the data when reduced-error pruning is used.
- NumFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree.in this study default numFolds is 3

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     EIClifeinsurancedataset_clustered-weka.filters.unsupervised.attribute.Remove-R1,8
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
              Cluster
Test mode:    split 66.0% train, remainder test

=== Classifier model (full training set) ===
```

Figure 21 Run Information of Decision Tree J48 Of Experiment #6

The decision tree generated in percentage split (66% for training and the rest for testing) test option was the same as the DT generated in the cross validation test option.

```
J48 pruned tree
------------------

AGE = '(0-14]'
|   COVER_TYPE = DEATH: cluster1 (15.0)
|   COVER_TYPE = SAI: cluster1 (2.0)
|   COVER_TYPE = DEATENDOW: cluster1 (7.0)
|   COVER_TYPE = PNFE: cluster0 (13.0)
|   COVER_TYPE = CAI: cluster1 (13.0)
|   COVER_TYPE = HIEYE: cluster1 (4.0)
|   COVER_TYPE = HIDENTAL: cluster1 (4.0)
|   COVER_TYPE = HIPREGN: cluster1 (2.0)
|   COVER_TYPE = HIMAT: cluster1 (2.0)
|   COVER_TYPE = HIHOSPSICK: cluster1 (5.0)
AGE = '(14-28]'
|   INSURED_VALUE = '(min value-0.115407]': cluster1 (9519.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (4.0/1.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster1 (4.0)
AGE = '(28-41)'
|   INSURED_VALUE = '(min value-0.115407]'
|   |   COVER_TYPE = DEATH: cluster1 (192.0)
|   |   COVER_TYPE = SAI: cluster0 (0.0)
|   |   COVER_TYPE = DEATENDOW: cluster0 (0.0)
|   |   COVER_TYPE = PNFE: cluster0 (191.0)
|   |   COVER_TYPE = CAI: cluster0 (191.0)
|   |   COVER_TYPE = HIEYE: cluster0 (0.0)
|   |   COVER_TYPE = HIDENTAL: cluster0 (0.0)
|   |   COVER_TYPE = HIPREGN: cluster0 (0.0)
|   |   COVER_TYPE = HIMAT: cluster0 (0.0)
|   |   COVER_TYPE = HIHOSPSICK: cluster0 (1.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (11437.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster0 (16.0)

Number of Leaves  :     25

Size of the tree :      30
```

Figure 22 Pruned tree of J48 DT Model of experiment#6

The discovered rules under **Experiment#6** and their interpretations are presented below.

Rule #7 IF AGE = '(0-14]'AND COVER_TYPE = PNFE THEN cluster0 (13.0)

Lower age customers whose life insurance policy covers for Pre- needed funeral expense are labelled under cluster0. Based on the information thirteen instances assigned to cluster1. In addition all instances that satisfy this condition are correctly classified.

Rule#8 IF AGE = '(0-14]' COVER_TYPE = CAI THEN cluster1 (13.0)

Policy holders whose age are below fourteen years old and their policy covers for Complementary or Comprehensive accident insurance are classified under as cluster1 label. Risks covered by such insurance policy covers are death, total or, partial or permanent disabilities. Thirteen instances are classified under this rule. All instances are classified correctly

Rule#9 IF AGE = '(0-14]' COVER_TYPE = HIPREGN THEN cluster1 (2.0)

Two instances labelled under to cluster1 class are extracted from rule that implicate customers whose age are below fourteen and their policy covers Pregnancy check-up

Rule#10 IF AGE = '(14-28]' AND INSURED_VALUE = '(0.132349-inf)' THEN cluster1 (4.0)

Insurers whose age category found between 14-28 and the insured value are greater than 77366.5 Birr are then labelled as cluster1class. The four instances assigned to this class label are correctly predicted as instances class

Ru1e#11 IF AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = DEATH THEN cluster1 (192.0)

One hundred ninety two instances are classified as class of the second cluster, clsuster1. The rule discovered indicates that if policyholders age is greater 28 but the not exceed the maximum age of 65 that EIC insurance regulation obliges and the sum insurers policy value less than 67719 Birr and the risk covered by their policy is death alone (life insurance only) then instances satisfy this condition are assigned under cluster1

The time taken to build the model was 0.3 seconds. And also, time taken to test model on training split is 0.03 seconds. During this experiment the tree size and the number of the leaves were the same as with the first experiment (**Experiment#5)**. The evaluation DT model of constructed by using percentage split test option is described in confusion matrix as follows.

	PREDICTED CLASS		
Actual Class	Class=cluster0 one	Class= cluster1	Total
Class=cluster0	A =(TP) 4007	B=(FP) 2	4009
Class=cluster1	C=(FN) 0	D=(TN) 3342	3342
Total	4007	3344	351

Table 21 Confusion Matrix of Experiment#6

Confusion matrix of **Experiment#6** revealed that, among 7351 (100%) instances which were used for testing the model, 7349 (99.9728%) instances were correctly classified and 2 instances (0.0272 %) instances is incorrectly classified .All (4007) instances, which are true classes of cluster0, are predicted under the same class (cluster0). 3342 instances (true negative), which are actual class of cluer1 are predicted under the same class (cluster1). And two instances, which belong to cluster0, are predicted under cluster1 class label.

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP}{TP+FP} * 100 = 7349/7351 = 99.9728\%$$

Experiment #7

In this experiment the same test option, percentage split(66% for training and for testing the dataset , was used . But the parameter of the J48 DT was changed by setting minNumObj =5. Therefore the result of this model is described as follows.

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 5
Relation:     EIClifeinsurancedataset_clustered-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R7
Instances:    21622
Attributes:   7
              INSURED_VALUE
              AGE
              SEX
              COVER_TYPE
              PAYMENT_STATUS
              duration
              Cluster
Test mode:    split 66.0% train, remainder test
=== Classifier model (full training set) ===
J48 pruned tree
------------------
AGE = '(0-14]'
|   COVER_TYPE = DEATH: cluster1 (15.0)
|   COVER_TYPE = SAI: cluster1 (2.0)
|   COVER_TYPE = DEATENDOW: cluster1 (7.0)
|   COVER_TYPE = PNFE: cluster0 (13.0)
|   COVER_TYPE = CAI: cluster1 (13.0)
|   COVER_TYPE = HIEYE: cluster1 (4.0)
|   COVER_TYPE = HIDENTAL: cluster1 (4.0)
|   COVER_TYPE = HIPREGN: cluster1 (2.0)
|   COVER_TYPE = HIMAT: cluster1 (2.0)
|   COVER_TYPE = HIHOSPSICK: cluster1 (5.0)
AGE = '(14-28]': cluster1 (9527.0/3.0)
AGE = '(28-41]'
|   INSURED_VALUE = '(min value-0.115407]'
|   |   COVER_TYPE = DEATH: cluster1 (192.0)
|   |   COVER_TYPE = SAI: cluster0 (0.0)
|   |   COVER_TYPE = DEATENDOW: cluster0 (0.0)
|   |   COVER_TYPE = PNFE: cluster0 (191.0)
|   |   COVER_TYPE = CAI: cluster0 (191.0)
|   |   COVER_TYPE = HIEYE: cluster0 (0.0)
|   |   COVER_TYPE = HIDENTAL: cluster0 (0.0)
|   |   COVER_TYPE = HIPREGN: cluster0 (0.0)
|   |   COVER_TYPE = HIMAT: cluster0 (0.0)
|   |   COVER_TYPE = HIHOSPSICK: cluster0 (1.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (11437.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster0 (16.0)

Number of Leaves  : 23
Size of the tree :    27
```

Figure 23 Pruned tree of DT Model of experiment#7

Under this experiment only one new rule was generated. The new rule discovered is presented as follows.

Ru1e#12 IF AGE = '(14-28]': THEN cluster1 (9527.0/3.0)

Policyholders, whose age are between 14-28 and the sum of insured value fall between (67719 to 77366.5 Birr are classified under clusters1. Of 9527 instance assigned to this class, three of them are incorrectly classified instance incorrectly classified. After the parameters modification, the size of trees and the number of leaves were declined to 23 and 27 respectively. But the confusion matrix of experiment#7 showed the same results observed in experiment#6.

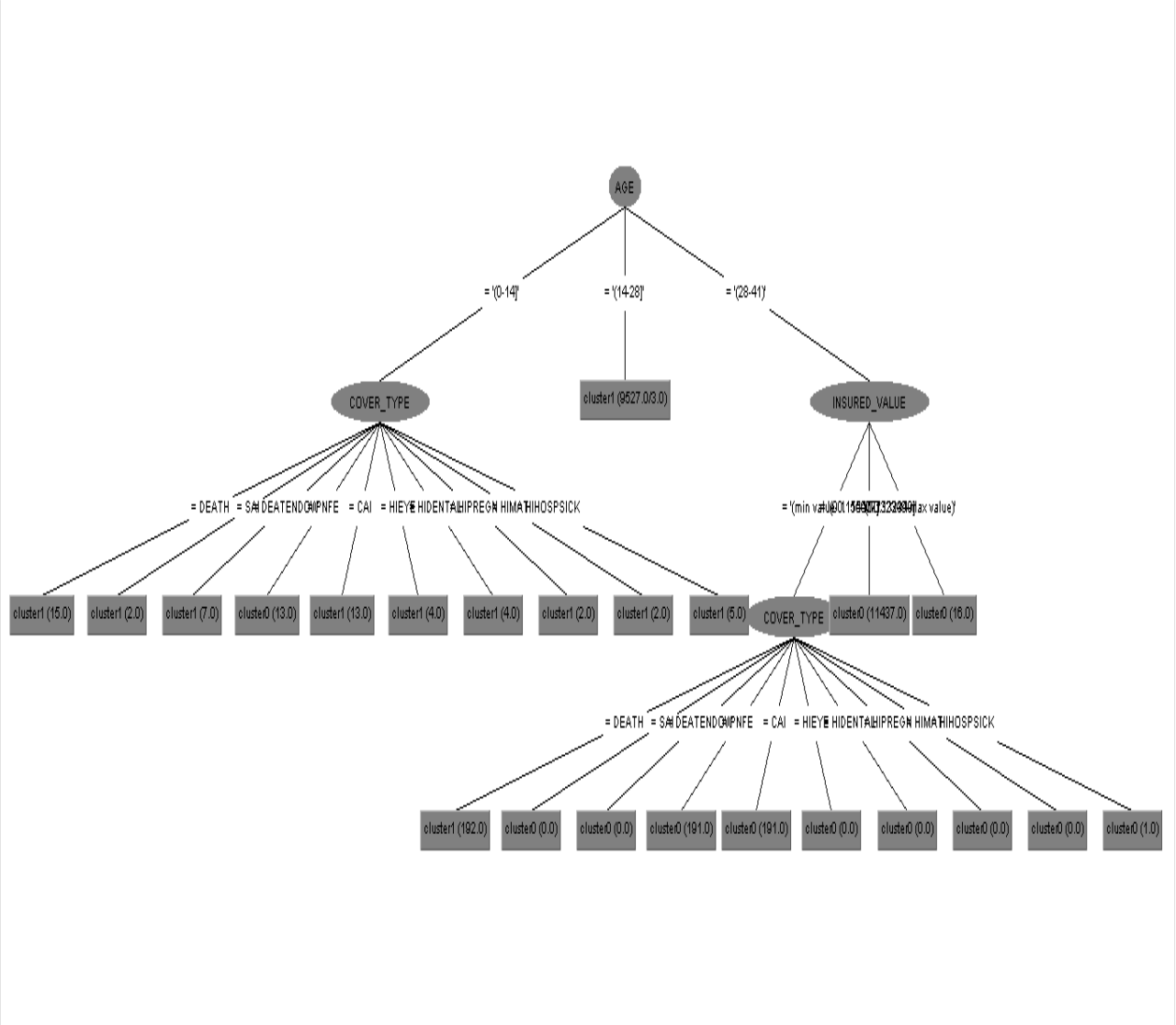


Figure 24 Decision tree of Experiment #7

Experiment #8

During this experiment the minimum number of objects per leaf (minNumObj) is changed from 10 to 15. And the test result is described below.

```
=== Run information ===

Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 15
Relation:   EIClifeinsurancedataset_clustered-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R7
Instances:  21622
Attributes: 7
            INSURED_VALUE
            AGE
            SEX
            COVER_TYPE
            PAYMENT_STATUS
            duration
            Cluster
Test mode:  split 66.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

AGE = '(0-14]': cluster1 (67.0/13.0)
AGE = '(14-28]': cluster1 (9527.0/3.0)
AGE = '(28-41]'
|   INSURED_VALUE = '(min value-0.115407]'
|   |   COVER_TYPE = DEATH: cluster1 (192.0)
|   |   COVER_TYPE = SAI: cluster0 (0.0)
|   |   COVER_TYPE = DEATENDOW: cluster0 (0.0)
|   |   COVER_TYPE = PNFE: cluster0 (191.0)
|   |   COVER_TYPE = CAI: cluster0 (191.0)
|   |   COVER_TYPE = HIEYE: cluster0 (0.0)
|   |   COVER_TYPE = HIDENTAL: cluster0 (0.0)
|   |   COVER_TYPE = HIPREGN: cluster0 (0.0)
|   |   COVER_TYPE = HIMAT: cluster0 (0.0)
|   |   COVER_TYPE = HIHOSPSICK: cluster0 (1.0)
|   INSURED_VALUE = '(0.115407-0.132349]': cluster0 (11437.0)
|   INSURED_VALUE = '(0.132349-max value)': cluster0 (16.0)

Number of Leaves  : 14

Size of the tree :      17
```

Figure 25 J48 Pruned tree of DT Model of experiment#8

During experiment#8 new rules were discovered. The explanation of knowledge discovered is presented as follows,

Ru1e#13 IF AGE = '(0-14]' THEN cluster1 (67.0/13.0)

If the age of the insured is between 0 and fourteen (children) are classified under to the second cluster1, cluster1. 67 instances were assigned to this class label. Among them, 13 instances are misclassified. Table 22 presents the confusion matrix for experiment#8.

Actual Class	PREDICTED CLASS		
	Class=cluster0	Class= cluster1	Total
Class=cluster0	A =(TP) 4003	B=(FP) 6	4009
Class=cluster1	C=(FN) 0	D=(TN) 3342	3342
Total	4003	3348	7347

Table 22 Confusion Matrix of Experiment#8

From the confusion matrix, it was observed that the model classified 7345 (99.9184 %) instances correctly and 6 (0.0816%) instances incorrectly classified. As a result, all instances categorized under class of cluster0 are predicted under the same class (cluster0). 3342 instances (true negative), which are actual class of cluer1 are predicted under the same class (cluster1). And 6 instances, which belong to cluster0, are predicted under cluster1 class label.

The result of the model under **Figure 25** revealed that the tree size of the model was reduced to 17 and the number of leaves was declined to 14.

5.5 Evaluation of the Decision Tree Models

The performance of the decision tree or classifier models are measured based on the accuracy rate revealed in the correctly classified instances. It can also be measured by the error rate revealed in the incorrectly classified instances. Hence, different parameters were used in order to select the best classifier model that showed high prediction accuracy. The following table summarizes the performance of the models with respect to the test Option selected, the minimum number of instances per leaf, the number of pruned trees revealed and the time taken to build the models.

Test Option	Minnumobj	No. Of Trees	Accuracy rate	Time To Build The Model
Cross validation	**2**	**30**	**99.9815 %**	**0.28 seconds**
Percentage split	**2**	**30**	**99.9728 %**	**0.3 seconds**
Percentage split	**5**	**27**	**99.9728 %**	**0.05 seconds**
Percentage split	**15**	**17**	**99.9184 %**	**0.02 Seconds**

Table 23 Evaluations of Decisions Tree Classifier Models

Table **23** presented the results of the decision tree models to evaluate and select the best model. Accordingly, among four experiments of decision tree models, the first experiment, configured with cross validation test option to showed higher accuracy rate (**99.9815 %**) to predict the customer value. And experts' were more interested in rules generated under this experiment results.

5.6 Evaluation of Domain Experts

The subjective evaluation of domain experts is very important to support the decision making in marketing function of the EIC's life insurance division. Moreover, the significance of cluster models results requires the knowledge of the domain experts to transform to business knowledge. The result of the experiments undertaken to build descriptive and predictive models were discussed with the domain experts found in EIC life insurance in order to approve whether it add business intelligence for the business practices in marketing . Besides, the descriptive model developed by clustering algorithm models (k-means) and predictive decision tree classification algorithm (J48) were undertaken by consulting the domain experts. Therefore, the parameters of the k-means and J48 algorithms were adjusted to their optimal values in order to discover best solution for the problem identified (customer segmentation based on their value).

5.6.1 Descriptive Models

Depending on the experts' consultation, the clustering models were built in four experiments by fine-tuning the parameters of the k-means clustering algorithm. The models built by k- means t algorithms were measured by to SSE, time taken to build the model and number of iterations. the experts were more interested by the result and model generated under , *Experiment#3,* which was selected as the best cluster model; hence it looks more meaningful to the business practices of the

corporation. Distinguishing attributes values are revealed age and insured value attributes. The subjective opinion domain experts undertaken were based on the current trends of the business.

Regarding patterns revealed in the cluster model, domain experts said that most policy are purchased by older age customers, which tend to stay longer time than lower age policy holders and the insured value of such customers is higher than the other age group. The reason that they contribute higher value is connected with the insured age and insured value. As a result, policy a holder whose age is greater than 28 and insured value is above average contributes high value for the corporation. The cluster model revealed that female policy holders' purchase more life insurance policy than the male. Experts explained that, it could be happened because of the premium payment is discounted for female policy holders. Hence, the marketing department would focus to attract male customers. And also, the experts confirmed that customers, mostly likely tend to purchase policies, which cover the risks of DEATH, PNFE (Pre-Needed Funeral Expense) and CAI (Comprehensive Accident Insurance) as life insurance fundamentally protects against losses caused by death, disabilities and old-age.

5.6.2 Predictive Models

The classification techniques works on predefined groups or segments. in this study, the predictive models was built by using the clustering results. depending on the business problems, life insurance policy holders of EIC segmented into homogeneous groups based on the their value. The J48 classification algorithm was used to define the characteristics of each cluster along with their belonging classes. To be specific, the study revealed that the customer value is affected by the policy holders insured value (sum assured) and age. Although, the classification models are evaluated using accuracy rate and time to build the model, the domain experts subjective evaluations is necessary to check whether the patterns revealed might add a value to their business domain. Most of the rules discovered are discussed under the experiments of DT.

Domain experts are satisfied with the patterns and rules revealed in the experiments. The interpretation made by the domain experts considered the business of the corporation. The experiments conducted with the default setting of the parameters and run by percentage split and cross validation test option were more attractive to domain experts. Experts described that life insurance policies are mainly designed based on different plans (term, endowment and whole life

insurances), that are designed to cover only economic losses caused by death, disability and old age. There are also supplementary contracts. As a result, the risk covered by the policies are also differs based on the cover type requested by policy holders. In addition, the amount of premium paid by the policy holders depending up on the cover type. In the experiment, most policy holders whose age is greater than 28 years and who paid insured value above the average contributes high value. On the contrary, majority of policy holders whose age is less than 28 years and who paid insured value below the average contributes low value. More importantly, experts assured that age is the primary consideration as far as premium is concerned. In this study context, the sum of the actual made by premium payment of policy holders are reflected in the insured value of attributes. Hence, the subjective judgments of the experts are alike to the revealed results of the DT models, which discovered that age and insured-value attributes as most important attributes that help to segment customers based on their value the attributes.

In general, the data mining techniques could be used to segment life insurance customers' of EIC based on their value. Additionally, the marketing department of EIC could gain benefit by applying the DM techniques to make effective decision in CRM.

CHAPTER SIX

SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1 SUMMARY

The findings of the study revealed the significances of data mining technologies for customer segmentation in EIC. EIC is one of the largest insurance providers in Ethiopia, so it has large number of policyholders registered for non-life and life insurances, and the customers' information has been stored in the database of the corporation called INSIS (divided in to GENERAL AND LIFE INSIS). In order to assess their profitability and gain competitive advantage, the corporation has to identify its valuable customers. Thus, the corporation needs to implement appropriate customer segmentation approaches to adjust or modify its CRM strategies in order to prevent loss of its customers and revenue.

In conducting the study, the researcher attempted to build, select and evaluate clustering and classification models in order to apply data mining techniques and algorithms to segment the customers of life insurance in EIC at Life Addis district. It was identified that the data mining technology enables to segment customers' with similar characteristics and extract useful patterns that indicate most valuable customers.

As it was indicated in chapter four, the research was conducted by accompanying policyholders' dataset with their Lifetime Value. As result, the current values of each customer were calculated and labelled according to the profits that they contributed to the company over time. Therefore, the researcher examined the potential of data mining techniques to segment customers' value into low and high values and to reveal important patterns.

K-means clustering algorithm was applied to reveal underlying segment (into k numbers of clusters) and to analyse related characteristics that policy holders exhibits accordingly. Then, the developed models k- means clustering models were used to identify groups of life insurance policy holders who contribute high or low value to the corporation. Four experiments were made

to segmenting customers of life insurance into 2 clusters (by setting k value to 2) through changing the numbers of seed value to increase intra-cluster similarity and reduce the inter-cluster similarity of clustering model. During the experiment each clustering models, attribute called value-seg, which represents customers' value as attuned as dependent variable. And later, the attribute was used to understand the value of customers' revealed in each cluster of the selected model. The evaluation of clustering models considered criteria such as SSE (Sum Squared Error), number of iterations and time taken to build the model and more weight was given to SSE measurement. Among the models developed through different test parameters **Experiment#3** registered high performances because it indicates the lowest SSE, minimum time and minimum number of iterations.

The output of the chosen clustering model was used as an input for the decision tree classification model because the similar characteristics of customers revealed in each cluster can be used to classify and predict customers according to their value. DT models supports to realize and to label the segments based on the common characteristics of the members. The DT tree J48 algorithm was used to build customer segmentation model that classify and predict customers of life insurance based on their value. Among four models of decision tree, high accuracy (99.9815 %) was achieved in the first experiment with cross validation the test option.

The results of the study revealed that the performances of decision tree J8 algorithm models were decreasing, when the minimum number of object per leaf parameter is changed from 2 to 5 and from 5- to 15 parameters. In the same way, there were reductions in the number of trees indicated in the models. On the contrary, high accuracy was registered in decision tree models when the parameters of the experiments tuned with default values.

In general, the results of the research pointed out that the customer segmentation models built by using classification and clustering data mining techniques are necessary for the LAD and marketing department of EIC in order to identify the valuable segments of customers and other factors underlying variations of the customers' values.

6.2 CONCLUSION

In this study an attempt is made to reveal the high potential of data mining applications for customer segmentation, referring to the optimal usage of data mining methods and techniques to thoroughly analyse the collected historical data and to segments life insurance customers of EIC based on their value. The study was undertaken through five main stages called business understanding, data understanding, data preparation, modelling and evaluation.

During the business understanding phase, to get insight with the problem domain and formulate data mining goal, interviews were conducted with the domain experts, relevant documents were analysed. In the second phase, with consultation of experts 12 month year historical data consists of 16 attributes and 27845 records in 4 tables were collected from LIFE INSIS data base of EIC life insurance at LAD. The customers' value was computed using the CLV model. Again with domain experts 7 attributes and 21622 records were selected for final datasets.

Through modelling phase, the combination of clustering and classification data mining techniques were conducted to build customer segmentation model based on customers value computation. To achieve the objective of the study, K- means clustering algorithm is applied to identify the characteristics of life insurance policy holders who contribute high or low value to the corporation. Different user-defined parameters were applied to achieve good clustering model.

According to the revealed results, majority (55%) of the customers were clustered under high value. In this group, most of policy holders are female, whose age is greater than 28 and who paid sum insured value above the average. In addition, cover types such as DEATH, CAI and PNFE was mostly purchased by female customers. In the other hand, the youth and adult policy holders who paid sum insured value less than average contribute low value for the firm.

Based on the results of clustering model, J48 algorithm of decision tree was implemented to predict the potential value of customers and to identify the most significant variables that could help to segment life insurance customers based on their values. Accordingly, the age and insured value attributes revealed the high tendency to group customers and to split the decision tree

models. The classifier results of decision tree models shows that policyholders most likely purchased insurance for the coverage of CAI and PNFE insurances contributed high value for the company. During the study, decision tree models accuracy rate of the decision tree models shows slight differences when run information were set with different parameter and tests option. High performance (99.9815 %,) is registered when cross validation test option and default parameter is set in the run information.

Even though it is difficult to generalize depending on the revealed results, it was identified that insured-value and age attributes revealed encouraging patterns and can also be considered as the most important attributes that can support to segment customers of life insurance of EIC based on their value the attributes.

This study revealed that a good customer segmentation model can be built by combining K-means clustering and J48 classification algorithms. Besides, J48 decision tree algorithm showed good quality to visualize the clusters and to understand the clusters that lead to profile customers. The decision tree models generated interesting and useful rules that satisfy the business rule of the study subject. Some of them are presented below.

IF AGE ='(28-41)' AND INSURED_VALUE = '(0.115407-0.132349] THEN cluster0 (11437.0)

This rule indicates that policyholders age is greater 28 but the not exceed the maximum age of 65 that EIC insurance regulation obliges and the sum insured value is between 67719 to 77366.5 are labelled as cluster0. 11437.0 instances are satisfying this condition. And all of the instances are correctly classified. The domain experts' opinion and the current trends of the business shows, most of older age customers from different organization are registered for life insurance policies and they have a tendency to stay longer than lower and adult customers' age customers. Thus, experts confirmed that those customers most likely provide high value for the corporation

IF AGE = '(0-14]' AND COVER_TYPE = DEATENDOW THEN cluster1 (7.0)

Lower age policyholders, whose insurance covers for death and endowment policy are characterized under this rule. The funds made purchase Endowment life insurance policies are mostly accumulated to the end of a specified period. Such a fund could be used to purchase or supplement retirement pension to finance children's university education, to start a small business and a host of other purposes .In the meantime, the policy afford cover for dependents. In general such type of life insurance primarily required for saving and the insurance protection

being only incidental. The experiment shows that all instances labelled under this class are correctly classified. According to the explanation of domain experts, the profit generated from lower age customers are not significant due to the minimum premium paid for this cover type.

IF AGE = '(14-28]' AND INSURED_VALUE = '(min value-0.115407]' THEN cluster1 (9519.0)

This rule shows that , Middle age policyholders of age between 14 -28 and the sum insurance value is less than 67719 Birr are labelled under cluster1 .Of 9519 instances assigned to this class label , all instances are correctly classified. According to domain experts' consultation, the revenue generated from middle age policy holders is lower in relation to the older age customers

Generally, the results of the study revealed that the said data mining techniques are relevant to develop customers' segments based on their value. Besides, the outcome of this study indicate that data mining algorithms have powerful ability to build value-based customer segmentation and to find out useful patterns that were unknown previously. Therefore, marketers can be beneficiaries of this technology to make effective customer relationship management strategy.

Moreover, it was identified that the technology is useful for business experts to visualize the patterns created by DM algorithm and make substantial revenue for the corporation by incorporating the model with their business concern to attain a better business advantage. Furthermore, the study revealed that the policy makers of the company can use the technology for better decision making and to make adjustments in their strategic plan and CRM such as modifying their customer retention strategies. In doing so, the corporation marketing subdivisions can improve their policy offerings to generate more profits from customers who are classified as high value and have higher tendency to build long-term relationships with the corporation.

6.3 RECOMMENDATIONS

Based on the findings of the study, the researcher would like to suggest some important issues which require further investigations.

- The research was aimed at computing customer's lifetime value for individual policy holders; hence, the researcher considered only individual life policy holders for the data-sets construction and segmentation process. Therefore, further researches should be conducted in the area of group life insurance using appropriate value measurement model.
- During the computation of CLV, the sum of assured value was considered to calculate only the current value of the customer; as a result, further investigation needs to be conducted by considering premium payments made by the policy owners.
- In this study, the classification models were built based the result of the clustering algorithm. The performance clustering algorithm may affect the model built by the decision tree models; thus, further studies needs to be conducted by choosing the initial cluster centroids to other points to reveal remarkable patterns and hidden knowledge in the data.
- In addition, the researcher suggested further studies to take an account of more attributes to increase the quality of the segments; therefore, it can lessen the amount and results of instances cases, which were misclassified or unclassified.
- However this study aimed at building descriptive and predictive models using customer value, the time frame is limited due to the available data. Further research could use longer time periods historical for the datasets construction.
- The attributes such as SECTOR (business type), OCCUPATION-ID (occupation type) and MARITAL-STATUS of the initial datasets were excluded from the final data because they had registered high Missing Values (>74%). Further investigations may discover new rules and find new patterns by including these attributes into new data sets and by incorporating them with cover type and insurance type attributes.

Reference

[1] H. Aeron, A. Kumar and J. Moorthy "*Data mining frameowrk for customer lifetme value based segmentation".* Journal of database marketing and customer stsrategy management 2012,

[2] H.Woubishet. *The Application Of Data Mining To Support Customer Reationship Management At Ethiopian Airlines.* Addis Ababa University. 2002

[3] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker and S. Sriram "Modeling Customer Lifetime-Value," *Journal of Service Research*, Volume 9, No. 2, November ;2006

[4] K. V. Rodpysh,, A. Aghai and M. Majdi ," Applying data mining techniques for Customer Relationship management", *International Journal of Information Technology, Control and Automation* Vol.2, No.3, July 2012

[5] K. Umamaheswari and S. Janakiraman "Role of Data mining in Insurance Industry", international journal of advanced computer technology, Vol.3 Issue-6, June-2014

[6] X. Du "*Data Mining And Modeling For Marketing Based Attributes Of Customer Relationship*", Vaxjo University, Sweden Msi Report , 2006

[7] T. Hintsay. *Predictive Modeling Using Techniques In Suport Of Insurance Risk Assessement.* Addis Ababa University . 2002.

[8] M. Fikre "Predictive Data Mining Technique In The Case of Ethiopian Insurance Corporation Addis Ababa University, 2005.

[9] K. Tsiptsis and A. Chorianopoulos, "*Data Mining Techniques in CRM: Inside Customer Segmentation*", John Wiley and Sons, Ltd., Publication, 2009.

[10] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinart, C. Shearer, and R. Wirth, CRISP–DM *Step-by-Step Data Mining Guide*, 2000. http://www.crisp-dm.org.

[11] Two Crows Corporation: Introduction *To Data Mining And Knowledge Discovery, Third Edition: 1999*

[12] J.A. Berry and S. Gordon Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. Second Edition, Wiley publishing, 2004.

[13] C. Rygielski, J. C. Wang and D. C. Yen, "Data Mining Techniques for Customer Relationship Management," Technology in Society, Vol. 24, No. 4, 2002, pp. 483-502.

[14] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. San Francisco . Morgan Kaufmann, Second Edition , 2006.

[15] K. J. Cios, W. Pedrycz, W. Swiniarski and L. A. Kurgan *Data Mining:A Knowledge Discovery Approach* New York : Springer Science+Business Media, LLC, 2007.

[16] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, From Data Mining To Knowledge Discovery: An Overview_*Advances In Knowledge Discovery And Data Mining*. AAAI Press / The MIT Press. 1996.

[17] D. L. Olson and D. Delen , *Advanced Data MiningTechniques*, ,Berlin Heidelberg: Springer-Verlag ,2008

[18] S. O. Danso "*An Exploration Of Classification Prediction Techniques In Data Mining:The Insurance Domain"* MSC .Dissertation, Bournemouth University.September, 2006

[19] L. Guo. "Applying Data mining Techniques in Property/Casualty Insurance", Casualty Actuarial Society Forum Casualty Actuarial Society - Arlington, Virginia, 2003.

[20] N. Sharma , A. Bajpai , R. Litoriya "Comparison the various clustering algorithms of Weka tools." *International Journal of Emerging Technology and Advanced Engineering* ,Vol. 2, pp.73- 80 , May 2012

[21] S-Y. Kim, T-S .Jung, E-H. Suh and H-S. Hwang "A Model For Evaluating The Effectiveness Of CRM Using The Balanced Scorecard." Journal Of Interactive Marketing 17 (2), 5–19, 2003

[22] T-S .Jung, E-H. Suh and H-S. Hwang. "An LTV Model And Customer Segmentation Based On Customer Value: A Case Study On The Wireless Telecommunication Industry", Expert Systems With Applications , (2004) 26 (2), 181–188.

[23] A. B. Devale and R. V. Kulkarni, "Application of data mining techniques in life insurance", International journal of Data mining and Knowledge management Process", Vol.2, No.12, pp. 31-40. July 2012

[24] E. Wodajo "Data mining application for combating corruption using corrupt activity data in federal ethics and anticorruption commission of Ethiopia, Addis Ababa University, 2012

[25] B. Reganie "Application Of Data Mining For Customer Segmentation: The Case of Buusaa Gonofa Microfinance Institution". Addis Ababa University, 2013.

[26] D. Mamo "*Application of Data Mining Technology To Support Fraud Protection: The Case Of Ethiopian Revenue And Custom Authority*" MSc thesis in Information Science, Addis Ababa University: Addis Ababa ,January 2013

[27] Y. Zhao and Y.Zhang "*Comparison of decision tree methods for finding active objects"* National Astronomical Observatories, Bejing, China. Kaufmann Publishers; 2007 pp: 278-315

[28] S-Y. Kim, T-S .Jung, E-H. Suh and H-S. Hwang *"Customer segmentation and strategy development based on customer lifetime value: A case study*." Expert Systems with Applications, vol 31, pp.101–107, 2006

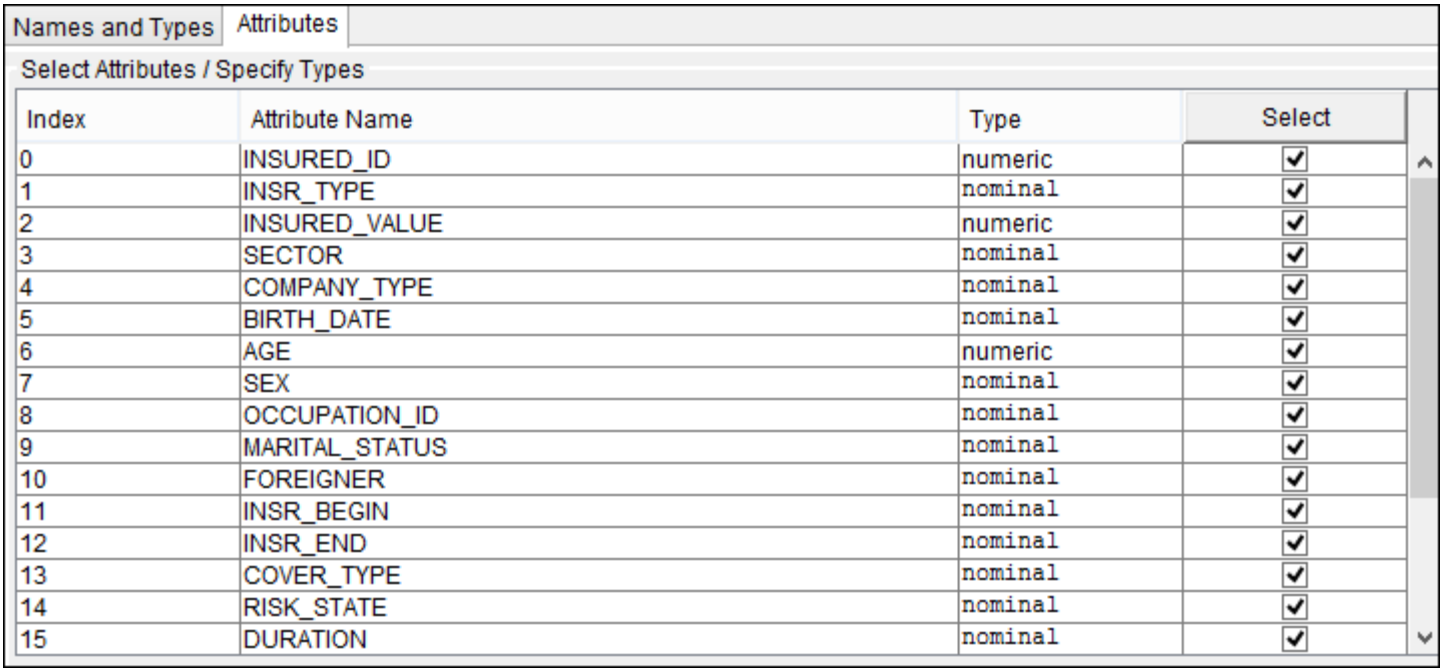
[29] P. Tan, M. Steinbach and V. Kumar .Introduction to Data Mining. Class lecture "Data Mining Cluster Analysis: Basic Concepts and Algorithms", 2004

[30] Y. Fu "Data Mining Tasks Techniques and Applications", University Of Missouri Rolla, 2005

[31] Y. K. Singh *Fundamental of Research Methodology and Statistics,* New Age International(P) Ltd, New Delhi ,2006

[32] Ethiopian Insurance Corporation "Insurance of Persons", EIC, 2012.

Appendix 1 List of attributes selected for initial dataset with DataPreparator 1.7 tool



Names and Types | Attributes

Select Attributes / Specify Types

Index	Attribute Name	Type	Select
0	INSURED_ID	numeric	☑
1	INSR_TYPE	nominal	☑
2	INSURED_VALUE	numeric	☑
3	SECTOR	nominal	☑
4	COMPANY_TYPE	nominal	☑
5	BIRTH_DATE	nominal	☑
6	AGE	numeric	☑
7	SEX	nominal	☑
8	OCCUPATION_ID	nominal	☑
9	MARITAL_STATUS	nominal	☑
10	FOREIGNER	nominal	☑
11	INSR_BEGIN	nominal	☑
12	INSR_END	nominal	☑
13	COVER_TYPE	nominal	☑
14	RISK_STATE	nominal	☑
15	DURATION	nominal	☑

Appendix 2 Rules generated from the decision tree models.

Generated Rules	Discovered knowledge	Interpretation
Rule#1	**IF** AGE = '(0-14]' AND COVER_TYPE = DEATH **THEN** cluster1 (15.0)	This rule shows that policyholders whose age is between 0 to 14 and their insurance policy covers for death only are classified under cluster1. Fifteen instances that fulfils this rule criteria and all of them are correctly classified
Rule#2	**IF** AGE = '(0-14]' AND COVER_TYPE = SAI **THEN** cluster1 (2.0)	Customers whose age category fall between zero and fourteen and their insurance cover type is fall under (SAI) Supplementary accident insurance – death and permanent disability are classified under cluster1. Two instances are assigned to this class. And also all records classified under this class are correctly classified.
Rule#3	**IF** AGE = '(0-14]' AND COVER_TYPE = DEATENDOW **THEN** cluster1 (7.0)	Under this rule Seven instances are mapped to cluster1 class label to indicate customers whose insurance policy covers death and endowment life insurance policies and their age group is found between 0 and fourteen. The rule indicates that those life insurance policies are purchased most likely children. The funds made purchase Endowment life insurance policies are mostly accumulated to the end of a specified period. Such a fund could be used to purchase or supplement retirement pension to finance children's university education, to start a small business and a host of other purposes .In the meantime, the policy afford cover for dependents. In general such type of life insurance primarily required for saving and the insurance protection being only incidental. The experiment shows that all instances labeled under this class are correctly classified
Rule#4	**IF** AGE = '(0-14]' AND COVER_TYPE = PNFE **THEN** cluster0 (13.0)	Lower age customers which their life insurance policy covers for Pre- needed funeral expense are labeled under cluster0. Based on the information Thirteen instances assigned to cluster1. In addition all instances that satisfy this condition are correctly classified.
Rule#5	**IF** AGE = '(0-14]' COVER_TYPE = CAI **THEN** cluster1 (13.0)	Policy holders whose age group are below fourteen years old and their policy covers for Complementary or Comprehensive accident insurance are classified under as cluster1 label. Risks

		covered by such insurance policy covers are death, total or, partial or permanent disabilities. Thirteen instances are classified under this rule. All instances are classified correctly
Rule#6	**IF** AGE = '(0-14]' COVER_TYPE = HIEYE **THEN** cluster1 (4.0)	Lower age policyholders and their insurance policy can cover the expenses of optical or Eye glass is labeled under cluster1 class. Four instances are assigned to this rule and all instances that fulfill this condition are correctly classified.
Rule#7	**IF** AGE = '(0-14]' COVER_TYPE = HIDENTAL **THEN** cluster1 (4.0)	If the age of the insured is between 0 and fourteen and the insurance benefits the insured ones for dental check-up then cluster1.
Rule#8	**IF** AGE = '(0-14]' COVER_TYPE = HIPREGN **THEN** cluster1 (2.0)	Two instances labeled under to cluster1 class are extracted from rule that implicate customers whose age are below fourteen and their policy covers Pregnancy check-up
Rule#9	**IF** AGE = '(0-14]' COVER_TYPE = HIMAT **THEN** cluster1 (2.0)	If the policy owner's age is under fourteen and their policy covers for Maternity benefit, they are classified under cluster1. Two records that satisfy the above condition are classified under this class and all of instance correctly classified
Rule#10	**IF** AGE = '(0-14]' COVER_TYPE = HIHOSPSICK **THEN** cluster1 (5.0)	Under this rule, five instances of life insurance customers labelled as cluster1 class. The rule indicates policyholders whose age fall between 0 and 14 and risk covered by their policy all instances assigned to cluster1 are correctly classified
Rule#11	**IF** AGE = '(14-28]' AND INSURED_VALUE = '(min value-0.115407]' **THEN** cluster1 (9519.0)	Middle age policyholders of age between 14 -28 and the sum insurance value is less than 67719 Birr are labeled under cluster1 class .Of 9519 instances assigned to this class label , all instances are correctly classified.
Rule#12	IF AGE = '(14-28]' AND INSURED_VALUE = '(0.115407-0.132349]' **THEN** cluster0 (4.0/1.0)	Insurers whose age is between 14-28 and the sum of insured value fall between (67719 to 77366.5] are classified under clusters1. Of four instance assigned to this class , one instance incorrectly classified
Rule#13	IF AGE = '(14-28]' AND INSURED_VALUE = '(0.132349-inf)' **THEN** cluster1 (4.0)	Insurers whose age category found between 14-28 and the insured value are greater than 77366.5 Birr are then labeled as cluster1class. The four instances assigned to this class label are correctly predicted as instances class

Ru1e#14	IF AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = DEATH **THEN** cluster1 (192.0)	One hundred ninety two instances are classified as class of the second cluster, clsuster1. The rule discovered indicates that if policyholders age is greater 28 but the not exceed the maximum age of 65 that EIC insurance regulation obliges and the sum insurers policy value less than 67719 Birr and the risk covered by their policy is death alone (life insurance only) then instances satisfy this condition are assigned under cluster1
Rule#15	IF AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = PNFE **THEN** cluster0 (191.0)	If the age of insurers is greater than 28 and less than 65 and insured value of the customers below 67719 Birr and the risk covered by their policy implies for pre needed funeral expense, then class label fall under cluster0. Based on the above rule one hundred ninety one policy owners assigned this class and all of the instances are correctly classified.
Rule#15	**IF** AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = CAI **THEN** cluster0 (191.0)	policy holders whose greater than 28 years and their the total insured value is below 67719 Birr and their policy covers for Comprehensive accident insurance are classified under as cluster0 label. Of 191 instances assigned to the this class label , all of them predicted correctly
Rule#16	**IF** AGE = '(28-41)' AND INSURED_VALUE = '(min value-0.115407]' AND COVER_TYPE = HIHOSPSICK **THEN** cluster0 (1.0)	policy holders whose greater than 28 years and their the total insured value is below 67719 Birr and their plocy covers medical expenses of Hospitalization and sickness classified under as cluster0 label. One instances assigned to the this class label and it is correctly classified
Rule#17	**IF** AGE = '(28-41)' AND NSURED_VALUE = '(0.115407-0.132349] **THEN** cluster0 (11437.0)	If the age the insurer id greater than 28 years and total value of the sum assured is between 67719 to 77366.5 are labeled as cluster0. 11437.0 instances are satisfying this condition. And all of the instances are correctly classified
Rule#18	**IF** AGE = '(28-41)' AND INSURED_VALUE = '(0.132349-inf) **THEN** cluster0 (16.0)	Adults and older age Policyholders whose sum assured value is greater than 77366.5 are classified under cluster0
Rule#19	**IF** AGE = '(14-28]'**: THEN** cluster1 (9527.0/3.0)	Policyholders, whose age are between 14-28 and the sum of insured value fall between (67719 to 77366.5 Birr are classified under clusters1. Of 9527 instance assigned to this class, three of

		them are incorrectly classified instance incorrectly classified. After the parameters modification, the size of trees and the number of leaves were declined to 23 and 27 respectively. But the confusion matrix of experiment#7 showed the same results observed in experiment#6.
Rule#20	**IF** AGE = '(0-14]' **THEN** cluster1 (67.0/13.0)	If the age of the insured is between 0 and fourteen(children) are classified under to the second cluster1 , cluster1. 67 instances are assigned to this class label. among 13 instances are misclassified .

Appendix 3 Unstructured Interview guide questions

A) Unstructured Interview guide questions prepared for EIC LAD managers

- What is life insurance? When does it start?
- What is the mission, vision and objective of the department?
- How do you attract new customers , advertising or how do you retain new customer
- What is the average cost of incurred to attract for customers
- How does the agreement will be made with your customers?
- Is there any documentation or system which can handle the customers' data?
- What are the main services included or covered under life insurance?
- Premium mode? In how many time gaps those customers are obligated to pay?
- Who encode the data to the system of INSIS is there any other system?
- How do you evaluate the customers' value?

B) Interview guide questions for EIC ICTM Deputy Manager

1. Does your organization have a database system? When did it starts
2. What kinds of data are handled by are by the DB
3. How many employees are hired to work on the database?

Appendix 4: Transcriptions of the Interview Data

A) Transcription of Unstructured Interview conducted with EIC LAD managers

Researcher (R): What is life insurance? When does it start?

Interviewee (I): Life insurance is sharing of loss against unexpected loss. It starts forty years ago.

R: what is the Mission, vision objective of the division?

I: Generally, the objective of the life insurance is to reimburse the insured one against losses caused by death, becoming old or old age or disabilities, which might be total or partial disabilities. The objective of our division is to satisfy our customers by increasing the accessibility life insurances and providing finest customer service.

R: How do you attract new customers, advertising or how do you retain new customer

I: There many sales person found in the marketing department that are trained to attract customers.

R: What is the average cost of incurred to attract for customers?

I: It however as it covers the cost made for sales person between from the initially payment of the premium. It is actually administered by the premium system of our corporation. In fact, the premium payments of the insured are not only used to cover the expenses of expenses of acquisition, also it covers the administration expense.

R: How does the agreement will be made with your customers?

I: The contractual agreement which signed among the corporation, marketing agents and the policy holders regarding to the policy type, the plan or duration of insurance, mode of premium payment, beneficiary chosen, additional supplementary contracts benefits desired, etc.....

R: Is there any documentation or system which can handle the customer data?

I: Definitely, the customer data are stored on the INSIS data base. Specifically the life insurance customer data are stored in the LIFE INSIS data base separately

R: What are the main services included or covered under life insurance?

I: Life insurance has developed along three different classes or lines of insurance such as individual and group insurance. Life insurance is available in each of these classes under basically three plans or forms. These plans are term, endowment and whole life.

R: Premium mode? In how many time gaps those customers are obligated to pay?

I: It is also important to know that premiums, if the policy owner wishes, may be paid more frequently than annually. Life insurance premiums are customarily expressed on an annual basis, payable at the beginning of each succeeding policy year. In fact, this was one of the first assumptions established by life insurance actuaries as a basis for premium computation. For a slight additional payment most polices permit payment on a semi-annual, quarterly, or monthly basis, if the policy owner prefers. These various frequencies are referred to as "mode" of premium payment. Regarding the payments, the cost of a life insurance policy is stated in terms of an annual premium. It is possible to pay by monthly, quarter of half-yearly installments as well.

R: Who encode the data to the system of INSIS? Is there any other system?

I: There are a number of trained IT specialists who administer the customer information. Our customer information is

R: How do you evaluate the customers' value?

I: Currently, we better say that we evaluate our customers based the risk assessment system, which assess the risk that the insured involved and there is also a premium system which concerns the number and timing of claims that the insurer must be prepared to pay. However, there is no such a system which specifically measures our customer's value. .

B) Transcription of Unstructured Interview conducted with EIC ICTM Deputy Manager

R: Does your organization have a database system? When did it start?

I: Absolutely. Our corporation has a data base that has been fully implemented starting from July 1, 2011. This makes EIC the first in kind for the insurance industry in the Country. There two packages developed by Oracle Company known as LIFE INSIS and GENERAL INSIS, which serves to LIFE INSIS

R: What kinds of data are handled are by the DB

I: The data base stores information such as the claim, underwriting, policy holders' information, sales agent, insurances policy, and financial data.

R: How many employees are hired to work on the database?

I: Generally, there are 26 workforces are hired to operate and support the system. These employees are trained IT specialists. Among them, 20 employees work on application of the system while the others are hired to work on the hardware part.

Appendix 4: Document Analysis Checklist

No	Issues/ points of analysis	Reponses	
		Yes	No
I	**Issues Related to Life insurance in the insurance of person's document**		
1.	Are the meaning and purpose life insurances stated in the document?		
2.	Does the document clearly indicate the types of life insurances?		
3.	Does it include the forms/plans of life insurance?		
4.	Does it consist of contracts of life insurances?		
II	**Issues concerning researches in the company**		
1.	Are there researches concerning data mining in EIC library?		
2.	Are there researches related to the application of data mining in the life insurance domain in the library?		
II	**Issues concerning data base of EIC(life INSIS)**		
1.	Does the data base customers' information that related with customers' transactional, demographic, life insurance policy and others which can explain the policy holders' information?		