Analysing Citizen Profiles with Data Mining

W. A. Mohotti

139172C

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology.

April 2016

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student

W. A Mohotti

Signature of Student

Date: 10 - 04 - 2016

Supervised by

Name of Supervisor

S. C. Premaratne

Signature of Supervisor

UOM Verified Signature

Date: 10/04/2016

Dedication

We dedicate the output of this research work and thesis to government policy setters who are trying to uplift lifestyles of Sri Lankans by addressing the citizens' problems. Also we specially dedicate this system to all those who generously contributed their valuable time, advising and helping in doing this research, specially to my supervisor Mr. S.C. Premaratne. In Sri Lanka, area of analyzing citizen profiles is not effectively done with appropriate techniques. It is with this thought in mind that we have done this research. I hope the research and the findings described below will provide a useful insight for analyzing lifestyle data to provide solutions to issues attach with citizens' life patterns.

Acknowledgement

First of all I would like to thank my project supervisor Mr. S.C. Premaratne who spent his valuable time for guiding this research to make it a success. Furthermore, my next big thank goes to Prof. Asoka Karunandha who taught us Research Methodology and Literature Review and thesis writing subjects which were the basis for this research.

Not only that my thanks should go to all the lecturers in M.Sc in Information Technology degree program of Faculty of IT, who gave their hands to sharpen our knowledge and ideas throughout these two years as they were the illumination which lit up our path ways to success.

Apart from the people who were directly involved, many more helped to make this project a success. Department of Census and Statistics contribute to this research by giving their HIES 2012/2013 data. So thank you all for your great support. Finally, I would like to thank all the batch mates of the M.Sc. in IT degree program who gave their valuable feedbacks to improve the results of the research.

Abstract

There is an exponential growth in issues attached with lifestyles of Sri Lankans over the past few decades. These may contribute to low down the life quality within citizens. In Sri Lanka, there are no adequate researches in the field of analyzing lifestyle data. Though there are few researches which have analyzed the causes for the socio-economic problems, such approaches are not capable of handling big data effectively and not efficient in predicting or describing the issues attach with lifestyle.

Hence, the research has been conducted to analyze citizen profiles in effective way to explore different lifestyle issues. It is hypothesized that analyzing citizen profiles can be done through data mining according to the output want to achieve through predictive or descriptive techniques. The solution takes HIES data set as the input and predict the factors attach with a particular lifestyle issue or describe specific lifestyle issue with its associative causes. Having received the input, this approach preprocessed the dataset to remove the anomalies. Then build data models to represent the lifestyle issue by extracting attributes from HIES data set. Then proceed with pattern recognition for the issues. The important patterns recognized through this approach will be useful for government and policy makers to set up appropriate government policies to uplift the life quality of citizen. The overall design of the research consists of two research question, one question used predictive mining based solution and other one is based on descriptive mining. Classification in data mining was used in finding the factors and their relationships that associated with no schooling and dropouts as those were predictive mining tasks. Clustering is used to explore the relationship between chronic diseases and family.

The overall research is designed using WEKA data mining tool and SPSS statistical tool. Finally, the data models build for citizen profile analysis using data mining techniques are evaluated for their performance using measurements such as value for accuracy, error rate, training time, TP rate, FP rate and ROC measurement.

Contents

List of Figures

School Dropouts

List of Tables

Chapter 1

Introduction

1.1. Prolegomena

Exponential growth of issues attach with lifestyles which need careful attention has recorded new dimension in analyzing citizen profiles in detail. In particular, huge volume in lifestyle data and variety in lifestyle data create the need of analysing big data to solve these issues [13][21][24].This trend has created a research challenge to analyze, forecast and describe the pattern behind these life style data. At present, in Sri Lanka, such analyses are done with manual techniques and normal statistical methods, giving very inefficient and inaccurate results. We have conducted a research to offer an in depth analysis of citizen profiles to explore lifestyle issues. Our solution has recorded high accuracy level.

1.2. Background & Motivation

In any country, major driving factor behind citizens' profile is its lifestyle. Life style means a way individuals live their lives within households and societies, which shows how they deal with their physical, psychological, social, and economic environments on daily life. Therefore lifestyle of a citizen is a composition of motivations, needs, and wants which is influenced by factors such as culture, family, reference groups, and social class. According to existing literature well-established European countries focus on factors such as income, health, education, labour market, household structure and living conditions to determine the living style of Europe [4].Sri Lanka as an Asian developing country also pay attention to same factors such as education, health, and housing in determining development of citizen's life style according to its major financial document, central bank report [9]. Hence in generally, life style of a citizen can be measured by factors such as their socio-economic condition, education level and health status.

A good understanding about citizens of the country using those lifestyle measurements is a crucial factor when determining effective economic and social policies by the government. Furthermore socio-economic segmentation of their citizens based on features like income, education or health is a key concern area for a

government to set appropriate future plans to uplift citizens' life quality. On the other hand, pay attention to real patterns and trends reveal by the citizen profiles can be used to find out issues in monetary allocation by the government. Also those patterns can be used to find out misuse of government funds and irregularities. Hence, proper analysis of data related to citizen's current life style and finding significant reasons that lead to life style will be a crucial matter for a development of a country.

But analyzing useful factors for better decision making is a problem due to huge amount of data generated and associated with lifestyles. Though Central bank and Department of Census and Statistics collect data for the major attributes in lifestyles, those institutes are not paying attention to hidden pattern behind them. In most of the currently existing researches in the world, these factors have been examined carefully using principal component analysis and clustering techniques to identify their dominant combinations for activity patterns[15],[17],[18]. Then those resulted association rules are used to improve citizen satisfaction [2]. But in Sri Lanka, those data are not effectively use to find out the socio-economic issues. If properly analyzed and investigate those gigantic life style data which have been kept in institutes such as department of census and statistics, governance of the country can be improved and controlled. Those analyses can wipe out misuse of state fund allocation, monitor the progress of development activities which lay down by the government and can track the areas which government should focus to uplift the life quality.

Therefore the research proposes a method for analyzing citizen profiles in Sri Lanka to investigate the hidden issues. This analysis is supported by information extracted from a Household Income and Expenditure survey (HIES) of Department of Census and Statistics [11]. A set of typical life cycle profiles are planned to explore from the dataset based on citizens' income and expenditure details given for the survey. Those profiles taken from HIES dataset will further examine for predictive analysis and descriptive analysis. Predictive mining process will be used to predict the probability of an outcome or future behavior. Also citizen profiles which describe the dominant factors attach with particular lifestyle behavior will use for descriptive mining.

1.3. Aims & Objectives

1.3.1 Aim

- Aim is to gain a wide knowledge of data model construction and application of data mining techniques for life style profiling of a country like Sri Lanka.

1.3.2 Objectives

- Examine which life style pattern reveal issues in national development and important conditions of household profiles
- Identification of major attributes and activities contributes to profiling of life style in Sri Lanka.
- Acquire the knowledge regarding data mining and other relevant technologies for analyzing household life styles and activity profiles.
- Construct models for analysing life style profiles by means of correct data mining technique.
- Evaluate models to investigate accuracy of profiling.
- Discover linkages of household demographics and 'lifestyle' choices of citizens to their behaviour patterns.
- Foresee the future direction of socio-economic patterns of Sri Lanka.

1.4 Proposed Solution

As better decision making of government depends on proper analysis of life styles of citizens, with in this research we proposed a method using data mining. Researchers have identified data mining as a best solution for digging useful hidden patterns within large repositories of data using the support of different software tools with its ability to deal with large number of dynamic variables simultaneously.

As the initial step, find out the factors that contribute to determine the lifestyle's of citizens. After that define the sub research questions which provide higher impact of development of Sri Lanka. Then data set obtained by the department of census and statistics are preprocessed and prepared for further analysis. According to selected sub research questions find out main socio-economic, health and education attributes that provide dominant contribution to determine that particular condition.

As data mining is a standard process meant of discovering correlations, patterns, trends or relationships by searching through large data sets stored in data stores, public databases, and data warehouses, it is selected as the basic methodology for this research. Data mining process had to follow basic life cycle sub processes which mentioned in Figure 1.1 to build appropriate models and to generate predictions.

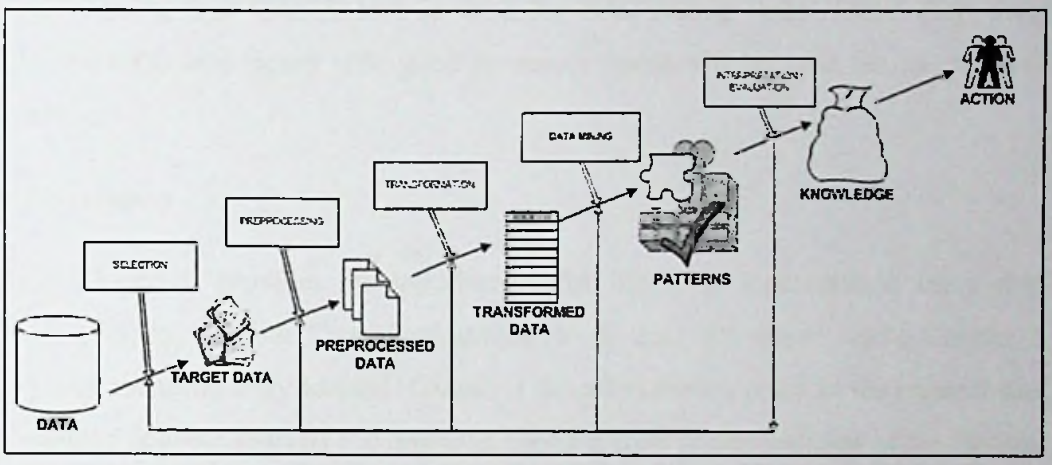

Figure 1.1: Steps in Data mining process

Selecting attributes data which is required to solve the particular sub question is the first step in data mining process. In their focusing on subset of samples on which discovery is to be performed. Then by removing noise, outliers and missing values dataset is prepared for further analysis. Then data are transformed or consolidated into forms appropriate for mining using features like attribute construction and smoothing. Finally build a data model that can be used to find causes and effects for a research sub question. Through the built data model find out the attributes that go hand in hand with particular condition. These identified association rules and loop holes can be used as inputs for government policy setting.

After building the data model we can determine the attributes which are having high correlation with the condition under experiment, probability of a particular condition and causes for a particular effect. This belongs to descriptive data mining task of the research. As the predictive data mining task build the data model in a way which cater further prediction for future scenarios. Then will determine the occurrence of particular effect in presence of particular lifestyle attribute and its percentage.

1.5 Resource Requirements

Access to common resources like books, journals, research papers and magazines about data mining and life style profiling are used for literature review. Also micro data of Household Income and Expenditure survey (HIES) obtained from Department of Census and Statistics is used to build data models. SPSS tool is used for data preprocessing and WEKA tool is used for data mining tasks. Other than these resources PC or a laptop with good processor speed will be used for the research purpose.

1.6 Summary

This Chapter 1 provides the introduction for life style segmentation using data mining. Then, Chapter 2 presents similar work done by others while Chapter 3 explains the technology adapted. Chapter 4 describes the approach of the research and Chapter 5 is about analysis and design. Chapter 6 is on implementation of the System. Then, Chapter 7 discussed about the evaluation of the system. Finally, Chapter 8 presents the conclusion and further work. List of references provides as the last section.

Chapter 2

State of the art of exploring issues in citizen profiles

2.1 Introduction

This chapter describes about the currently existing methods and measures for life style segmentation in Sri Lankan context and worldwide. Especially this chapter focuses on measures that used in Sri Lanka according to central bank report and describes about the loop holes in analyzing citizen profiles which are not address by the annual central bank report. Also this chapter focuses on different methods used around the world to investigate about lifestyle patterns that are useful for development of a country.

2.2 Lifestyle and household profiling

The distribution of wealth, resources and services across households has been an underlying consideration in government concerns on various issues, including development plans, taxation and social welfare. Results generate by analyzing those data are used to build government development policies to uplift life quality of households.

Internationally there are around 4,152 household surveys with economic and social variables which span across most of the countries in the world. Among them 266 of which are household surveys on income and expenditures [36].The main aim of analyzing household survey data is to prepare estimation for fields such as education, health, transportation and communication. Furthermore, insight about the issues such as types of education perceived by household members or cost and benefits of health services used by family members can be used to uplift lifestyle of citizens in a country. Hence, to obtain those decisions, living standards measurements are calculated by examining lifestyles of households.

World Bank is an international body which uses data about occupation, poverty and access to social services, such as health care and education to calculate measurements attach with life quality. It's living standards measurement study(LSMS) survey collects wide-ranging data on most aspects of household welfare such as consumption, income from activities in the labour market, household enterprises or

agriculture, asset ownership, migration, health, education, nutrition, fertility, savings and credit, and anthropometrics. LSMS was intended to help country's statistical institutes to analyze household survey data for policy needs, and provide policy makers with data that can be used to understand the factors of detected social and economic outcome[36],[32].

Not only that, when consider the focus area of Organization for Economic Co-operation and Development (OECD) which establish to promote policies that will improve the economic and social well-being of people around the world also pay attention to field such as communication, transportation, education and energy.

2.2.1 Lifestyle and household profiling in Europe

Problems such as in Europe, how much income poverty is there? Is inequity increasing? Does a job assurance escape from income poverty? How welfare help to cope with the economic crisis? ; are issues need to be addressed for the development of any Europe country. Therefore, major areas that should be focused for the socio-economic statistics that address social and economic issues in Europe are income, living conditions of Europe's households, health, labour market and education [12].

There are number of similarities in demographic trends in Europe. By analyzing those Europe household data can find answers for critical questions such as how many households considered 'at–risk-of poverty', presence of longstanding illness or disability, distribution of labour earnings and provision of public services. Furthermore patterns such as educational intensity of employment, level of fertility for household structure and variation of health by socio-economic status are used to analyze in Europe countries to determine their lifestyle. Hence, main focus areas of Europe for citizen profiling are employment, income inequality or poverty, housing, health, education, deprivation and social exclusion. Careful analysis of those fields allows telling about how the workers of Europe earn their living, about the living arrangements of Europeans, about their social participation, and about the ways in which their incomes are affected by taxes and transfers.

2.2.2 Lifestyle and household profiling in Asia

The current concern over chronic poverty and rising income inequality in many Asian countries of the developing world, highlight the need for deeper thoughtfulness of not simply the numbers of the poor but also the nature of poverty. To focus on those critical issues most of the Asian countries depend on fields such as normal demographics data, income, education, occupation, location of residence [28],[34]. According to Asian bank report where most of the Asian countries are members, mainly focus on income, education, health, agriculture and other social protection schemas to improve Asian's life standard [3].

2.2.3 Lifestyle and household profiling in Sri Lanka

The improvement of economic and social infrastructure continued to be a priority area in the national development agenda to support the economic growth process in Sri Lanka .According to Central bank report of Sri Lanka, government has identified efficient and effective healthcare sector with reduced regional differences and improved equity in accessibility to healthcare services will be led to an improved state of public health in the country. Furthermore it stated Sri Lanka needs highly accessible and quality, primary and secondary education that provide the foundation for the knowledge based economic and social development.

Also it mentioned the Sri Lanka's effort to achieve admirable usage of communication and information technology services in economic activities to improve life quality of citizens. Moreover government recognized the importance of improved road and railway network will provide forward linkage to the development process. Therefore to improve the lifestyle of Sri Lankan citizens currently government carrying out projects for road and railway development, power plants, education sector development and public health sector development with help of foreign financial assistance[9]. As mentioned in central bank report main focus areas of Sri Lanka to determine the lifestyle are communication, power consumption ,transportation ,water supply and irrigation , education , health sector ,housing & urban development and environment.

2.2.3.1 Issues with analyzing citizen profiles in Sri Lanka

Department of census and statistics is a government authority which use to collect and analyze data to determine development of Sri Lanka. It conducts several surveys such as demographic & health survey, school census, population census and Household Income and Expenditure Survey (HIES).

HIES survey provides information on household income and expenditure to measure the levels and changes in living conditions of the people. Data collected from this survey is used to observe the consumption patterns to compute various other socio-economic indicators such as poverty price indices. HIES is conducted over year to capture seasonal variations of income and expenditure patterns in Sri Lanka. The HIES questionnaire consider nine priority areas to collect household information covering the demography, school education, health, food and non-food expenditure, income, inventory of durable goods, access to facilities in the area and debts of the households, housing information and agriculture holdings & Livestock.

Using HIES data, department of census and statistics calculate figures such as average monthly household income, median monthly household income, average monthly per-capita income, average monthly income receiver's income, average monthly household expenditure, household size, number of income receiver's per household. Furthermore ,using HIES data department of census and statistics estimate household population, distribution of school attendance, health status of household population, income inequality using gini coefficient, expenditure for food and non-food ratio In Sri Lanka.

Though department of census and statistics explore HIES data set to find out important economic and social indices, it does not pay attention to find out hidden pattern behind those data. Trends reveal by questions such as "What are the factors that cause for the popularity of main transportation mode use by public?", "What is the significant energy consumption pattern?", "What are the reasons for the popularity of a particular communication tool used by citizens?" are not address by HIES dataset. Furthermore it does not disclose pattern in household consumption & expenditure by type of good such as services or non-durable goods or durable goods.

Another major loop hole of HIES analysis is, it doesn't try to reveal issues and mismatches of economic development programs that are conducted and ongoing. Basically central bank report is reporting about number of development projects and monetary allocation for them. Most of the cases, it highlighted the positive side of them. Therefore it is necessary to find out whether there are weaknesses in those projects, are those projects beneficial for citizens , did government spend money for actually valuable projects and what are the root causes behind social issues. Thus, not addressing these issues and not revealing patterns behind them is a crucial mistake within data analysis of HIES.

2.3 Methods for lifestyle data analysis

In citizen lifestyle analysing, major task is to investigate the responses in survey data. Though we consider sample survey or full survey, lifestyle data of citizens in a country will be huge in size and contain variety of information which ultimately falls into big data category. Handling this type of data set should be done with a special care to interpret the results accurately. Table 2.1 provides a comparison of currently existing methods for analyzing big data [7],[16],[20].

Statistics	Database Queries	Machine Learning	Data Mining
Providing a theory for estimating probabilities of predictions and distinguish between significant findings	Analyze the data in a database and return raw data that satisfies certain constraints	Collection of algorithms and techniques used to design systems that learn from data	Deals with the extraction of previously unknown and interesting information from raw data and returns models of the data in question
		Has mathematical and statistical basis that does not take into account domain knowledge and data pre-processing.	Cover full process from data pre-processing to deployment of findings: Facilitating data visualization, data understanding, and reducing the dimension

Deal with structured data in order to solve structured problem			Data mining is designed to deal with structured data in order to solve unstructured business problems
Results are software and researcher independent			Results are software and researcher dependent (absence of implementation standards)
Inference reflects statistical hypothesis testing			Inference reflects computational properties of data mining algorithm
Depend on assumption			Include generalization capabilities
			May have long training process

Table 2.1: Comparison of methods for big data analysis

In lifestyle survey data analysis, it is necessary to preprocess data to correct missing values and to remove outliers after selecting the attribute from the problem domain. Then using smoothing like techniques we may need to consolidate into forms appropriate for analyzing. Thereby we can build a data model to represent a particular issue in lifestyle profiles. Finally we have to evaluate our model to test the accuracy of its results. This step by step process correctly matches with the phases in data mining process. Data mining consist of business understanding, data understanding, data preparation, modeling, and evaluation and deployment phases [35]. Therefore data mining methodology will be a better method to analyze lifestyle data of citizens to find out hidden issues behind citizen profiles as it belongs to unstructured problem category.

2.3.1 Currently existing methods in the world

Existing literature reveals that, different data mining techniques were used in lifestyle data analysis according to the output try to achieve as in Figure 2.1[15][21][24][25]. Predictive mining tasks used techniques such as classification and regression. On the other hand, descriptive tasks used techniques such as clustering and association rules.

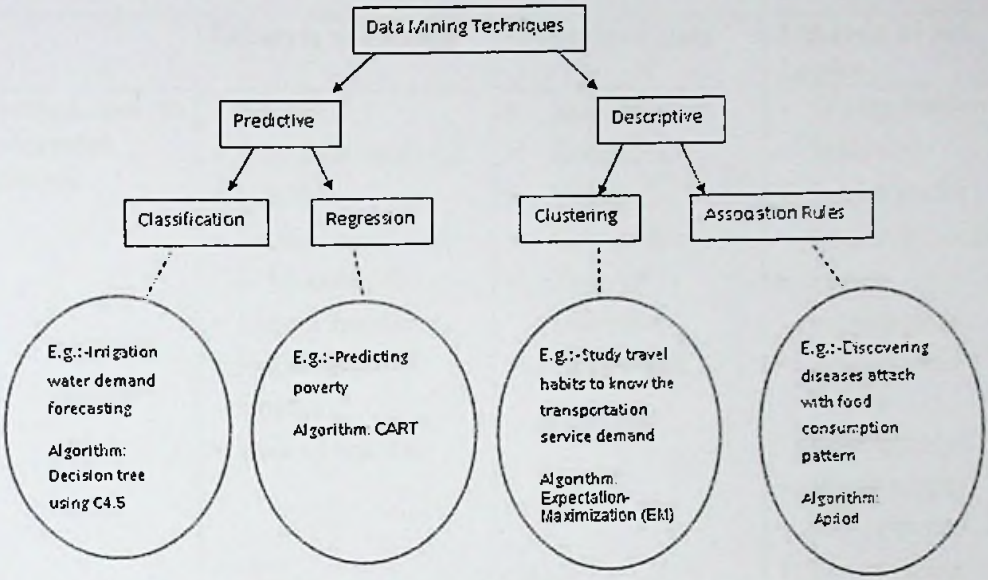

Figure 2.1: Different data mining techniques used for lifestyle data

Also for Sri Lanka, these types of data analysis using citizen lifestyle data are equally important to improve the life quality of citizens. However, these things are completely missing in Sri Lankan context as lack of awareness or poor initiative by the government. There may be other minor reasons like affordability and sustainability that cause not to have these types of applications within Sri Lanka as it is a developing country.

2.4 Research Question

It is evident from the literature that, in-depth analysis in lifestyle data by considering driving measures of lifestyles remains a research challenge for Sri Lankan context. Hence, we intended to solve this problem of analyzing citizen profiles to explore the problems attach with lifestyles by data mining based solution.

2.5 Summary

This chapter provided complete description of existing measure in determining lifestyles of citizens as given below in Table 2.2.

	Lifestyle of Europe	Lifestyle of Asia	Lifestyle of Sri Lanka
Factors use to determine lifestyle	• Income • Education and skill • Health • Living conditions of households • Labour market • Household structure • Role of the state	• Income • Education • Health • Occupation • Type of residence • Agriculture	• Transportation • Education • Health sector • Communication • Power consumption • Housing and Urban Development • Water supply and irrigation • Environment

Table 2.2: Comparison of measure use for citizen profiles

After comparing factors determining Europeans lifestyles, Asians lifestyle and Sri Lankans lifestyle common factors such as income, education, health, transportation, communication and power consumption are identified as citizen profiling measures. Though these fields are covered by HIES dataset, further analyzing them to address social and economic issues is currently lacking area which is very important. Moreover this chapter discussed about different big data analysis methods and existing lifestyle analysis applications in the world. Next chapter will elaborate data mining technology in detail which uses to analyse citizen profiles.

Chapter 3

Technology adapted

3.1. Introduction

Chapter 2 discussed the existing methods and attributes for analyzing citizen profiles to identify the issues in lifestyles. This chapter presents data mining technology which is selected to analyze citizen profiles effectively in detail. This chapter highlights the effectiveness of selected technology that distinguishes it from the technologies applied in existing literature.

3.2. What is Data Mining?

Nowadays we come across with large, complex set of data which generated by computers, networks and humans. Government agencies, scientific institutes and businesses dedicated enormous amount of resources to collect and store these data. Among them only small amount is used because, in many cases volume is too large to manage, or the data structures themselves are too complicated to analyse effectively. Ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming important in this competitive world. Therefore the data mining process is emerged. The entire process of applying a computer based methodologies including new techniques for discovering knowledge can be treated as "Data Mining".

Data mining is a process of analyzing big data from different perspectives and summarizing them to useful information by the means of different techniques. Formally this is defined as "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [14]". the overall process of finding useful information from raw data involves the sequential line up of steps such as developing an understanding of the application domain, creating a target data set based on a smart way of selecting data by focusing on a subset of variables or data samples, data cleaning and preprocessing, data reduction and projection, choosing the data mining task, choosing the data mining algorithm, data mining, interpreting mined patterns and consolidating discovered knowledge. Hence, process of data mining can be consider as total solution which consist of phases business understanding, data

understanding and preparation, modeling, evaluation and deployments in interactive manner as in Figure 3.1 [22].

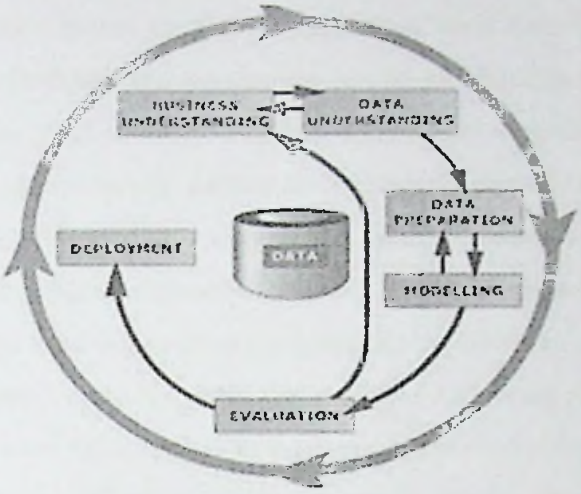

Figure 3.1: Cross industry standard process for data mining

Data mining is having two primary goals of being predictive or descriptive. According to the task, different techniques are available in data mining. For predictive tasks, techniques such as classification, regression and deviation detection are used. Meanwhile techniques such as association rules, cluster analysis are used for descriptive tasks as in Figure 3.2. Predictive algorithms determine models or rules to predict the values of variables when given input data. On the other hand descriptive algorithms determine models to summarize the data in some manner. Therefore selecting the most appropriate mining technique depends on the goal which users going to achieve.

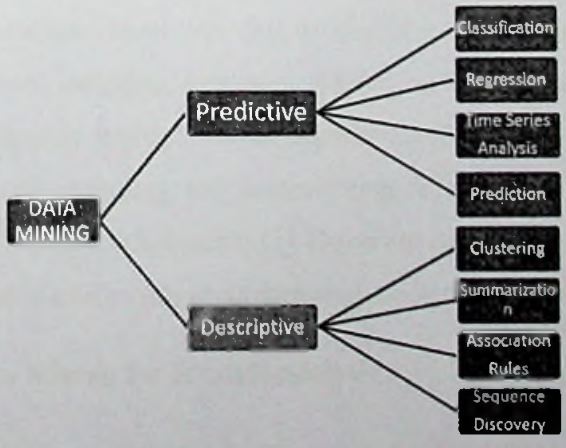

Figure 3.2: Data mining Techniques classification

Researchers have identified data mining as interesting, beneficial subject area, because of successful KDD applications. When considering business applications of data mining, market basket analysis in marketing, fraud detection in finance, defect findings in manufacturing and governance can be identified as effective applications [24]. In science, data mining achieve remarkable success in fields such as telecommunication, astronomy, pattern discovery in biology. There are so many other application areas of data mining such as web content mining, web usage mining and stream data mining. Most of these phenomena are totally absent in many developing countries. Though these things are equally applicable for developing countries as well as the industrialized countries, with the available socioeconomic and sociopolitical requirement there are some barriers to implement them in developing countries. Issues such as sustainability, affordability and community identity may be some causes that directly affect to developing countries in developing information systems [29].

3.3. Reasons for using data mining for citizen profile analysis

Data associated with citizen lifestyle patterns are huge in size and vary in complexity. Therefore these data which kept in computers of government agencies after doing different lifestyle surveys time to time can be treated as big data. Big data is a term given to data which is huge in volume, variety and velocity. Literature review which compares different data analysis methods used by researches proved that data mining is the best option for big data analysis.

When comparing data mining with traditional methods of querying a database, latter require predefine variables. Moreover data mining is a complicated type of database querying which allows including new and greater number of variables. Also data mining allows examining multiple areas simultaneously. So data mining can be defined as a process of analyzing and summarizing data from different perspectives and converting it into useful information [1].Therefore data mining can be used as the most suitable method to citizen life style data analysis according to literature review.

3.4. How to use data mining for lifestyle analysis

In lifestyle data analysis we may need to find out factors related to particular socio-economic issue and the relationship of causes for a particular phenomenon. Furthermore we need to find out in detail view how some problems occur and

combination of causes that lead to some phenomena. As citizen profiling is conducting to avoid socio-economic issues, it is essential to find out the root causes for them and their nature. Therefore this research utilizes different data mining techniques in predictive stream and descriptive stream according to selected research sub question.

For an example, research sub question to reveal the factors causes for school dropouts and no-schooling can use predictive data mining as an effective technique. For analyzing school dropouts primarily need to take data such as demographic data of students, income of households and education level of parents from surveys such as Household Income and Expenditure Survey (HIES). Then using data preprocessing techniques in data mining can reduce the uncertainty of data by removing erroneous data and missing values. As the next task, by analyzing the significance of factors major data fields can be chosen. To exactly find out the correlation between these data fields which contribute to school dropout or no-schooling, classification technique in data mining can be utilized. Revealing hidden pattern behind school dropout can be effectively done through classification as it is a predictive mining task.

3.5. Summary

This chapter presented data mining as the technology proposed to analyze citizen profiles to identify the issues in lifestyle. In this sense, it is pointed out how the data mining offers an efficient and accurate solution for citizen profile analysis. The next chapter shows a novel approach of analyzing citizen profiles through technology presented here.

Chapter 4

A novel approach for citizen profiling analysis

4.1. Introduction

Chapter 3 discussed the technology for analyzing citizen profiles to identify the issues in lifestyles. This chapter presents our approach to analyze citizen profiles in detail using data mining under several headings, namely, hypothesis, input, output, process, users and features. This chapter highlights the key features that distinguish our novel approach from the existing approaches for citizen profile analysis in Lanka.

4.2. Hypothesis

Using data mining techniques can predict or explore issues attach with lifestyles. Predictive data mining can be used to predict the factors affecting different social issues. Descriptive data mining can be used to explore current situation demonstrated in lifestyles.

4.3. Input

As the initial input for this process, data obtain from Household Income and Expenditure Survey of department of census and statistics is used.

4.4. Output

As the output of this process different data patterns related to the citizen lifestyle has can be reveled according to the sub research question identified. Prediction will be given as output for the research question attach with predictive tasks. Summarization will be given for the research question attach with descriptive tasks.

4.5. Process

In this process of analyzing citizen profiles with data mining all the standard steps in knowledge discovery process which include data selection to evaluation are carried out. Throughout the process the data set is cleaned, formatted and prepared for mining and interpretation.

4.2.1. Data Selection

To determine the citizen's life condition of any country we need to keep track of basic needs, food, clothes and shelter. Therefore to cater that requirement Sri Lanka also focus on expenditures in household for food, clothes, fuel for cooking, transportation and housing facilities. Other than expenditures, both European and Asian countries focus on income as a major area to determine citizen's lifestyles. In Sri Lankan context also income is an important factor. The income for Sri Lanka can be calculated from different sections such as income from occupation, agriculture, nonagricultural means and Samurdhi like social protections schemes. Not only that education and health are other two main important sections which are equally applied to Sri Lankan context when determining citizen profiling. Apart from these major areas in any country their citizen's basic demographic data such as age, gender, household size like factors should be considered in determining citizen's life styles.

Though anthropometrics is used in world to determine lifestyles, in that kind of dynamics, only the height, weight like factors can be measured accurately. As a result of the difficulty of measuring and using, anthropometrics are not suited in Sri Lankan citizen profiling to determine citizen lifestyles. Central bank report of Sri Lanka and major surveys conducted by department of census and statistics also use basic demographic data, education, health, housing and income and expenditures as their major focus areas and excludes anthropometrics measurements.

Most surveys and censuses done across the world collect information on household characteristics, including those related to location, household composition such as household size, sex of the household head. Also they keep track of socio-economic characteristics such as household wealth, dwelling quality and type to determine citizen's lifestyle. These factors are categorized mainly into demographic data, education, health, housing and income and expenditure in Sri Lankan context when consider its major financial reports and surveys. Data from those major areas are used to analyse citizen profiling.

4.2.2. Data Preprocessing

For this research HIES data set is used as the major data source which contains citizen life style data from different areas of the country. Department of Census and Statistics

uses a questionnaire for that which is filled by their officials in the field. Sometimes this collected data may be in poor quality. But accuracy, completeness, consistency, timeliness, believability, value added, interpretability and accessibility are some of the characteristics that should involve with data taken to a research to draw a well-accepted conclusion.

Incomplete data may come from "not applicable" data value when collected, human or hardware or software problems, different considerations between the times when the data was collected and when it is analyzed. Noise in data is another problem which reduces the quality of data. Noisy data may come from human or computer error at data entry, and errors in data transmission. Another issue is inconsistent data which may come from functional dependency violation in linked data. Duplicate records also need data cleaning which lead to poor data quality. Therefore HIES dataset is preprocessed before further analysis.

4.2.3. Data Transformation

This is the step where data is transformed or consolidated into forms appropriate for mining by performing operations such as summary and aggregation. Due to availability of huge amount of data and immense need for tuning those data to useful information and knowledge to support government decisions, smoothing, aggregation, generalization and normalization like strategies used within data transformation process. Smoothing is used to remove noise from data, aggregation is used to summarization and data cube construction, generalization is used to concept hierarchy climbing and normalization is used to scale within a small, specified range. According to the selected data for the sub research question appropriate smoothing and aggregations are done.

4.2.4. Data mining

This is the essential part of this research which used intelligent methods in order to extract data patterns. Not only that but also discovering interesting knowledge includes finding association, changes, anomalies and significant structures from HIES data set. In Association, the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns. In Classification, the

methods are intended for learning different functions that map each item of the selected data into one of a predefined set of classes.

Prediction analysis in predictive mining is related to regression techniques. The key idea behind that is to discover the relationship between the dependent and independent variables, the relationship between the independent variables. Sequential Pattern analysis search for similar patterns in data transaction over a period. These patterns can be used by business analysts to identify relationships among data. Descriptive cluster analysis takes unorganized data and uses automatic methods to put this data into groups. Most of the mathematical models applied in regards to classification can be applied to cluster analysis as well. According to the task we try to achieve through research question, predictive or descriptive mining task is selected.

4.2.5. Evaluation/Interpretation

The data interpretation stage is the most critical one as it integrates knowledge from mined data. There are two essential issues. One issue is how to decide the business value from knowledge patterns discovered in the data mining stage. Another issue is which technique or visualization tool should be used to show the data mining results. Determining the business value from discovered knowledge patterns is similar to playing "puzzles" as different techniques can be used for same dataset with different algorithms and best out of them should be chosen to match with business purpose. Therefore evaluation of mined patterns should be done with respect to goal or objective to maximize the efficiency. In order to properly interpret knowledge patterns, it's important to choose an appropriate visualization tool. Many visualization packages and tools are available, including pie charts, histograms, plots, trees and distribution networks.

4.6. Users

Government and policy makers are the users who should involve with this citizen profile analysis and pay attention to patterns represented by the mining process to uplift citizens' life quality and to set up development programs.

4.7. Features

The solution proposed by this research can be used to analyse huge volume of lifestyle data which are in different forms in consistent manner. Through the predictive mining tasks this solution allows to make predictions for future instances. Moreover, descriptive tasks allow to describe pattern describe by the dataset. This solution provides the output by extracting previously unknown patterns dynamically.

4.8. Summary

This chapter presented our novel approach to analyze citizen profiles to identify the issues in lifestyle. In this sense, it is pointed out how the novel approach offers an efficient and accurate solution for citizen profile analysis using data mining. The next chapter shows the design of the novel approach presented here.

Chapter 5

Research Design for Analysing Citizen

5.1. Introduction

Chapter 4 presented the approach to analyses citizen profiles to identify the issues in lifestyles. This chapter elaborates the approach, and focuses on high level design and sub modules within in the design. Later part of this chapter pays attention to the interactions among the sub modules.

5.2. Research Design

The basic aim of science is to explain a natural phenomenon which is known as theory. In scientific approach, instead of trying to explain each and every separate behavior, the researcher seeks for a general explanation that encompasses and links together many behaviors. Hence, scientific research is systematic, controlled, experimental, public and critical investigation of natural phenomena. It is guided by theory and hypotheses about the presumed relations among such phenomena [27].

In this research of analyzing citizen profiles using data mining, we wish to discover how common particular forms of behavior occurs and their trends by taking large dataset which consist of social and economic data of Sri Lankan citizens. Hence, we have to carry out a quantitative data analysis using behavioural science research methodology. In analyzing citizen lifestyles we need systematic analysis and investigation of human behaviour through controlled, naturalistic observation and scientific experimentation to draw the conclusion. "The behavioral-science paradigm has its roots in natural science research methods. It seeks to develop and justify theories (i.e., principles and laws) that explain or predict organizational and human phenomena surrounding the analysis, design, implementation, management, and use of information systems [5]". Therefore, Behavioural Sciences approach is mainly suitable for this research.

In this research of analyzing citizen profiles using data mining, we have followed the steps in scientific method which consists of observation, study, defining problem, building theoretical framework, hypothesis development, experimental design, data gathering, data analysis and conclusion [23]. Also this research starts with the

observation of existence of social-economic issues in Sri Lanka. Then study was carried out to get information about the factors we should focus on to solve these kinds of problems and how similar problems are solved. Next exactly identify factors influencing socio-economic issues in Sri Lanka and sub research questions that we should focus on to avoid socio-economic issues. After selecting sub research questions to be addressed, under theoretical framework, network of variables or factors have recognized by considering the related work in that problem domain. Then hypothesis development has conducted and experimental design is done using SPSS and WEKA. Results given by different data mining techniques for a particular sub research question are based on data gathered for the sub research question. Those generated results or the data are further analysed for efficiency and accuracy. By considering the analyzed results for a particular research question conclusion can be given.

5.3. Top Level Design

Figure 5.1 illustrated an analytical framework for citizen profiling in Sri Lanka. As a coach holds by its wheels, main ingredients for the household profiling in Sri Lanka are basic demographic data, income and expenditure for needs and wants of citizens, education details, health details, housing information according to the major financial document and results of government surveys.

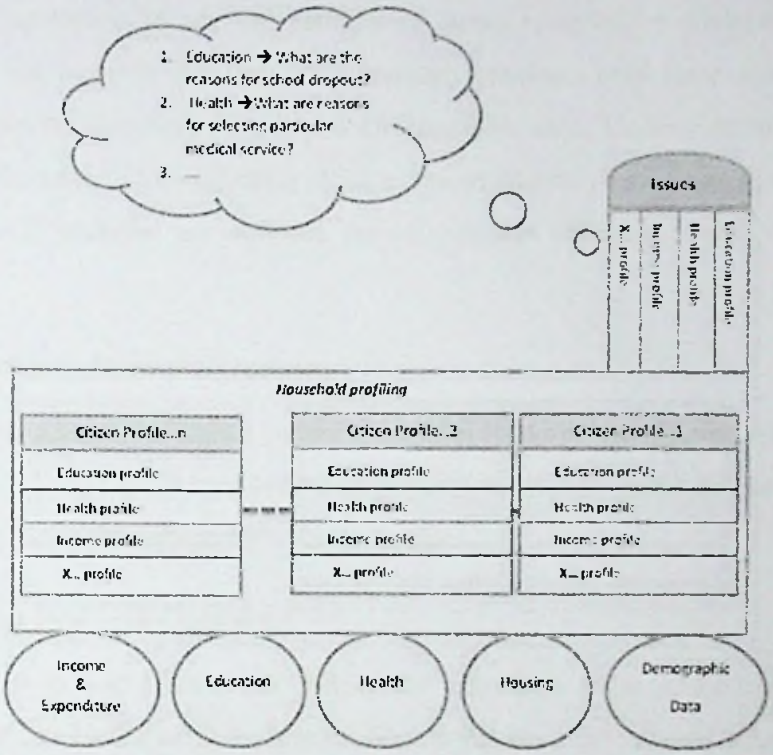

Figure 5.1: Analytical framework for citizen profiling in Sri Lanka

Furthermore, Figure 5.1 express that complete citizen profiles are made by sub profiles such as education profile, health profile and income profile according to major areas which define major requirements in lives of citizens. These major sectors that contribute to define the life quality are the important sections that government needs to handle with special care. These different sub profiles can be collectively taken from citizen profiles to analyse and identify social and economic issues with in country. Therefore, this framework represent the ability of citizen profiles to signify the issues in life styles that need to address by government.This framework is built according to the existing variables in HIES dataset.

5.4. Detailed Design of the Research

Analyzing Sri Lankan citizen profiles to find out the issues within lifestyles is identified as the primary research question in this research. According to the major financial document in Sri Lanka, life quality of Sri Lankans can be categorized into several categories. Major areas among them are Education, Health, Communication,

Transportation, Housing, which can be summarized to according to Figure 5.1. Based on that framework, to analyse patterns or issues revealed by lifestyles we create different sub research question. This research considers only three areas of citizen profile named Education, Health and Demographic data. Under Education category we identified the factors affecting children not to attend school or leave school early. From Health sections we explored the relationship between chronic diseases and income.

5.4.1. Primary Research Question

There is no adequate research carried out in Sri Lanka to identify the issues/patterns attach with lifestyles. Hence primarily propose a solution based on data mining to analyze lifestyle data attach with citizen profiles.

5.4.2. Sub Research Question 1

It is crucial to find out the real factors that contribute to no schooling and school dropout in Sri Lanka using citizen profiles to improve the level of education in Sri Lanka which is still an unexplored matter attach with citizens lifestyles.

5.4.3. Sub Research Question 2

There are no sufficient researches in Sri Lanka which accurately analyse huge volume of lifestyle data to determine the relationship between chronic diseases and family income though it is important subject to pay attention as a country.

5.5. Summary

This chapter provided details on research design and applicability of selected research method for the research. Furthermore this chapter focuses on top level design for the research and how sub research questions are structured with in the research. Subsequent section will be discussed about implementation details according to this design.

Chapter 6

Implementation

6.1. Introduction

In chapter 5, the top level design of the solution has been described in terms of what attributes are used to represent citizen profiles and, two sub research questions. This chapter describes the implementation of each sub research question regarding software, algorithms, method...etc. In that sense, this chapter is about how the system is implemented.

6.2. Solution for Sub Research Question 1:

As the solution for the first sub research question reasons for Sri Lankan children to dropout from school and no schooling and their trends are analysed using following steps.

6.2.1. Existing work in this domain

The analysis on causes of school dropouts in developed countries has emphasized factors such as drug use, alcohol consumption, and parents' psychiatric disorders. Also the factors such as socio-economic status of the family, gender, race, and age of the child are contributed for that. According to a research work carried out in Brazil, where a developing country, identified early parenthood, child labor and poverty after examining socio-economic background, education, health, social capital, home violence and employment as factors of school dropout by estimating a logit model [8].

In a study carried out for Uganda using dimensions such as rural-urban, gender, and age with logistic model analysis found that factors such as rural-urban divide, gender of household head and of pupil, age of the household head, household size, academic achievement of mother and father, distance to school, school fees payment, contribution to economy are causes for primary school dropout [26].

Sri Lanka also investigates about the determinants for out of school for students. That observations suggest age, gender, ethnic group, wealth or rural–urban divide, family poverty, child labour, poor health and disabilities, lack of parental encouragement and family problems, quality of schools and facilities, negative attitudes of teachers, uninteresting lessons, and harassment by teachers and peers, child abuse, natural

disaster and conflict as factors which affect children to be out of school [33]. By Considering these existing works and data available in HIES dataset, initially data are selected manually to build the model one by one using forward method.

6.2.2. Theoretical Framework using SPSS

Initial attribute selection process ends up with attributes such as age, household size, district, sector, ethnicity, religion, participation for economic activities, household head's gender, father's education, mother's education, income, structure of the house and disabilities.

6.2.2.1. Data Preprocessing

To avoid incomplete, noisy, and inconsistent data, data preprocessing is essential before establishing a theoretical framework using variables. Under data preprocessing data cleaning, data integration, data transformation, data discretization techniques are used in this research. In HIES data set there are missing data due to inconsistent with other recorded data, due to misunderstanding or less value at the time of entry. Consequently, tuples have no recorded value for several attributes such as house number, sex, age. In this research to solve this issue two techniques of ignoring the tuple or using a global constant to fill in the missing values have been used (Appendix A). In data integration, integrate metadata from different sources as HIES data set consist of set of .csv files for demographic data, housing, income and education data. For an example file containing demographic data has serial number to identify a person as Person_Serial_No and file containing school education is having same thing as R2_Person_Serial which need to integrate for identification of that entity. In data transformation, attribute or feature construction is used for new attributes construction from the given ones. As an example, different income types such as wages, agriculture income are combining into one and construct new attribute called income for a household. Also to uniquely identify a student has to create a serial number by combining attributes in the HIES data set named district, sector, Psu, Snumber, Hhno, Nhh (Appendix A).

6.2.2.2. Attribute Selection

For the data model construction, multinomial logistic regression is used as dependent variable is nominal with more than two levels with no-schooling and dropouts.

Furthermore, this is an extension of logistic regression, which analyzes binary dependents. We hypnotized that initial attributes (independent variables) are directly contribute to no schooling and dropout. In this analysis, the null hypothesis used is that there is no difference between the model without independent variables and the model with independent variables. If an attribute having significant level or probability which is less than or equal to the level of significance of 0.05, null hypothesis is rejected and select that attribute for the data model (Appendix B). The existence of a relationship between the independent variables and the dependent variable was supported in that situation. Further analysed by stepwise backward elimination to check where least significant attributes will removed until other non-significant attributes are contribute to model. By following that procedure attributes in Table 6.1 are selected for the data model of no-schooling and dropouts at the end.

Noschooling	Dropouts
Age	Age
	HouseholdSize
District	District
Religion	Religion
Is_Active	Is_Active
	HeadSex
	FatherEducation
	MotherEducation
Is_Ill_Disable11	Is_Ill_Disable11

Table 6.1: Attributes selected after multinomial logistic regression

6.2.3. Data Model using WEKA

WEKA is a collection of machine learning algorithms for data mining tasks. WEKA holds tools for data analysis tasks such as pre-processing, classification, regression, clustering, association rules, and visualization.

6.2.3.1. Classification as the data mining technique

Data mining can be divided into two tasks called predictive tasks and descriptive tasks. Predictive analytics is the section in data mining which consider about with forecasting probabilities and trends. A predictive model consists of a number of predictors, which are variable factors that are likely to affect future behavior or results. In predictive modeling, data is collected for the relevant forecasters, a statistical model is constructed, then forecasts are made and the model is validated as additional data becomes available.

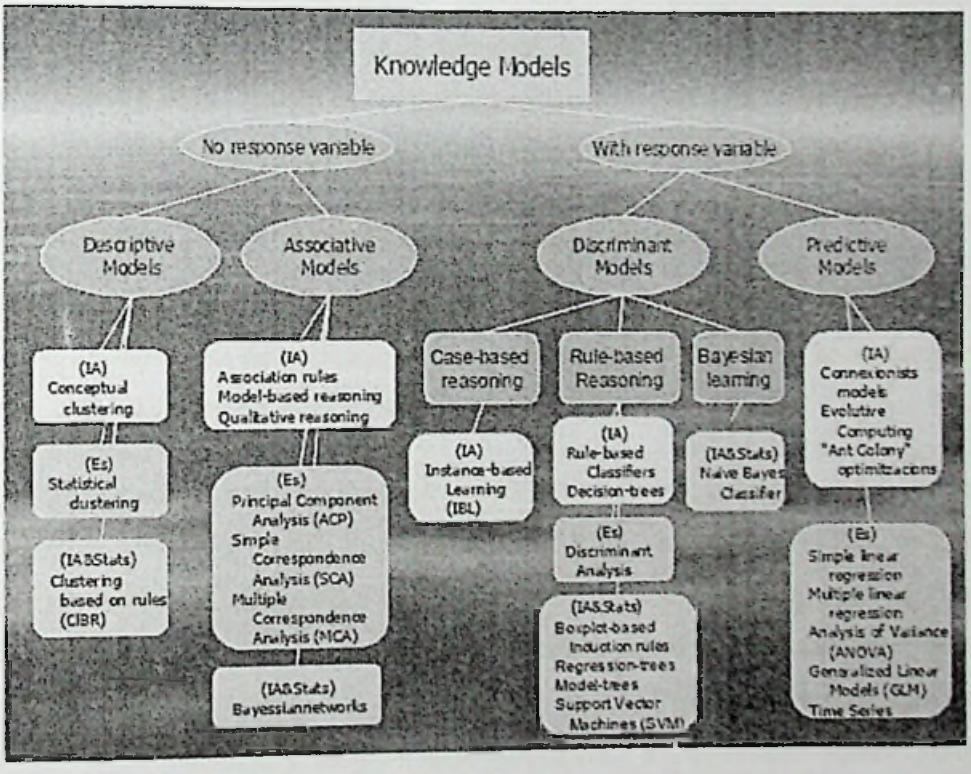

Figure 6.1: Different categories in data models

Knowledge models have basic difference between having and not having a reference variable to be described. Left part of the Figure 6.1 refers to non-supervised methods, without response variable, in where the main goal is a better reasoning of the target phenomenon and explanation is enough as a result. Whereas right part of the Figure 6.1 refers those supervised models concerned with to re-cognition, where a response variable is to be described and prediction is to be made. Among methods oriented to prediction main distinction is the nature of the response variable. While discriminant

methods describe or foresee qualitative variables, the classical predictive methods refer to quantitative response variables.

In here for predicting the factors that influence no schooling or dropouts we use predictive data mining. For No schooling and dropouts, separately predictors are selected from HIES dataset. Moreover, this research used classification which contains further level of subdivision as the predictive technique. Rule-based reasoning methods provide explicit knowledge models, which can be expressed by formal rules that applied for further prediction. In case-based reasoning methods the predictive models are implicit in historical data. The third option bayesian learning is a combination of prior explicit knowledge model and iterative enhancements based on future data. One technique from each paradigm is used to anayse citizen profiles. Selected classification techniques are J48 algorithm from decision tree, naïve bayes from Bayesian network classifier and IBk algorithm from K-nearest neighbours.

6.2.3.2. Decision Tree for School dropouts and no-schooling

This technique as well as Bayesian network classifier required numeric to nominal conversion of attributes and binning of continues attribute values to make algorithms working with improved efficiency (Appendix C).

J48 classifier model is a pruned decision tree in textual form that was produced on the full training data. For school dropouts there is only one split which is on the 'Is_Active' attribute which denote whether child is actively participate for the economic activities. In this case, 92.93% of 16931 training instances have been classified correctly. This indicates that the results obtained from the training data are optimistic. More specifically, model's mean output error and root mean squared error are 0.13 and 0.25 respectively. The reason why the errors are nearly 1 or 0 is correctness of its classification.

In J48 classifier model built for no schooling, the first split is on the 'Is_Active' attribute. Then split on 'Religion' which produces tree size of eight according to the number of nodes in the tree. Decision tree for no schooling classified 98.98% of 15090 training instances correctly which meant that results are optimistic. Moreover in the model, mean output error is 0.01 and root mean squared error is 0.09 which ultimately prove that most of the training instances are classified correctly.

6.2.3.3. Bayesian network classifier for School dropouts and no-schooling

Naive Bayes classifier gives the count of every feature according to class. In this case for school dropouts, 92.40% of 16931 training instances have been classified correctly which is acceptable for a good model. Also model has very low mean output error and root mean squared error which takes the value 0.09 and 0.22 which indirectly prove that correctness of the model is high.

Bayesian network for no-schooling classified 99.02% of 15090 training instances correctly. This indicates that the results obtained from the training data are positive. Furthermore, mean output error is 0.01 and the root mean squared error is 0.09 which support the fact that most of the training instances are classified correctly.

6.2.3.4. K-nearest neighbours for School dropouts and no-schooling

KNN is a method which uses to calculate distance from attributes values in model building process. Therefore class label is converted to nominal value before building the model (Appendix C).

In this case 91.78% of 16931 training instances have been classified correctly for school dropouts. This also indicates that the results are optimistic. In addition model shows a mean output error of 0.08 and root mean squared error of 0.28.

Similarly, K-nearest neighbours classified 98.75% of 15090 training instances correctly for no-schooling. Also that model has only outputs mean output error of 0.01 and root mean squared error of 0.11. Therefore most of the training instances are classified correctly.

6.3. Solution for Sub Research Question 2:

As the solution for the second sub research question explore the relationship between chronic diseases and family income using following steps.

6.3.1. Existing work in this domain

This study of Chung and co-workers examine the relationship between chronic diseases and economic outcome [10]. This work has studied socioeconomic status on the general health of individuals. Moreover, Chung and others have identified cancer, heart attack, heart disease, lung disease, stroke, arthritis, diabetes, hypertension,

psychological problems, asthma, memory loss, learning disability as some of the major chronic diseases which have high impact on family income. However, this study does not focus on patterns on how chronic diseases affect income.

There is a similar study on this issue in US [31]. This study has focused on impact of health status on economic status statistically. Nevertheless, this has concentrated on health and economic status the near elderly population without working-age population. But, it is important to consider family income for a working-age sample by considering the impact of chronic diseases.

This situation is crucial in developing countries such as Sri Lanka. However, there are no enough studies on this issue in Sri Lankan context. Therefore, it is important to focus on impact of chronic diseases on family income to uplift the quality of the life. By Considering these existing works and data available in HIES dataset, initially total income for the family is calculated and with the chronic diseases individual members having in the family to build the model.

6.3.2. Theoretical Framework

As initial attribute monthly income is calculated using SPSS tool for each family by considering main income, main agricultural income, other agricultural income, non-agricultural income and other income attributes in the HIES dataset. Then chronic disease attribute is integrate with the data model.

6.3.3. Data Modeling

6.3.3.1. Data Preprocessing

After combining different types of income to take the total income we had to remove the outliers. According to central bank report income distribution mean income of the lowest category was Rs. 6700 and mean income for the highest category was Rs. 174,376[9].Hence, by assuming others are outliers remove them and set the range of total income in between 6700-174,376. Also for missing data remove those tuple. For an example if income component are missing code or if chronic disease type is missing remove those tuples.

6.3.3.2. Clustering as the data mining technique

Simply, clustering can be defined as finding groups of objects such that the objects in a group will be similar or related to one another and different from or unrelated to the objects in other groups. Two main types of Clustering methods can be seen in the field of data mining: hierarchical method and partitioning methods [30]. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. For this research we have used three algorithms belonged to portioning methods which are popular.

6.3.3.2.1. Clustering using Kmean Algorithm

In KMean clustering main idea is to define k number of centers, one for each cluster. In there, try to place center as much as possible far away from each other. The next step in there is to take each point belonging to a given data set and associate it to the nearest center. The first step is completed when no point is pending. Then again calculate k new centroids as center of the clusters taking instances from the previous step. After having k new centroids, a new binding has to be done between the same data set instances and the nearest new center. Iteratively this has been done. Due to this cycle, we can identify the k centers that change their location step by step until no more changes are done [30].

When measuring cluster validity several numerical measures are applied to judge various aspects of cluster validity. As an internal index, Sum of Squared Error (SSE) is used in there. SSE is used to measure the goodness of a clustering structure without respect to external information. In here we choose 7 clusters where SSE is 6.78526048886785.When go beyond more than seven clusters we can only see a small change in Sum of Squared Error. Hence, seven is decided as the optimum number of clusters is all these three algorithms.

Through K-Mean algorithm we have identified seven clusters. First cluster contain roughly income range from 6700 to 17500 and have most of the chronic diseases and Blood pressure as the major disease. People with roughly family income range 17500 to 29000 belong to the second cluster and those people also having most of the diseases. When consider the second cluster, Diabetics dominate that one. In the third

cluster, we can see people with roughly income range 29000-43450 and those people are having Blood pressure, Diabetics, Asthma, Heart Conditions/Diseases and Arthritis. In the third cluster also Blood pressure got the highest number of instances. We can see four more clusters which are range from 43450-63250, 63250-87000, 87000-122400 and 125000-170800 where having Blood pressure, Diabetics, Asthma and, Blood pressure is the major disease in all these clusters. (Appendix E).

6.3.3.2.2. Clustering using Expectation-Maximization

Expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood. In EM, "the data set is usually modeled with a number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set [30]".

Through EM algorithm we have identified seven clusters. First cluster contain roughly income range from 6700 to 12800 and have most of the chronic diseases and Blood pressure as the major disease. People with roughly family income range 12800 to 26500 belong to the second cluster and those people also having most of the diseases and Blood pressure as the major disease. In the third cluster, we can see people with roughly income range 26500-42300 and those people are also having Blood pressure, Diabetics, Asthma, Heart Conditions/Diseases and Arthritis. In the third cluster also Blood pressure got the highest number of instances. We can see four more clusters which are range from 42300-62000, 62000-87000, 87000-130500 and 130500-170800 where having Blood pressure and Diabetics. Blood pressure is the major disease in all these clusters except in last one where it has Diabetics. (Appendix E).

6.3.3.2.3. Clustering using MakeDensityBasedClusterer Algorithm

Density based clustering algorithm has played a strong role in finding nonlinear shapes structure based on the density. The intention of these approaches is to identify the clusters and their distribution parameters. "This method can be used for discovering clusters of arbitrary shape which are not necessarily convex [6]".

Through MakeDensityBasedClusterer algorithm we have identified seven clusters. First cluster contain roughly income range from 6700 to 17900 and have most of the chronic diseases and Blood pressure as the major disease. People with roughly family income range 17900 to 29200 belong to the second cluster and those people also

having most of the diseases and Diabetics is the major disease. In the third cluster, we can see people with roughly income range 29200-43250 and those people are also having Blood pressure, Diabetics, Asthma, Heart Conditions/Diseases and Arthritis. In the third cluster also Blood pressure got the highest number of instances. We can see four more clusters which are range from 43250-62000, 62000-87100, 87000-123400 and 123400-170200 where having Blood pressure, Diabetics, Asthma and, Blood pressure is the major disease in all these clusters (Appendix E).

6.4. Summary

Implementation chapter provide the full path in constructing data models for addressing research sub questions. Furthermore this chapter gives detail description about using SPSS and WEKA tool to build the data model and attribute selection. Next chapter will be on discussion about evaluation.

Chapter 7
Evaluation

7.1. Introduction

This chapter focuses on how testing strategies carried out for the research sub question in terms of the evaluation measurements for the selected data mining technique such as percentage of accuracy, TP rate and ROC area for classification.

7.2. Evaluation for Classification

For evaluating a classifier quality we can use confusion matrix. The confusion matrix helps us to find the various evaluation measures such as accuracy, recall, and precision to evaluate data mining classifiers. These measurements and their definition are given in following Table 7.1.

Measure	Formula	Intuitive Meaning
Precision	TP / (TP + FP)	The percentage of positive predictions that are correctly predicted.
Recall / Sensitivity	TP / (TP + FN)	The percentage of positive labeled instances that were predicted as positive labels.
Specificity	TN / (TN + FP)	The percentage of negative labeled instances that were predicted as negative labels.
Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions those are correct.

Table 7.1: Evaluation measurements for classifiers

Using aforesaid measurements classifiers provide few set of measures called TP rate, FP rate, F-measure and ROC area. TP rate is equal to sensitivity, while FP rate equal to one minus specificity. F-measure calculated by precision and recall. The area under a ROC curve quantifies the overall ability of the test to discriminate between usefulness and uselessness. A truly useless test has an area of 0.5. A perfect test has an area of 1.00. Usually better models are having higher TP-rate, lower FP rate and ROC space close to 1.00.

Comparison of the confusion matrixes and the weighted averages in the classification model used for school dropout scenario in sub research question 1 are given in the following Table 7.2.

Technique	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
Decision Tree- J48	0.929	0.56	0.935	0.929	0.914	0.684
Bayesian Network Classifier- Naïve Bayes	0.924	0.354	0.921	0.924	0.922	0.948
K-nearest neighbors–IBk	0.918	0.342	0.917	0.918	0.917	0.81

Table 7.2: comparison of different classification methods to determine school dropout

Table 7.3 shows that Naïve Bayes and IBk have almost equal accuracy measures, but IBk having lower value than Naïve Bayes. This is critical in ROC Area measure in which Naïve Bayes has higher accuracy on the school dropout dataset. So, Naïve Bayes is better method for Schooling-dropouts dataset. Furthermore algorithm IBk having almost lower value than J48 except for ROC Area. So J48 is better method for school dropout dataset. Moreover, algorithm Naïve Bayes having lower value than J48 for FP Rate and higher value for ROC Area. So Naïve Bayes is better than J48 method for Schooling-dropouts dataset. Hence, performance of methods can be mentioned in ascending order as Ibk , J48 and Naïve Bayes.

In schooling-noschooling dataset accuracy parameters have shown in Table 7.3. According to the table, J48 and Naïve Bayes have almost equal accuracy measures except ROC Area measure in which Naïve Bayes has higher accuracy on the schooling-noschooling dataset. So, Naïve Bayes is the better method for schooling-noschooling dataset. Moreover algorithm IBk having lower value than J48 except for ROC Area. So J48 is better method for schooling-noschooling dataset. Also algorithm IBk having lower value than Naïve Bayes except for ROC Area. So Naïve Bayes is better method for schooling-noschooling dataset. Thus, performance of methods can be mentioned in ascending order as Ibk, J48 and Naïve Bayes for schooling-noschooling data model.

Technique	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
Decision Tree- J48	0.99	0.931	0.985	0.99	0.986	0.53
Bayesian Network Classifier- Naïve Bayes	0.99	0.931	0.986	0.99	0.986	0.829
K-nearest neighbors– IBk	0.988	0.852	0.984	0.988	0.986	0.85

Table 7.3: comparison of different classification methods to determine no-schooling

7.3. Evaluation for Clustering

In clustering, time to build the model and within cluster sum of squared errors measurement are used to evaluate a data model [19]. To evaluate the accuracy of data model, datasets is deployed in Weka tool and then clustering algorithms are applied to the dataset with classes to cluster evaluation option.

To decide the number of clusters analyses the within cluster sum of squared errors (SSE) measurement as in Table 7.4 by using KMean algorithm. Then project the number of clusters with respect to SSE and when SSE become stabilized as in Figure 7.1, that number of clusters chosen as the number of optimum clusters. Hence, 7 clusters were selected as the number of optimum clusters in determining relationship between chronic diseases and family income.

Within cluster sum of squared errors (SSE)	Number of Clusters
75.3412485560538	2
36.08672870118752	3
21.59519383130758	4
13.117956741879242	5
9.356318638835468	6
6.785260488867854	7
5.491883010809618	8
4.790596279407836	9
3.698849855672156	10

Figure 7.4: SSE vs. Number of clusters

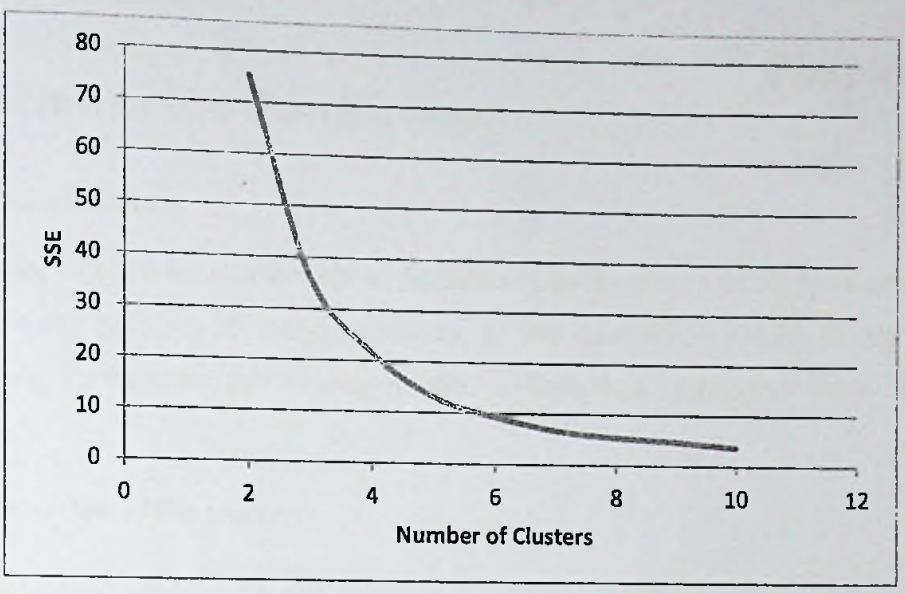

Figure 7.4: Scattered graph of SSE vs. Number of clusters

The time complexity for K-means, EM and MakeDensityBasedClusterer are shown in the Table 7.5 .It is shown that K-means has required minimum time to make cluster for the datasets in comparison with other clustering algorithms. Hence, K-means has maximum accuracy.

Algorithm	Kmean	EM	MakeDensityBasedClusterer
Time(seconds)	0.52	8.23	0.54

Table 7.5: Time taken by clustering algorithms to make clusters for given data set

7.4. Summary

This chapter concludes with test results used to evaluate the data model. Final chapter will summarize the overall research and highlights the significance findings of the research.

Chapter 8

Conclusion and Further work

8.1. Introduction

This chapter provides an overview of the research and how we provide the solution to address the problem of analyzing citizen profile data which belong to big data category. Furthermore this chapter focuses of limitations and further work of this research.

8.2. Overview of the research

By analyzing citizen profiles of any country including Sri Lanka we can find out the issues inherent with their citizen life style. Life style describes the way of living of individuals with in households on day-to-day life. Both Asian and European lifestyles mainly focus on areas such as income, education, health, communication and transportation. Sri Lanka also consider these areas as important fields in describing their citizens lifestyles' according to its' main financial reports and socio-economic indicators and collected data related to them within government institutes. But proper analysis was not carried out by the government by using these data to uplift life quality. Around the world, there are many examples where developing countries as well as developed countries use lifestyle data to identify root causes for a particular social or economic issue. There are some case study observations in Sri Lankan context about these socio-economic issues but realistic solutions are completely missing and existing researches are not adequate.

Hence, as a developing country Sri Lankan government must set appropriate future development plans to uplift citizen's life quality by analyzing socio-economic factors carefully. Furthermore socio-economic segmentation should be analyzed by paying attention to find out hidden pattern behind lifestyle data. To achieve these objectives, Household Income and Expenditure Survey (HIES) dataset is used here to analyze deeply as it cover core fields in lifestyles targeted by Sri Lanka.

When comparing existing solutions given around the world which are most of the time statistical based, data mining is identified as a novel approach to analyse with its ability to analyse big data set dynamically and effectively. Once, data mining is

identified as a best approach to find out hidden patterns within this household dataset, different data mining techniques in data mining have been used to address the sub research questions. Major paradigms in data mining are predictive techniques and descriptive techniques according to the output we try to achieve. To determine the accuracy of the drawn solution different algorithm within the selected techniques are used and compared their efficiency before selecting the best approach to made the conclusions.

8.3. Problem encountered & limitations

This is a secondary research where we used sample data set from Household Income and Expenditure Survey (HIES) of Department of census and statistics. Hence, number of research problems we can focus on is limited according to the data available with HIES. Also sample size is not enough to give the conclusions in district level.

Clustering solution used to explore the relationship between chronic diseases and family income shows a less accuracy level when used continues data as family income. In the clustering solution, to improve the accuracy level categorical data can be used instead of continues data. However, when used categorical data for the income we cannot directly identify the distinct clusters within that data set.

8.4. Further work

This research only address two sub research questions attach with citizen profiles. Those two questions focus on demographic data, education and health basically. According to the analytical diagram in Figure 5.1, citizen profiles are based on housing, income and expenditure in addition to the areas addressed by this research. By addressing those areas can find the hidden pattern such as in power consumption, water demand and demand for different services. For those problems association rules, classification or clustering like different techniques can be used to according to the output we try to achieve. Hence, this research can be extended further to address different issues in Sri Lankan lifestyle.

8.5. Summary

This chapter concludes the thesis by describing the solution given with data mining to analyze the lifestyle data and how it can be enhance further to improve the level of accuracy in predicating /exploring lifestyle data to analyze issues attach with citizens' lifestyle.

Reference

[1] Adnan, M.H.M., Husain, W., Rashid, N.A.A, (2012). Data Mining for Medical Systems: A Review. International Conference on Advances in Computer and Information Technology. , pp.17-22

[2] Ahmadvand, A. M., Bidgoli, B. M., & Akhondzadeh, E. (2010, January). A hybrid data mining model for effective citizen relationship management: a case study on Tehran municipality. In e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E'10. International Conference on (pp. 277-281). IEEE

[3] Asian Development Bank.(2013).Asian Development Bank Annual Report 2013. [ONLINE] Available at: http://www.adb.org/sites/default/files/institutional-document/42741/adb-annual-report-2013.pdf. [Accessed 09 July 15].

[4] Atkinson, B., & Marlier, E. (2010). Income and living conditions in Europe

[5] Ayanso, A., Lertwachara, K., & Vachon, F. (2011). Design and behavioral science research in premier IS journals: evidence from database management research. In Service-oriented perspectives in design science research (pp. 138-152). Springer Berlin Heidelberg

[6] Berkhin, P. (2006). A survey of clustering data mining techniques. In*Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

[7] Business Intelligence Solutions.(2015). Data Mining vs. Statistics. [ONLINE] Available at: http://www.bisolutions.us/Data-Mining-vs-Statistics.php. [Accessed 09 July 15].

[8] Cardoso, A. R., & Verner, D. (2006). School drop-out and push-out factors in Brazil: The role of early parenthood, child labor, and poverty.

[9] Central Bank of Sri Lanka.(2013). Central Bank of Sri Lanka Annual Report 2013.[ONLINE] Available at: http://www.cbsl.gov.lk//pics_n_docs/10_pub/_docs/efr/annual_report/ar2013/english/content.htm. [Accessed 09 July 15].

[10] Chung, Y., 2013. Chronic Health Conditions and Economic Outcomes.

[11] Department of Census and Statistics. (2013). Household Income and Expenditure Survey -2012/2013 Final Results. [ONLINE] Available at:

http://www.statistics.gov.lk/HIES/HIES200213FinalBuletin4.pdf. [Accessed 09 July 15].

[12] Ec.europa.eu, (2015). [online] Available at: http://ec.europa.eu/eurostat/documents/3217494/5722557/KS-31-10-555-EN.PDF/e8c0a679-be01-461c-a08b-7eb08a272767 [Accessed 11 Nov. 2015].

[13] Einsele, F., Sadeghi, L., Ingold, R., & Jenzer, H. (2015). A Study about Discovery of Critical Food Consumption Patterns Linked with Lifestyle Diseases using Data Mining Methods

[14] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining

[15] Fernandez-Villaverde, J., & Krueger, D. (2007). Consumption over the life cycle: Facts from consumer expenditure survey data. The Review of Economics and Statistics, 89(3), 552-565

[16] Hamel, L., & Hall, T. (2005).A brief tutorial on database queries, data mining, and OLAP. The Encyclopedia of Data Warehousing and Mining, 401.

[17] Hildebrandt, M., & Gutwirth, S. (2008). Profiling the European citizen. Heidelberg: Springer

[18] Jiang, S., Ferreira, J., & González, M. (2013, January). ANALYZING HOUSEHOLD LIFESTYLES, MOBILITY AND ACTIVITY PROFILES: A CASE STUDY OF SINGAPORE. 92nd Annual Meeting

[19] Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms.*Biotechnology & Biotechnological Equipment*, *28*(sup1), S44-S48.

[20] Junling, H. (2013). About Data Mining: Data Mining vs. Machine Learning. [ONLINE] Available at: http://www.aboutdm.com/2013/02/data-mining-vs-machine-learning.html. [Accessed 09 July 15]

[21] Khan, M. A., Islam, Z., & Hafeez, M. (2011, December). Irrigation water demand forecasting: a data pre-processing and data mining approach based on spatio-temporal data. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 183-194). Australian Computer Society, Inc.

[22] Koh, H. C.,Tang,G., (2011). Data Mining for Medical Systems: A Review. Journal of Healthcare Information Management. 19 (2), pp.64-72 -19

[23] Lee, H. (2007). Essentials of Behavioral Science Research

[24] Mahrsi, M. K. E., Etienne, C. O. M. E., Johanna, B. A. R. O., & Oukhellou, L. (2014, January). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. In ACM SIGKDD Workshop on Urban Computing (p. 9p)

[25] Mina, C. D., & Barrios, E. B. (2009). Profiling poverty with multivariate adaptive regression splines. Philippine Institute for Development Studies.

[26] Okumu, I. M., Nakajjo, A., & Isoke, D. (2008). Socioeconomic determinants of primary school dropout: the logistic model analysis

[27] Orodho, A. J., & Kombo, D. K. (2002). Research methods. Nairobi: Kenyatta University, Institute of Open Learning

[28] Publications, S. (2015). Home &middot Science Publications. [online] Thescipub.com. Available at: http://thescipub.com/PDF/ajassp.2009.2036.2042.pdf [Accessed 11 Nov. 2015].

[29] Rahman, H. (Ed.). (2008). Data mining applications for empowering knowledge societies. IGI Global.

[30] Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer US

[31] Smith, J.P., 1999. Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status [WWW Document]. URL https://www.aeaweb.org/articles.php?doi=10.1257/jep.13.2.145 (accessed 1.24.16).

[32] UNESCO Institute for Statistics. (2014) GUIDE TO THE ANALYSIS AND USE OF HOUSEHOLD SURVEY AND CENSUS EDUCATION DATA. [ONLINE] Available at: www.uis.unesco.org/Library/Documents/hhsguide04-en.pdf. [Accessed 09 July 15].

[33] UNICEF, (2013). Out of School Children in Sri Lanka, Summary Report. [online] UNICEF Sri Lanka. Available at: http://www.unicef.org/srilanka/2013_OSS_Summery_E.pdf [Accessed 10 Nov. 2015]

[34] Warc.com, (2015). Lifestyle segmentation of the Chinese consumer. [online] Available at: http://www.warc.com/fulltext/esomar/80217.htm [Accessed 11 Nov. 2015].

[35] Wirth, R., &Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29-39)

[36] World Bank Group. (2015). Assessing Sector Performance and Inequality in Education –Chapter 2. [ONLINE] Available at: http://siteresources.worldbank.org/EXTEDSTATS/Resources/3232763-1252439241095/ADePTBookChap-2.pdf [Accessed 09 July 15].

Appendix A

Data Preprocessing

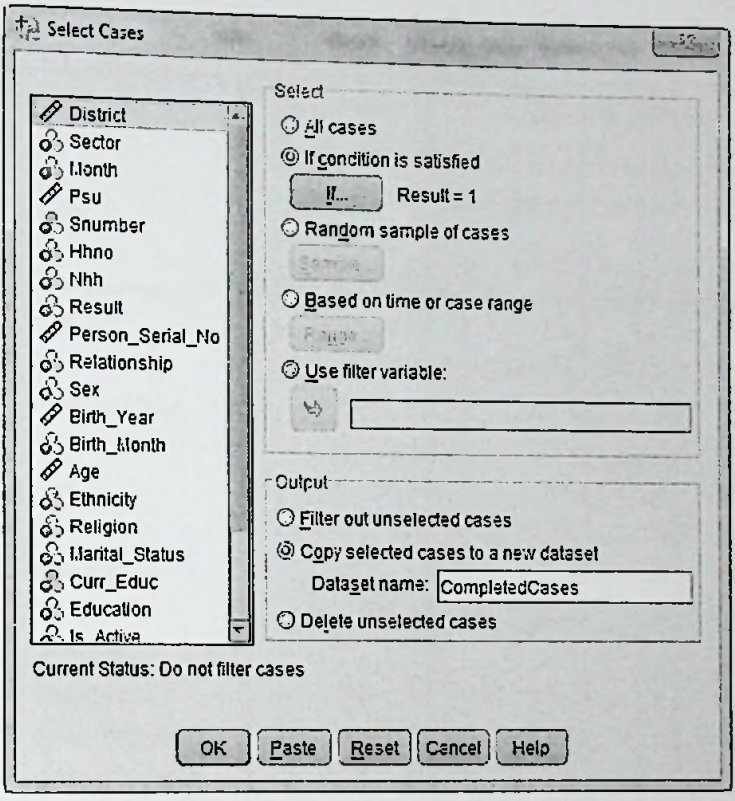

Figure 9.1: ignoring the tuple

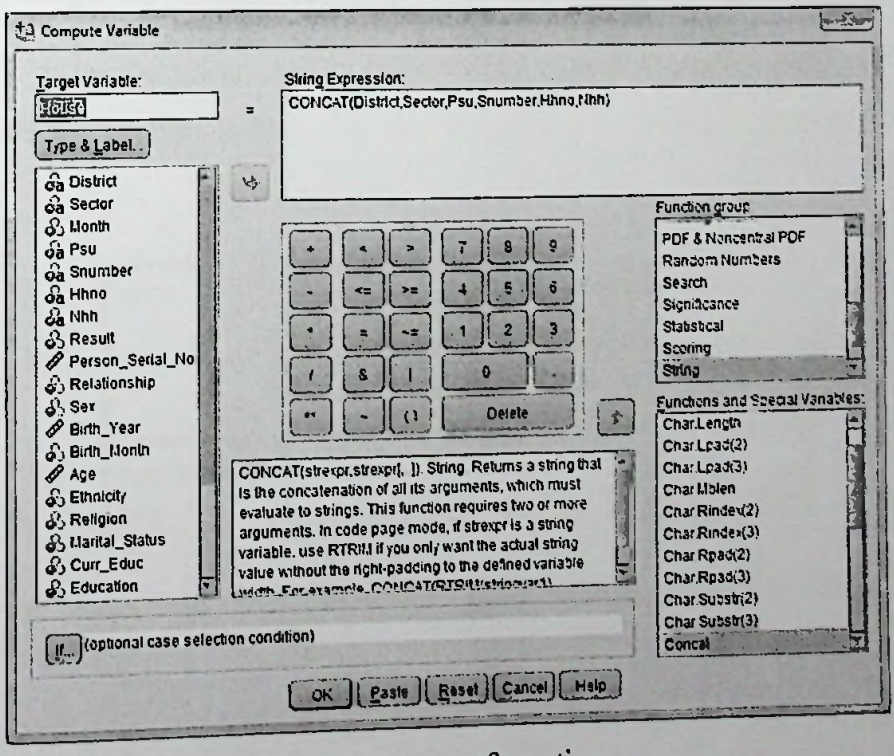

Figure 9.2: data transformation

Appendix B

Attribute Selection using SPSS

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	5776.207[a]	.000	0	.
Age	9245.118[b]	3468.912	2	.000
HouseholdSize	5782.583[b]	6.376	2	.041
District	5889.220[b]	113.013	48	.000
Sector	5778.631[b]	2.425	4	.658
Sex	5780.354[b]	4.147	2	.126
Ethnicity	5784.428[b]	8.221	12	.768
Religion	5789.936[b]	13.730	8	.089
Is_Active	6603.556[b]	827.350	2	.000
HeadSex	5785.470[b]	9.263	2	.010
FatherEducation	5898.866[b]	122.659	42	.000
MotherEducation	5977.752[b]	201.545	42	.000
income	5779.287[b]	3.080	4	.545
Structure	5803.527[b]	27.321	20	.126
Natural_Calamity	5777.901[b]	1.694	2	.429
Is_Ill_Disable11	6036.825[b]	260.618	2	.000

Figure 9.3: Likelihood Ratio Test considering schooling and dropouts both

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	13677.353			
Final	5776.207	7901.146	194	.000

Figure 9.4: Model fitting information to represent statistical significance of model

Appendix C

Preprocessing with WEKA

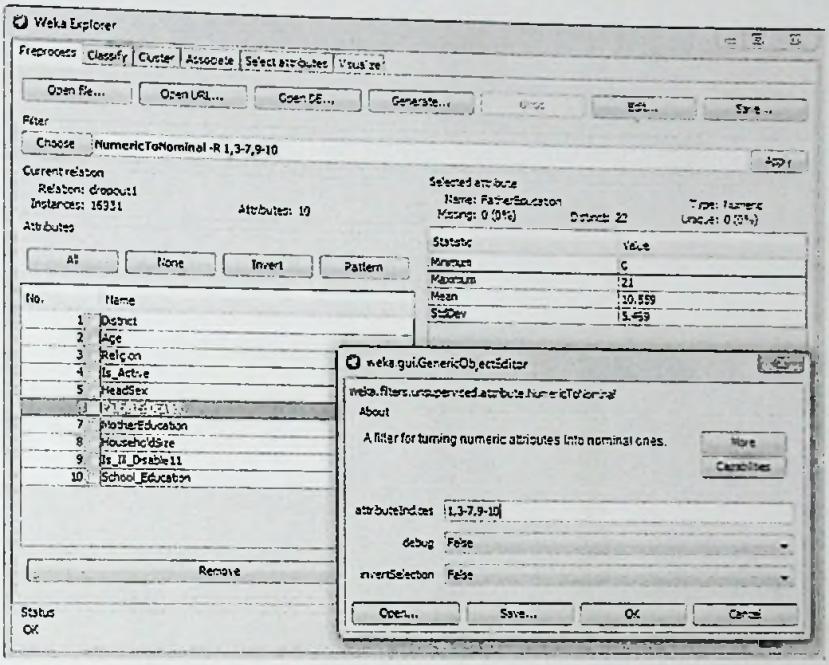

Figure 9.5: Preprocessing with filters

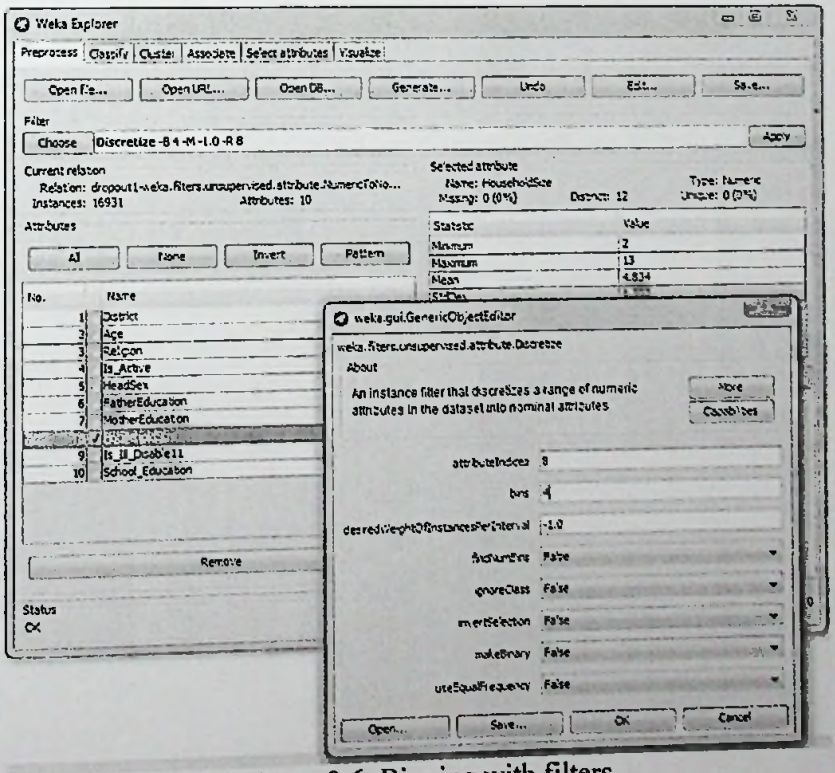

Figure 9.6: Binning with filters

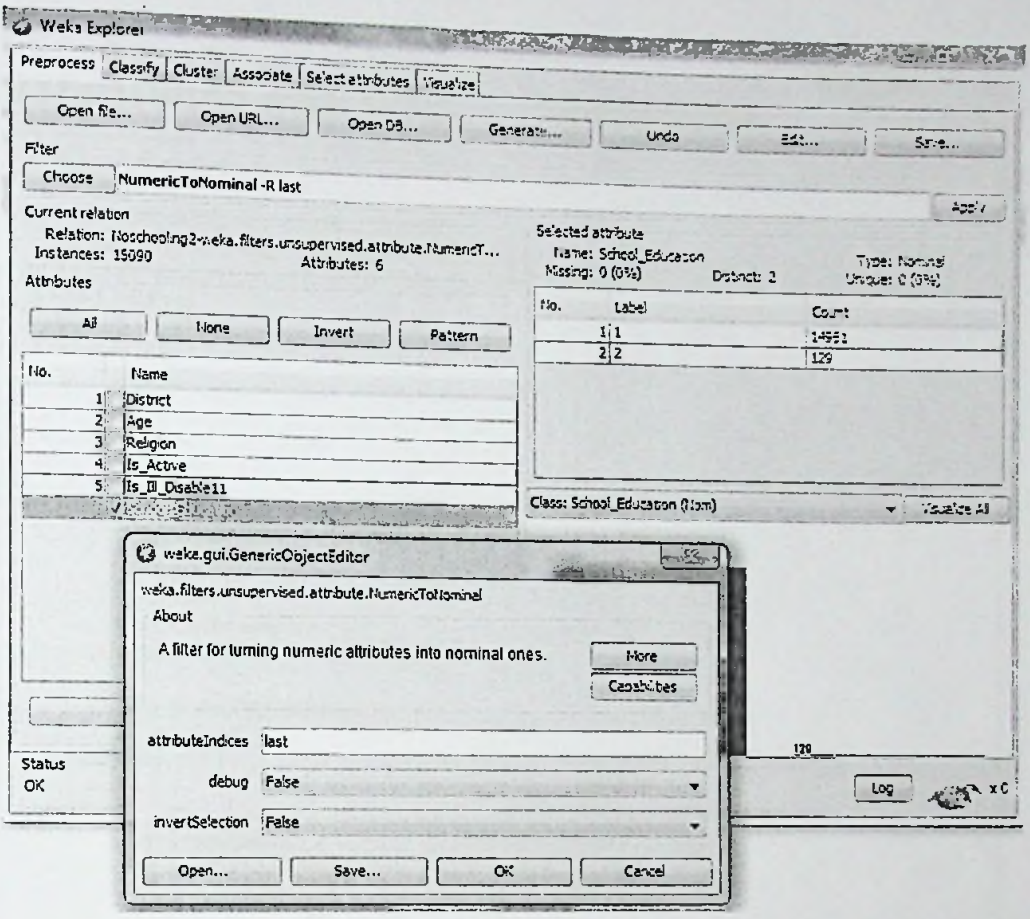

Figure 9.7: Convert class label to nominal for KNN

Appendix D

Data Mining with WEKA-Classification

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     dropout1-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-7,9-10-
weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Discretize-
B4-M-1.0-R8
Instances:    16931
Attributes:   10
              District
              Age
              Religion
              Is_Active
              HeadSex
              FatherEducation
              MotherEducation
              HouseholdSize
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

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

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

Is_Active = 1: 3 (709.0/9.0)
Is_Active = 2: 1 (16222.0/1270.0)

Number of Leaves  :         2

Size of the tree :    3


Time taken to build model: 0.31 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5350               92.9303 %
Incorrectly Classified Instances       407                7.0697 %
Kappa statistic                          0.5094
Mean absolute error                      0.1374
Root mean squared error                  0.256
Relative absolute error                 67.3125 %
Root relative squared error             81.1504 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   1         0.631     0.926     1         0.962     0.684     1
                   0.369     0         1         0.369     0.539     0.684     3
Weighted Avg.      0.929     0.56      0.935     0.929     0.914     0.684

=== Confusion Matrix ===

   a   b   <-- classified as
5112   0 |   a = 1
 407 238 |   b = 3
```

Figure 9.8: Decision Tree for School Dropouts

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: dropout1-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-7,9-10-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2-weka.filters.unsupervised.attribute.Discretize-B4-M-1.0-R8
Instances: 16931
Attributes: 10
District
Age
Religion
Is_Active
HeadSex
FatherEducation
MotherEducation
HouseholdSize
Is_Ill_Disable11
School_Education
Test mode:split 66.0% train, remainder test

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

Naive Bayes Classifier

Attribute	Class 1 (0.88)	3 (0.12)
District		
11	1384.0	160.0
12	1233.0	169.0
13	831.0	89.0
21	701.0	92.0
22	404.0	65.0
23	555.0	51.0
31	919.0	133.0
32	845.0	90.0
33	507.0	70.0
41	573.0	82.0
42	261.0	44.0
43	221.0	33.0
44	241.0	42.0
45	393.0	57.0
51	674.0	135.0
52	708.0	96.0
53	478.0	46.0
61	737.0	74.0
62	497.0	71.0
71	530.0	61.0
72	336.0	50.0
81	572.0	65.0
82	396.0	50.0
91	529.0	112.0
92	461.0	58.0
[total]	14986.0	1995.0

Age

'(-inf-10.333333]'	6443.0	17.0
'(10.333333-15.666667]'	5768.0	125.0
'(15.666667-inf)'	2753.0	1831.0
[total]	14964.0	1973.0

Religion

1	8525.0	910.0
2	3087.0	544.0
3	2059.0	312.0
4	1290.0	206.0
9	5.0	3.0
[total]	14966.0	1975.0

Is_Active

1	10.0	701.0
2	14953.0	1271.0
[total]	14963.0	1972.0

HeadSex

1	12579.0	1600.0
2	2384.0	372.0
[total]	14963.0	1972.0

FatherEducation

0	49.0	24.0
1	175.0	51.0
2	319.0	113.0
3	445.0	117.0
4	629.0	135.0
5	1012.0	199.0
6	538.0	118.0
7	744.0	126.0
8	1179.0	151.0
9	793.0	101.0
10	3152.0	248.0
11	1309.0	74.0
12	601.0	23.0
13	1265.0	49.0
14	24.0	1.0
15	261.0	8.0
16	99.0	1.0
17	5.0	1.0
18	3.0	1.0
19	273.0	117.0
20	3.0	1.0
21	2105.0	333.0
[total]	14983.0	1992.0

MotherEducation

0	51.0	15.0
1	121.0	36.0
2	303.0	108.0
3	330.0	86.0
4	530.0	174.0
5	781.0	215.0
6	542.0	137.0

```
7                634.0   124.0
8                843.0   122.0
9                951.0   123.0
10               4172.0  358.0
11               1660.0  101.0
12               772.0   35.0
13               2000.0  55.0
14               32.0    1.0
15               288.0   1.0
16               79.0    1.0
17               3.0     1.0
18               2.0     1.0
19               416.0   186.0
20               4.0     1.0
21               469.0   111.0
[total]          14983.0 1992.0

HouseholdSize
'(-inf-4.75]'    6914.0  802.0
'(4.75-7.5]'     7491.0  1032.0
'(7.5-10.25]'    519.0   133.0
'(10.25-inf)'    41.0    7.0
[total]          14965.0 1974.0

Is_Ill_Disable11
1                457.0   97.0
2                14506.0 1875.0
[total]          14963.0 1972.0


Time taken to build model: 0.04 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5320               92.4092 %
Incorrectly Classified Instances      437                7.5908 %
Kappa statistic                       0.5992
Mean absolute error                   0.0949
Root mean squared error               0.2278
Relative absolute error               46.4906 %
Root relative squared error           72.2145 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.964     0.394     0.951       0.964   0.958       0.948     1
                 0.606     0.036     0.681       0.606   0.642       0.948     3
Weighted Avg.    0.924     0.354     0.921       0.924   0.922       0.948

=== Confusion Matrix ===

   a   b   <-- classified as
4929  183 |   a = 1
 254  391 |   b = 3
```

Figure 9.9: Naïve Bayes Classifier for School Dropouts

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A
\"weka.core.EuclideanDistance -R first-last\""
Relation:     dropout1-weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances:    16931
Attributes:   10
              District
              Age
              Religion
              Is_Active
              HeadSex
              FatherEducation
              MotherEducation
              HouseholdSize
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5284               91.7839 %
Incorrectly Classified Instances       473                8.2161 %
Kappa statistic                          0.5823
Mean absolute error                      0.0838
Root mean squared error                  0.2862
Relative absolute error                 41.0367 %
Root relative squared error             90.7335 %
Total Number of Instances             5757

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.955    0.38       0.952      0.955    0.954      0.81      1
                   0.62     0.045      0.637      0.62     0.628      0.81      3
Weighted Avg.      0.918    0.342      0.917      0.918    0.917      0.81

=== Confusion Matrix ===

   a   b   <-- classified as
4884 228 |   a = 1
 245 400 |   b = 3
```

Figure 9.10: K-nearest neighbours for School Dropouts

```
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-6-
weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last
Instances:    15090
Attributes:   6
              District
              Age
              Religion
              Is_Active
              Is_Ill_Disable11
              School_Education
Test mode:split 66.0% train, remainder test

=== Classifier model (full training set) ===
J48 pruned tree
------------------
Is_Active = 1
|   Religion = 1: 1 (8.0/1.0)
|   Religion = 2: 2 (11.0/2.0)
|   Religion = 3: 2 (1.0)
|   Religion = 4: 2 (0.0)
|   Religion = 9: 2 (0.0)
Is_Active = 2: 1 (15070.0/118.0)

Number of Leaves  :         6
Size of the tree :    8

Time taken to build model: 0.14 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5079               98.9866 %
Incorrectly Classified Instances        52               1.0134 %
Kappa statistic                          0.101
Mean absolute error                      0.0171
Root mean squared error                  0.0986
Relative absolute error                 96.7818 %
Root relative squared error            100.3851 %
Total Number of Instances             5131

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999    0.94      0.991     0.999     0.995      0.53     1
                 0.06     0.001     0.375     0.06      0.103      0.53     2
Weighted Avg.    0.99     0.931     0.985     0.99      0.986      0.53

=== Confusion Matrix ===
   a    b   <-- classified as
 5076    5 |    a = 1
   47    3 |    b = 2
```

Figure 9.11: Decision Tree for No schooling

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation: Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-R1,3-6-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R2
Instances: 15090
Attributes: 6
District
Age
Religion
Is_Active
Is_Ill_Disable11
School_Education
Test mode:split 66.0% train, remainder test

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

Naive Bayes Classifier

Attribute	Class 1 (0.99)	2 (0.01)
District		
11	1384.0	14.0
12	1233.0	10.0
13	831.0	6.0
21	701.0	2.0
22	404.0	5.0
23	555.0	7.0
31	919.0	13.0
32	845.0	13.0
33	507.0	6.0
41	573.0	2.0
42	261.0	1.0
43	221.0	5.0
44	241.0	1.0
45	393.0	2.0
51	674.0	9.0
52	708.0	5.0
53	478.0	5.0
61	737.0	5.0
62	497.0	5.0
71	530.0	6.0
72	336.0	7.0
81	572.0	4.0
82	396.0	9.0
91	529.0	11.0
92	461.0	1.0
[total]	14986.0	154.0

```
Age
 '(-inf-10.333333]'        6443.0   68.0
 '(10.333333-15.666667]'   5768.0   28.0
 '(15.666667-inf)'         2753.0   36.0
 [total]                  14964.0  132.0

Religion
 1                         8525.0   60.0
 2                         3087.0   50.0
 3                         2059.0   14.0
 4                         1290.0    9.0
 9                            5.0    1.0
 [total]                  14966.0  134.0

Is_Active
 1                           10.0   12.0
 2                        14953.0  119.0
 [total]                  14963.0  131.0

Is_Ill_Disable11
 1                          457.0   62.0
 2                        14506.0   69.0
 [total]                  14963.0  131.0

Time taken to build model: 0.06 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        5081               99.0255 %
Incorrectly Classified Instances        50                0.9745 %
Kappa statistic                          0.1053
Mean absolute error                      0.0165
Root mean squared error                  0.0952
Relative absolute error                 93.5984 %
Root relative squared error             96.8788 %
Total Number of Instances             5131

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.999    0.94      0.991     0.999     0.995      0.829    1
                 0.06     0.001     0.5       0.06      0.107      0.829    2
Weighted Avg.    0.99     0.931     0.986     0.99      0.986      0.829

=== Confusion Matrix ===

   a    b   <-- classified as
5078    3 |   a = 1
  47    3 |   b = 2
```

Figure 9.12: Naïve Bayes Classifier for No Schooling

=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A
"weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R
first-last\""
Relation: Noschooling2-weka.filters.unsupervised.attribute.NumericToNominal-Rlast
Instances: 15090
Attributes: 6
District
Age
Religion
Is_Active
Is_Ill_Disable1 1
School_Education
Test mode:split 66.0% train, remainder test

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

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification

Time taken to build model: 0.03 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	5067	98.7527 %
Incorrectly Classified Instances	64	1.2473 %
Kappa statistic	0.1737	
Mean absolute error	0.0155	
Root mean squared error	0.1134	
Relative absolute error	87.9197 %	
Root relative squared error	115.4595 %	
Total Number of Instances	5131	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.996	0.86	0.992	0.996	0.994	0.85	1
	0.14	0.004	0.25	0.14	0.179	0.85	2
Weighted Avg.	0.988	0.852	0.984	0.988	0.986	0.85	

=== Confusion Matrix ===

a	b	<-- classified as
5060	21	a = 1
43	7	b = 2

Figure 9.13: K-nearest neighbours for No schooling

Data Mining with WEKA-Clustering

Appendix E

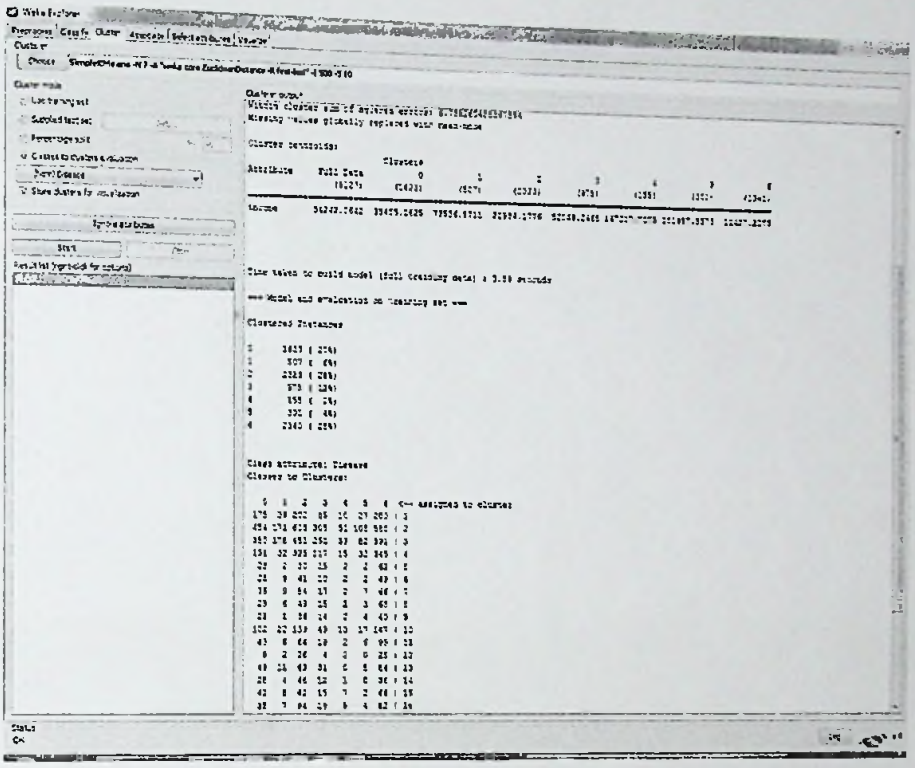

Figure 9.14: KMean clustering Results Window

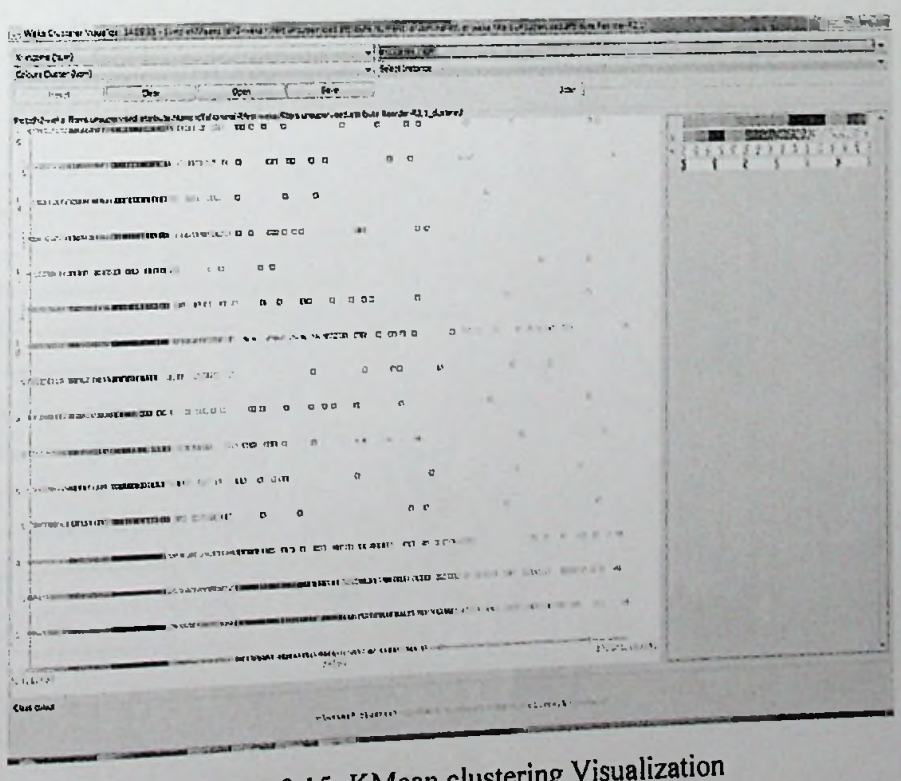

Figure 9.15: KMean clustering Visualization

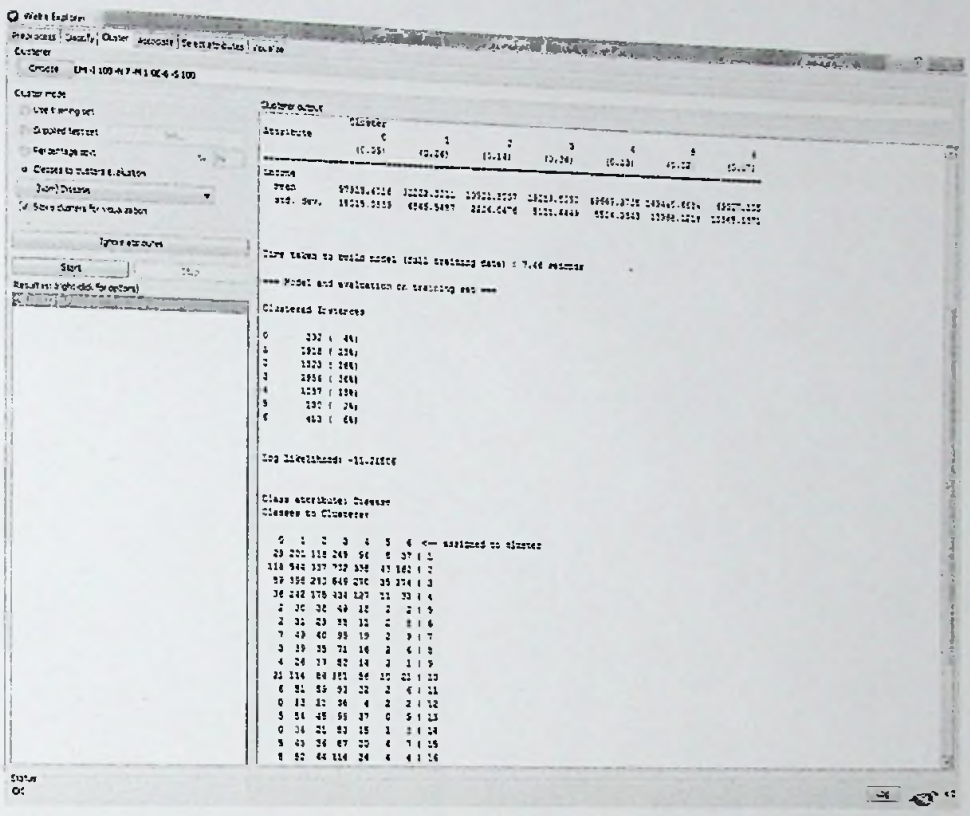

Figure 9.16: EM clustering Results Window

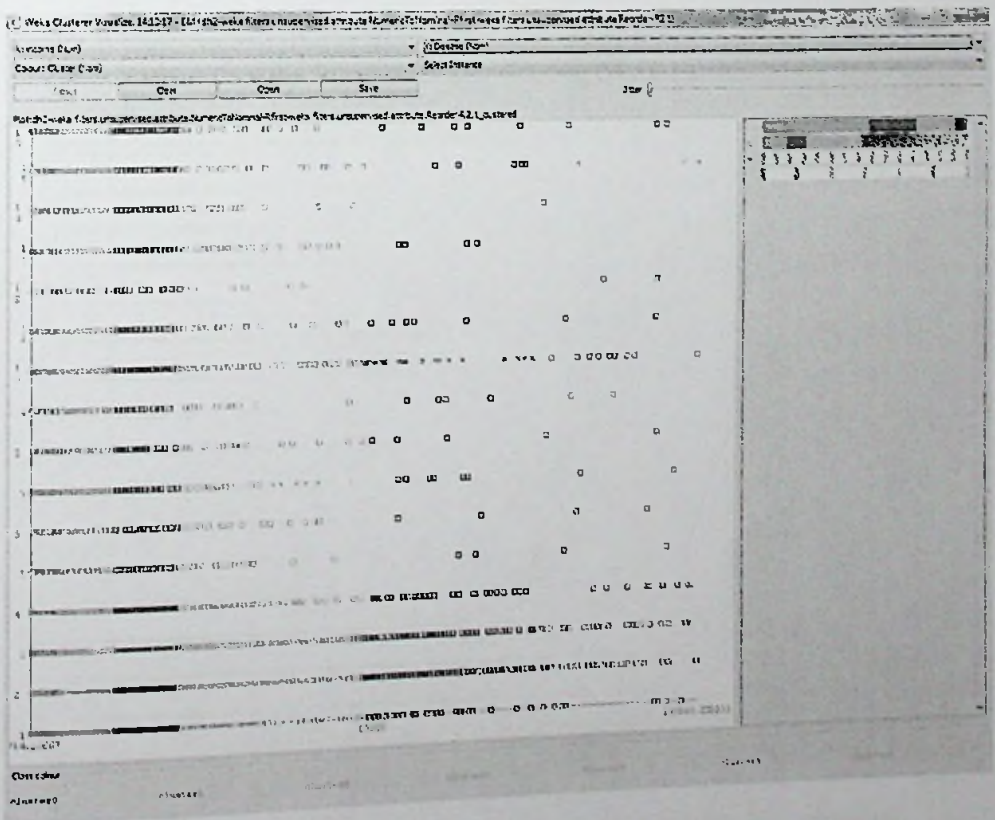

Figure 9.17: EM clustering Visualization

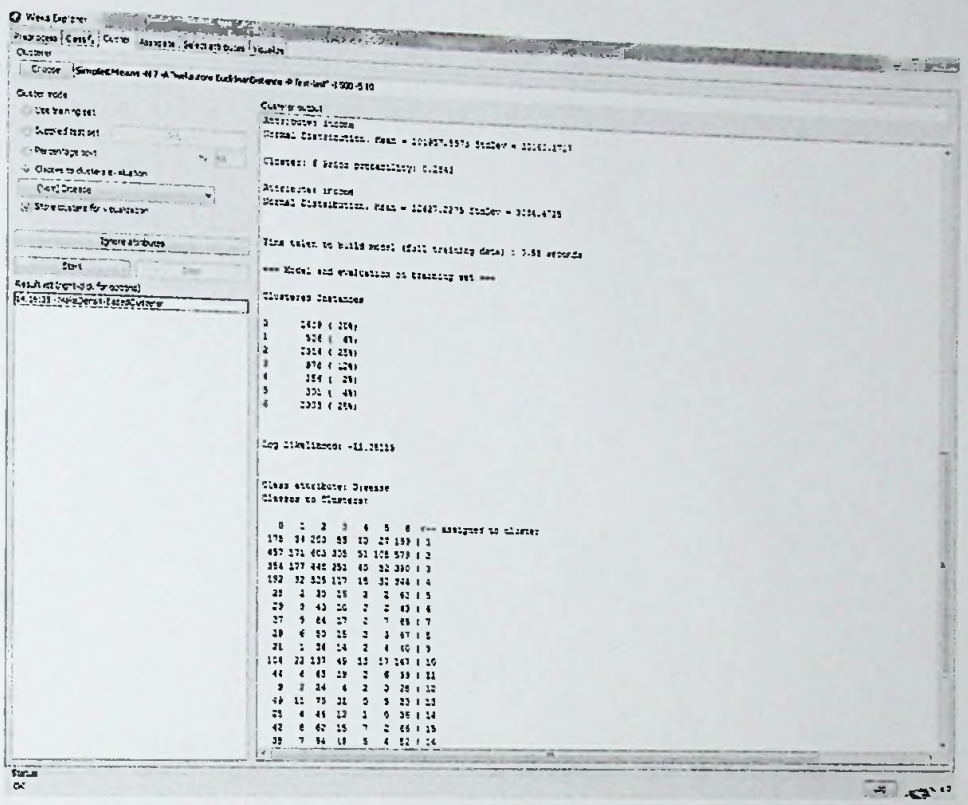

Figure 9.18: MakeDensityBasedClusterer clustering Results Window

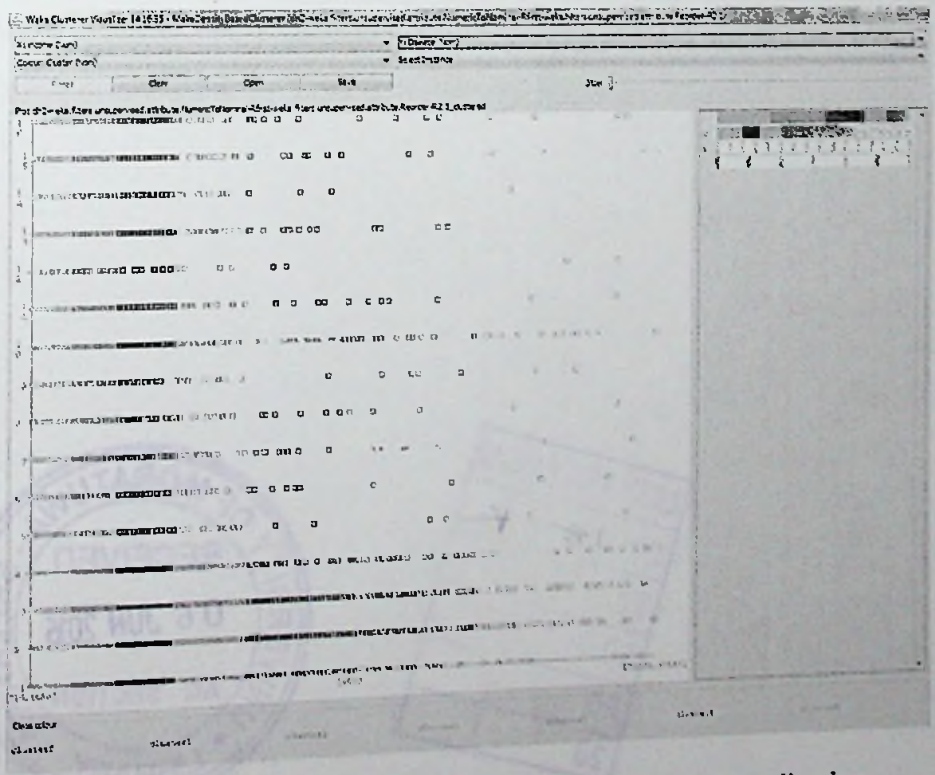

Figure 9.19: MakeDensityBasedClusterer clustering Visualization